The 5th Workshop on Building and Using Comparable Corpora

Special Theme: "Language Resources for Machine Translation in Less-Resourced Languages and Domains"

LREC2012 Workshop
26 May 2012
Istanbul, Turkey

Editors

Reinhard Rapp	University of Leeds and University of Mainz
Marko Tadić	University of Zagreb, Faculty of Humanities and Social Sciences
Serge Sharoff	University of Leeds
Pierre Zweigenbaum	LIMSI-CNRS and ERTIM-INALCO, Orsay

Workshop Organizing Committee

Reinhard Rapp	University of Leeds and University of Mainz
Marko Tadić	University of Zagreb, Faculty of Humanities and Social Sciences
Serge Sharoff	University of Leeds
Pierre Zweigenbaum	LIMSI-CNRS and ERTIM-INALCO, Orsay
Andrejs Vasiļjevs	Tilde, Riga

Workshop Programme Committee

Srinivas Bangalore	AT&T Labs, USA
Caroline Barrière	National Research Council Canada
Chris Biemann	Microsoft / Powerset, San Francisco, USA
Lynne Bowker	University of Ottawa, Canada
Hervé Déjean	Xerox Research Centre Europe, Grenoble, France
Andreas Eisele	DFKI, Saarbrücken, Germany
Rob Gaizauskas	University of Sheffield, UK
Éric Gaussier	Université Joseph Fourier, Grenoble, France
Nikos Glaros	ILSP, Athens, Greece
Gregory Grefenstette	Exalead/Dassault Systemes, Paris, France
Silvia Hansen-Schirra	University of Mainz, Germany
Kyo Kageura	University of Tokyo, Japan
Adam Kilgarriff	Lexical Computing Ltd, UK
Natalie Kübler	Université Paris Diderot, France
Philippe Langlais	Université de Montréal, Canada
Tony McEnery	Lancaster University, UK
Emmanuel Morin	Université de Nantes, France
Dragos Stefan Munteanu	Language Weaver Inc., USA
Lene Offersgaard	University of Copenhagen, Denmark
Reinhard Rapp	Universities of Mainz, Germany, and Leeds, UK
Sujith Ravi	Yahoo! Research, Santa Clara, CA, USA
Serge Sharoff	University of Leeds, UK
Michel Simard	National Research Council Canada
Inguna Skadiņa	Tilde, Riga, Latvia
Monique Slodzian	INALCO, Paris, France
Benjamin Tsou	The Hong Kong Institute of Education, China
Dan Tufis	Romanian Academy, Bucharest, Romania
Justin Washtell	University of Leeds, UK
Michael Zock	LIF, CNRS Marseille, France
Pierre Zweigenbaum	LIMSI-CNRS and ERTIM-INALCO, Orsay, France

Workshop Programme

Saturday, 26 May 2012

09:00 – 09:10 **Opening**

Oral Presentations 1: Multilinguality (Chair: Pierre Zweigenbaum)

09:10 – 09:30 Philipp Petrenz, Bonnie Webber: *Robust Cross-Lingual Genre Classification through Comparable Corpora*

09:30 – 09:50 Qian Yu, François Yvon, Aurélien Max: *Revisiting sentence alignment algorithms for alignment visualization and evaluation*

Invited Projects Session (Chair: Serge Sharoff)

09:50 – 10:10 Inguna Skadiņa: *Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation* (ACCURAT, http://www.accurat-project.eu)

10:10 – 10:30 Andrejs Vasiļjevs: *LetsMT! – Platform to Drive Development and Application of Statistical Machine Translation* (LetsMT!, http://www.letsmt.eu)

10:30 – 11:00 **Coffee Break**

Invited Project Session (Contd.)

11:00 – 11:20 Núria Bel, Vassilis Papavasiliou, Prokopis Prokopidis, Antonio Toral, Victoria Arranz: *Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform* (PANACEA, http://panacea-lr.eu)

11:20 – 11:40 Adam Kilgarriff, George Tambouratzis: *The PRESEMT Project* (PRESEMT, http://www.presemt.eu)

11:40 – 12:00 Béatrice Daille: *Building Bilingual Terminologies from Comparable Corpora: The TTC TermSuite* (TTC, http://www.ttc-project.eu)

12:00 – 12:30 **Panel Discussion with Invited Speakers**

12:30 – 14:00 **Lunch Break**

Oral Presentations 2: Building Comparable Corpora (Chair: Reinhard Rapp)

14:00 – 14:20 Aimée Lahaussois, Séverine Guillaume: *A viewing and processing tool for the analysis of a comparable corpus of Kiranti mythology*

14:20 – 14:40 Nancy Ide: *MultiMASC: An Open Linguistic Infrastructure for Language Research*

Booster Session for Posters (Chair: Marko Tadić)

14:40 – 14:45 Elena Irimia: *Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for: English-Romanian language pair*

14:45 – 14:50 Iustina Ilisei, Diana Inkpen, Gloria Corpas, Ruslan Mitkov: *Romanian Translational Corpora: Building Comparable Corpora for Translation Studies*

14:50 – 14:55 Angelina Ivanova: *Evaluation of a Bilingual Dictionary Extracted from Wikipedia*

14:55 – 15:00 Quoc Hung-Ngo, Werner Winiwarter: *A Visualizing Annotation Tool for Semi-Automatical Building a Bilingual Corpus*

15:00 – 15:05 Lene Offersgaard, Dorte Haltrup Hansen: *SMT systems for less-resourced languages based on domain-specific data*

15:05 – 15:10 Magdalena Plamada, Martin Volk: *Towards a Wikipedia-extracted Alpine Corpus*

15:10 – 15:15 Sanja Štajner, Ruslan Mitkov: *Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness*

15:15 – 15:20 Dan Ştefănescu: *Mining for Term Translations in Comparable Corpora*

15:20 – 15:25 George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos, Marina Vassiliou: *Accurate phrase alignment in a bilingual corpus for EBMT systems*

15:25 – 15:30 Kateřina Veselovská, Ngụy Giang Linh, Michal Novák: *Using Czech-English Parallel Corpora in Automatic Identification of* It

15:30 – 15:35 Manuela Yapomo, Gloria Corpas, Ruslan Mitkov: *CLIR- and Ontology-Based Approach for Bilingual Extraction of Comparable Documents*

15:35 – 16:30 **Poster Session and Coffee Break** (coffee from 16:00 – 16:30)

Oral Presentations 3: Lexicon Extraction and Corpus Analysis (Chair: Andrejs Vasiļjevs)

16:30 – 16:50 Amir Hazem, Emmanuel Morin: *ICA for Bilingual Lexicon Extraction from Comparable Corpora*

16:50 – 17:10 Hiroyuki Kaji, Takashi Tsunakawa, Yoshihoro Komatsubara: *Improving Compositional Translation with Comparable Corpora*

17:10 – 17:30 Nikola Ljubešić, Špela Vintar, Darja Fišer: *Multi-word term extraction from comparable corpora by combining contextual and constituent clues*

17:30 – 17:50 Robert Remus, Mathias Bank: *Textual Characteristics of Different-sized Corpora*

17:50 – 18:00 **Wrapup discussion and end of the workshop**

Author Index

Invited Speakers

Núria Bel	University Pompeu Fabra, Barcelona, Spain
Béatrice Daille	University of Nantes, France
Adam Kilgarriff	Lexical Computing Ltd., UK
Inguna Skadiņa	Tilde, Riga, Latvia
Andrejs Vasiļjevs	Tilde, Riga, Latvia

Table of Contents

Preface

Following the four previous editions of the Workshop on Building and Using Comparable Corpora which took place at LREC 2008 in Marrakech, at ACL-IJCNLP 2009 in Singapore, at LREC 2010 in Malta, and at ACL-HLT 2011 in Portland, this year the workshop was co-located with LREC 2012 in Istanbul.

Although papers on all topics related to comparable corpora were welcome at the workshop, this year's special theme was "Language Resources for Machine Translation in Less-Resourced Languages and Domains". This theme was chosen with the aim of finding ways to overcome the shortage of parallel resources when building machine translation systems for less-resourced languages and domains. Lack of sufficient language resources for many language pairs and domains is currently one of the major obstacles in the further advancement of machine translation. Possible solutions include the identification of parallel segments within comparable corpora or reaching out for parallel data that is 'hidden' in users' repositories.

To highlight the increasing interest in comparable corpora and the success of the field, representatives from five international research projects were invited to present the important role of work on comparable corpora within a special session. These projects were ACCURAT (http://www.accurat-project.eu/), LetsMT! (https://www.letsmt.eu/), PANACEA (http://panacea-lr.eu/), PRESEMT (http://www.presemt.eu/), and TTC (http://www.ttc-project.eu/).

We would like to thank all people and institutions who helped in making this workshop a success. This year the workshop has been formally endorsed by ACL SIGWAC (Special Interest Group on Web as Corpus), FLaReNet (Fostering Language Resources Network), and META-NET (Multilingual Europe Technology Alliance). Our special thanks go to the representatives of the above mentioned projects for accepting our invitations, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the organizers of the hosting conference. Last but not least we would like to thank our authors and the participants of the workshop.

Reinhard Rapp
Marko Tadić
Serge Sharoff
Pierre Zweigenbaum

Robust Cross-Lingual Genre Classification through Comparable Corpora

Philipp Petrenz, Bonnie Webber

University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB, UK
p.petrenz@sms.ed.ac.uk, bonnie@inf.ed.ac.uk

Abstract

Classification of texts by genre can benefit applications in Natural Language Processing and Information Retrieval. However, a mono-lingual approach requires large amounts of labeled texts in the target language. Work reported here shows that the benefits of genre classification can be extended to other languages through cross-lingual methods. Comparable corpora – here taken to be collections of texts from the same set of genres but written in different languages – are exploited to train classification models on multi-lingual text collections. The resulting genre classifiers are shown to be robust and high-performing when compared to mono-lingual training sets. The work also shows that comparable corpora can be used to identify features that are indicative of genre in various languages. These features can be considered stable genre predictors across a set of languages. Our experiments show that selecting stable features yields significant accuracy gains over the full feature set, and that a small amount of features can suffice to reliably distinguish between different genres.

Keywords: Genre, Text Classification, Cross-Lingual, Comparable Corpora

1. Introduction

Automated text classification has become standard practice with applications in fields such as information retrieval and natural language processing. The most common basis for text classification is by topic (Joachims, 1998; Sebastiani, 2002), but other classification criteria have evolved, including sentiment (Pang et al., 2002), authorship (de Vel et al., 2001; Stamatatos et al., 2000a), and author personality (Oberlander and Nowson, 2006), as well as categories relevant to filter algorithms (e.g., spam or inappropriate contents for minors).

Genre is another text characteristic, often described as orthogonal to topic. It has been shown by Biber (1988) and others after him, that the genre of a text affects its formal properties. It is therefore possible to use cues (e.g., lexical, syntactic, structural) from a text as features to predict its genre, which can then feed into information retrieval applications (Karlgren and Cutting, 1994; Kessler et al., 1997; Finn and Kushmerick, 2006; Freund et al., 2006). This is because users may want documents that serve a particular communicative purpose, as well as being on a particular topic. For example, a web search on the topic "crocodiles" may return an encyclopedia entry, a biological fact sheet, a news report about attacks in Australia, a blog post about a safari experience, a fiction novel set in South Africa, or a poem about wildlife. A user may reject many of these, just because of their genre: Blog posts, poems, novels, or news reports may not contain the kind or quality of information she is seeking. Having classified indexed texts by genre would allow additional selection criteria to reflect this.

Genre classification can also benefit Language Technology indirectly, where differences in the cues that correlate with genre may impact system performance. For example, Petrenz and Webber (2011) found that within the New York Times corpus (Sandhaus, 2008), the word "states" has a higher likelihood of being a verb in letters (approx. 20%) than in editorials (approx. 2%). Part-of-Speech (PoS) taggers or statistical machine translation systems could benefit from knowing such genre-based domain variation. Kessler et al. (1997) mention that parsing and word-sense disambiguation can also benefit from genre classification. Webber (2009) found that different genres have a different distribution of discourse relations, and Goldstein et al. (2007) showed that knowing the genre of a text can also improve automated summarization algorithms, as genre conventions dictate the location and structure of important information within a document.

All the above work has been done within a single language. Recent work by one of the current authors (Petrenz, 2012) demonstrated a new approach to genre classification that is cross-lingual (CLGC) in that it trains a genre classification model solely on labeled texts from one language L_S and then uses this model to predict the genres of texts written in another language L_T. As such, CLGC differs from both poly-lingual and language-independent genre classification in requiring *no labeled training data in the target language* (L_T). Instead, it attempts to leverage the available annotated data in well-resourced languages like English in order to bring the aforementioned advantages to poorly-resourced languages. This reduces the need for manual annotation of text corpora in the target language.

What is new in the current work is that we show that there is even greater benefit to be gained from the use of a comparable corpus, comprising texts in several languages, in training a genre classifier for texts of the target language (L_T), different from any in the comparable corpus.

The paper is structured as follows: Section 2. describes prior work on genre classification, including our own. Section 3. describes our approach based on a comparable corpus, Section 4. describes the set of experiments we carried out and Section 5. discusses the results. Finally, Section 6. concludes with thoughts on taking this work forward.

2. Prior work

Work on automated genre classification was first carried out by Karlgren and Cutting (1994). Like Kessler et al. (1997) after them, they exploited hand-crafted sets of features, which are specific to texts in English. In subsequent research, automatically generated feature sets have become more popular. Most of these tend to be language-independent and might work in mono-lingual genre classification tasks in languages other than English. Examples include word based approaches (Argamon et al., 1998; Stamatatos et al., 2000b; Freund et al., 2006), PoS trigrams (Argamon et al., 1998) and PoS history frequencies (Feldman et al., 2009), image features (Kim and Ross, 2008), and character n-gram approaches (Kanaris and Stamatatos, 2007; Sharoff et al., 2010), all of which were tested exclusively on English texts. One of the few researchers to assess the language-independence of their approach was Sharoff (2007). Using PoS 3-grams and a variation of common word 3-grams as feature sets, Sharoff classified English and Russian documents into genre categories, although in both cases his experiments were mono-lingual.

The only work on CLGC to date has been that of Petrenz (2012). This makes use of a set of hand-crafted stable features to bridge the language gap between English and Chinese, and then a bootstrapping technique to exploit unlabeled data in the target language. The approach performs equally well or better than a baseline in which texts are automatically translated and a mono-lingual genre classifier applied to the result. However, classifiers were only trained on a single language (English or Chinese), rather than exploiting the additional knowledge that might be available in comparable corpora. The notion of *stable features* used by Petrenz and Webber (2011) to specify features that are unaffected (i.e., stable) in the face of changing topics, could be applied here to specify features that are stable in the face of changing languages.

Cross-lingual methods have been explored for other text classification tasks. The first to report such experiments were Bel et al. (2003), who predicted text topics in Spanish and English documents, using one language for training and the other for testing. Their approach involves training a classifier on language A, using a document representation containing only content words (nouns, adjectives, and verbs with a high corpus frequency). These words are then translated from language B to language A, so that texts in either language are mapped to a common representation.

Thereafter, cross-lingual text classification was typically regarded as a domain adaptation problem that researchers have tried to solve using large sets of unlabeled data and/or small sets of labeled data in the target language. For instance, Rigutini et al. (2005) present an EM algorithm in which labeled source language documents are translated into the target language and then a classifier is trained to predict labels on a large, unlabeled set in the target language. These instances are then used to iteratively retrain the classification model and the predictions are updated until convergence occurs. Using information gain scores at every iteration to only retain the most predictive words and thus reduce noise, Rigutini et al. (2005) achieve a considerable improvement over the baseline accuracy, which is a simple translation of the training instances and subsequent mono-lingual classification. They, too, were classifying texts by topics and used a collection of English and Italian newsgroup messages. Similarly, researchers have used semi-supervised bootstrapping methods like co-training (Wan, 2009) and other domain adaptation methods like structural component learning (Prettenhofer and Stein, 2010) to carry out cross-lingual text classification.

All of the approaches described above rely to some extent on statistical machine translation. This makes applications dependent on parallel corpora, which may not be available for poorly-resourced languages. It also suffers problems due to word ambiguity and morphology, especially where single words are translated out of context. A different method is proposed by Gliozzo and Strapparava (2006), who use Latent Semantic Analysis on a comparable corpus of texts written in two languages. The rationale is that named entities such as "Microsoft" or "HIV" are identical in different languages with the same writing system. Using term correlation, the algorithm can identify semantically similar words in both languages. The authors exploit these mappings in cross-lingual topic classification, and their results are promising. However, they also report considerable from using bilingual dictionaries.

While all of the methods above could technically be used in any text classification task, the idiosyncrasies of genres pose additional challenges. Techniques relying on automated translation of predictive terms (Bel et al., 2003; Prettenhofer and Stein, 2010) are workable in the contexts of topics and sentiment, as these typically rely on content words such as nouns, adjectives, and adverbs. For example, "hospital" may indicate a text from the medical domain, while "excellent" may indicate that a review is positive. Such terms are relatively easy to translate, even if not always without ambiguity. Genres, on the other hand, are often classified using function words (Karlgren and Cutting, 1994; Stamatatos et al., 2000b) like "of", "it", or "in", which are next to impossible to translate out of context, especially when morphological differences between the languages can mean that function words in one language are morphological affixes in another.

Although it is theoretically possible to use the bilingual low-dimension approach by Gliozzo and Strapparava (2006) for genre classification, it relies on certain lexical identities in the two languages. While this may be the case for topic-indicating named entities — a text containing the words "Obama" and "McCain" will almost certainly be about the U.S. elections in 2008, or at least about U.S. politics — it is less indicative of genre: The text could be *inter alia* a news report, an editorial, a letter, an interview, a biography, or a blog entry, although correlations between topics and genres would probably rules out genres like instruction manuals or product reviews. However, uncertainty is still large, and Petrenz and Webber (2011) show that it can be dangerous to rely on such correlations.

3. Approach

The experiments described in Section 4. exploit features that are comparable across languages and a corpus of comparable texts across the same set of languages. We describe

both here before going into detail about the experiments.

3.1. Stable features

Many types of features have been used in genre classification. They all fall into one of three groups: *Language-specific features* are cues which can only be extracted from texts in one language. An example would be the frequency of a particular word, such as "yesterday". *Language-independent features* can be extracted in any language, but they are not necessarily directly comparable. Examples would be the frequencies of the most common words. While these can be extracted for any language (as long as words can be identified as such), the function of a word on a certain position in this ranking will likely differ from one language to another. *Comparable features*, on the other hand, serve a similar role in two or more languages. An example would be type/token ratios, which, in combination with document length, represent the lexical richness of a text, independent of its language. If such features prove to be good genre predictors across languages, they may be considered *stable* across those languages. If suitable features can be identified, CLGC may be considered a standard classification problem.

The approach we propose, like the one in (Petrenz, 2012), makes use of stable features that are mainly structural rather than lexical (cf. Section 4.2.) since the latter tend to vary by topic and are thus *unstable* with respect to genre (Petrenz and Webber, 2011). It does not assume the availability of machine translation, supervised PoS taggers, syntactic parsers, or other supervised tools. The only resources required are a way to detect sentence and paragraph boundaries in both source and target languages (e.g., a simple rule-based algorithm or an unsupervised method), and a sufficiently large, unlabeled set of target-language texts.

3.2. Hypotheses related to comparable corpora

The experiments described in Section 4. are designed to test two hypotheses: First, a comparable corpus of texts written in different languages but from the same distribution of genres can be used to train a classification model that is more robust for cross-lingual classification tasks than a model trained on a mono-lingual training set whose genre-related differences might not be the same as those in the target language. Adding more languages to the training set will result in a classification model which can separate genres in multiple languages. This makes it more likely to perform well on the target language.

The second hypothesis is that selecting features based on the cross-lingual performance within a separate comparable corpus can prevent a classifier from over fitting to the idiosyncrasies of the training language. Using a supervised feature selection technique on a set of several languages may yield features that have predictive power in more than one language. Cross-lingual genre classification can be regarded a special case of a domain adaptation problem, where feature selection techniques have been applied successfully before (Pan et al., 2010). Here, we apply a simple feature-ranking method, using information gain to determine the value of a feature to predict genres. Information gain is defined as

$$IG(Class, Feature) = H(Class) - H(Class|Feature)$$

where $H(X)$ is the entropy of variable X. A subset of features can then be obtained by choosing the top k features in this ranking of n features. While the availability of domain knowledge would allow this parameter to be set manually, here we determine it automatically, by finding the maximum cross-validation accuracy on the comparable corpus, where each fold corresponds to training on a single language and testing on all remaining languages. While this involves an exhaustive search over all possible values of k, using the information–gain ranking greatly reduces the possible numbers of feature subsets from $2^n - 1$ to n.

Note that, unlike the method in Gliozzo and Strapparava (2006), discussed in Section 2., the current approach does not require the comparable corpus to include texts from the target language.

4. Experiments

4.1. Data

Our experiments use three publicly available corpora, each of which included texts from a single genre written in several languages: the Reuters volume 1+2 corpus (Rose et al., 2002), the Europarl corpus (Koehn, 2005), and the JRC-ACQUIS corpus (Steinberger et al., 2006). All three corpora contain a large number of texts in Danish, English, French, German, Italian, Portuguese, Spanish, and Swedish. (Although all three also contain texts in Dutch, there are comparatively few Dutch texts in the Reuters corpus, so Dutch texts are not used in our experiments.) We reorganized the source corpora to obtain a comparable corpus that contains texts in eight languages and three genres: newswire texts, transcribed speech, and legal texts. Note that the corpus is *comparable* since it contains texts from a fixed set of genres, but not necessarily topics.

Since the source corpora are in different formats, some preprocessing was necessary. The XML markup was removed from the Reuters newswire texts, and only the contents of the tags `<headline>`, `<byline>`, `<dateline>`, and `<text>` were kept. Paragraph markers were kept in the text. The texts in the Europarl corpus were divided up by speaker: that is, we considered each speech to be a distinct document. We then removed the `<speaker>` tags, but kept the paragraph markers. We ignored missing speeches: The only requirement was that each text contains at least one token. The JRC-ACQUIS corpus comprises several sub-genres within the legal domain, including treaties, agreements and proposals. We therefore restricted ourselves to using documents from CELEX[1] sector 3 (legislation), as this is the largest group within the corpus. We extracted the text within the `<body>` tags, again keeping the paragraph structure intact.

All texts were segmented into sentences using the unsupervised *Punkt* algorithm (Kiss and Strunk, 2006) implemented in the NLTK (Bird et al., 2009) framework. Since

[1] CELEX (Communitatis Europeae Lex) is a database for European Union law documents. All texts in the JRC-ACQUIS corpus are classified by CELEX sector and document type.

Europarl and JRC-ACQUIS are parallel corpora, we ensured that no translation of the same text was used in any two sets in our experiments. For Europarl texts, we always used the language that the speech was made in, which is indicated in the meta-data. For JRC-ACQUIS, the choice was random, since the corresponding journal is published in all European languages simultaneously.

Splitting the legislation texts of the JRC-ACQUIS yielded 1,942 documents in each of the eight languages. To keep the genre distribution in our corpus balanced, we randomly sampled 1,942 documents from both the Reuters and the Europarl corpora. The resulting eight sets each contained 5,826 texts from a single language. A list with identifiers of the texts we used for our experiments can be found on our website[2], along with scripts to extract and clean texts from the source corpora mentioned before. There is, to the best of our knowledge, no publicly available corpus containing texts written in several languages from a common set of genres. Therefore, the method described above can be seen as a suggestion to facilitate research into cross-lingual genre classification and provide a common data set to compare approaches.

4.2. Features

We hypothesized that our experiments could produce a set of features that would serve as stable genre predictors across a range of languages, not just for a single one as in (Petrenz and Webber, 2011). To this end, we selected as candidate features, ones that would hold for texts in many languages. These included the frequencies of 32 common punctuation symbols, as well as simple text statistics (document length, sentence length mean and variance, paragraph length mean and variance, single-sentence-paragraph count and frequency over all sentences, single-sentence-paragraph distribution value, type/token ratio[3], and number/token ratio).

Single-sentence paragraphs are typically headlines, datelines, author names, or other structurally interesting parts. Their distribution value indicates how evenly they are distributed throughout a text, with high values indicating single-sentence paragraphs predominantly occurring at the beginning and/or end of a text. It is computed by averaging over the distance of all such paragraphs from the $(n/2)$th token in a text of length n.

To this set, we added features based on concepts from information retrieval. We used tf-idf weighting and marked the ten highest-weighted words in a text as relevant. We then treated the text as a ranked list of relevant and non-relevant words, where the position of a word in the text determined its rank. This allowed us to compute an average precision (AP) value, which indicates the distribution of relevant words. A high AP score means that the top tf-idf weighted words are found predominantly in the beginning of a text. This follows the intuition that genre conventions may influence the location of important content words within a text. For example, Thomson et al. (2008) found that news articles in English, French, Japanese, and Indonesian are all structured according to the inverted pyramid principle (Pöttker, 2003), where important information appears in the beginning, followed by background information and other less important material. In addition, for each of the same ten words, we added its tf-idf value to the feature set, divided by the sum of all ten. These values indicate whether a text is very focused (a sharp drop between higher and lower ranked words) or more spread out across topics (relatively flat distribution).

Finally, we also added the frequencies in the text of the 25 most common words in the respective language. Common word frequencies have been shown to have discriminative power in mono-lingual genre classification tasks (Stamatatos et al., 2000a). However, since the i^{th} most common word in language A differs semantically from the i^{th} most common word in language B, we expected these features to be of little value for a cross-lingual task and that they might have a negative impact on prediction accuracies. We included them in the feature set to find out whether this is the case and if so, whether they are filtered out in the feature selection process of our method.

The final set comprised 78 features, three of which were discarded, as they had zero values for all texts in one or more languages. After extracting the full set of features from the texts, their values were standardized. This was achieved by subtracting from each feature value the mean over all texts and dividing it by the standard deviation, so that each feature had zero mean and unit variance. Standardization was done separately for each language, to balance out differences between them. Because this step exploits only unlabeled data in order to make feature values more comparable (i.e., it does not require genre labels), standardization can be applied to the target language feature set, as long as enough target language texts are available.

4.3. Experimental Frameworks

To generate baselines, we evaluated classification models which were trained on one language and tested on another. To this end, we trained a separate Support Vector Machine (SVM) model for each of the eight mono-lingual sets, using all 75 features. Each model was then tested on the seven languages that were not used to train it. This performance is achievable without the use of a comparable corpus.

To exploit the genre labels in more than just one language, we then merged the representations of seven language sets into a single training set, holding one language back for testing. An example of this is illustrated in Figure 1. Naturally, the merged multi-lingual training set contained seven times as many texts as any mono-lingual baseline. Since supervised classification results tend to improve with larger training set sizes, we removed this bias by splitting the merged set into seven disjoint training sets, keeping the language and genre distributions intact. Thus, for each target language, the SVM model was trained seven times and evaluated by computing the average accuracy.

To evaluate whether a comparable corpus can be used to identify stable features from a set of candidates, even if the

[2] http://homepages.inf.ed.ac.uk/s0895822/BUCC2012/

[3] As the type/token ratio is known to correlate with document size, we recorded the ratio for a sliding window of 300 tokens. For shorter texts, this was estimated by computing a percentage of the average type/token ratio at the end of the text and multiplying this with the average value for 300 tokens.

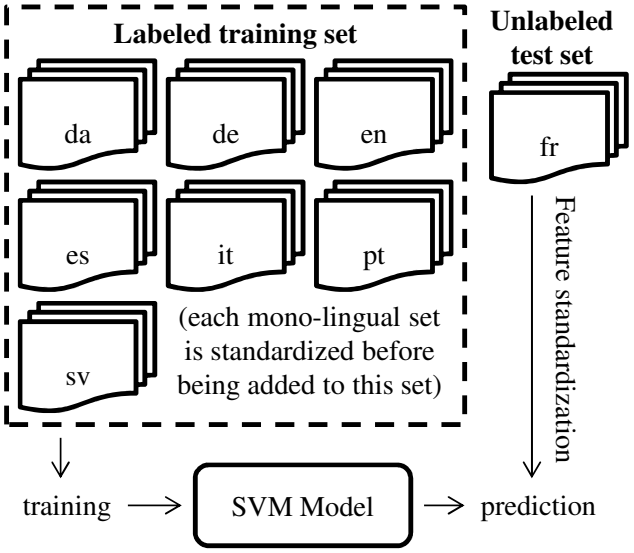


Figure 1: First experimental framework, example with French test set. Set of seven languages is used to train classification model. The full set of features is used.

set does not include texts written in the source or target languages, we conducted a second experiment. Here, we ranked features using a set of six languages. (Features were ranked by their information gain, as explained in Section 3.) Then, 6-fold cross-validation was used to determine the threshold parameter k. The feature sets of the seventh and eighth languages were reduced to the resulting subset, and then used for training and testing respectively. An example of this is illustrated in Figure 2. When compared with the baseline, the results will indicate to what extent feature selection on a separate comparable corpus can benefit cross-lingual genre classification applications.

5. Results and Discussion

Table 1 shows the classification accuracies for the 56 single language training experiments (i.e. baseline performances), as well as the accuracies yielded by the combined multi-lingual training set. The last row corresponds to the experimental framework illustrated in Figure 1. *For all eight target languages, accuracy based on the multi-lingual training set exceeded accuracy based on any of the seven mono-lingual baselines.* This significant (sign test; $p < 0.01$) improvement indicates that the knowledge represented by genre labels in different languages can be exploited to build robust cross-lingual genre classification models.

In the second experiment, we performed feature selection using the six languages that remained after choosing one language for training and a second for testing (cf. Figure 2). Table 2 shows the gains and losses in prediction accuracy when using only the top k features, as compared to the full feature set. For the 56 tasks, k ranged between 13 and 23, with the majority between 13 and 15. Most classification models benefited from this feature selection step. Although in some cases accuracy deteriorated, performance based on the reduced feature set was significantly better ($p < 1e^{-8}$), according to the sign test. Since these subsets were identified using a supervised ranking technique, the results in

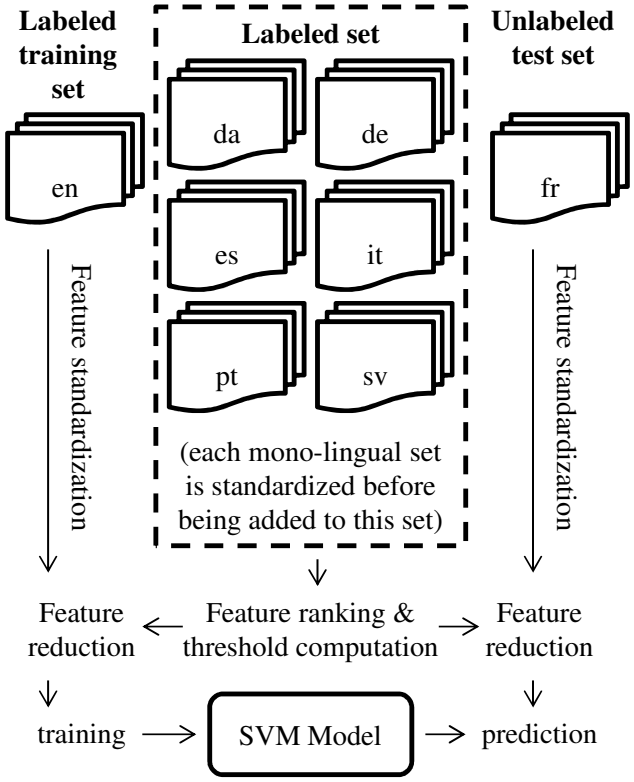


Figure 2: Second experimental framework, example with English training set and French test set. Set of six languages is used to rank features and determine the threshold k. The languages used for training and testing are not represented in this set.

Table 2 suggest that comparable corpora can also be used to identify features with strong discriminative powers for cross-lingual genre classification tasks. They also show that this is possible even if neither the source nor the target language is included in the comparable corpus.

An important question is whether the algorithm can find a good value for the threshold k. Using the results in Table 2, we picked the combination that gained the most from the feature reduction (training on Spanish texts, testing on German texts: es→de) and the one that suffered the most (training on Portuguese texts, testing on English texts: pt→en). We also picked the combination that used the largest number of features (training on Danish texts, testing on Italian texts: da→it). For these three combinations, we recorded the performance when removing features from the set one by one, starting at the performance of the full set shown in Table 1. Figure 3 illustrates the prediction accuracies as functions of the number of features used. The arrows indicate the threshold chosen by the algorithm. The es→de classifier performs clearly better when selecting between 12 and 22 features from the ranking. The threshold (14) happens to be a very good choice and yields significant[4]

[4]We assume that the number of misclassifications is approximately normally distributed with mean $\mu = e * n$ and standard deviation $\sigma = \sqrt{\mu * (1 - e)}$, where e is the percentage of misclassified instances and n is the size of the test set. The 95% confidence interval is then $\mu \pm 1.96 * \sigma$.

	da	de	en	es	fr	it	pt	sv	μ
Danish (da)	—	.959	.951	.961	.930	.965	.937	.971	.953
German (de)	.943	—	.925	.934	.897	.957	.933	.954	.935
English (en)	.948	.942	—	.961	.934	.962	.942	.972	.952
Spanish (es)	.960	.920	.952	—	.946	.963	.927	.973	.949
French (fr)	.961	.952	.965	.974	—	.973	.940	.967	.962
Italian (it)	.959	.963	.955	.962	.948	—	.949	.953	.956
Portuguese (pt)	.955	.948	.945	.954	.928	.954	—	.961	.949
Swedish (sv)	.965	.949	.948	.963	.911	.947	.928	—	.944
Multi-lingual	**.979**	**.968**	**.973**	**.979**	**.967**	**.980**	**.971**	**.986**	**.975**

Table 1: Prediction accuracies for the cross-lingual genre classification tasks. Rows 2-9 denote the training language, Columns 2-9 denote the testing language. The accuracies in row 10 were achieved by training the model on the seven languages which it was not tested on. Column 10 contains the average of each row. The best accuracy for each column is highlighted.

	da	de	en	es	fr	it	pt	sv
Danish (da)	—	+.005	+.013	+.009	**+.033**	+.011	+.013	−.009
German (de)	+.015	—	+.016	**+.031**	**+.035**	+.009	−.002	−.001
English (en)	**+.021**	+.018	—	**+.022**	**+.040**	+.010	+.005	+.010
Spanish (es)	+.005	**+.062**	**+.021**	—	**+.024**	+.017	**+.035**	+.004
French (fr)	+.015	+.016	+.011	+.017	—	+.000	+.018	+.010
Italian (it)	−.003	+.017	+.011	**+.025**	+.019	—	+.010	+.017
Portuguese (pt)	**+.024**	−.001	**−.026**	**+.025**	+.011	**+.022**	—	+.011
Swedish (sv)	+.009	+.011	**+.025**	+.019	**+.061**	**+.030**	+.017	—

Table 2: Difference in prediction accuracy after feature selection when compared to the corresponding results in Table 1. As in Table 1, rows 2-9 denote the training language, columns 2-9 denote the testing language. Differences of more than .02 are highlighted.

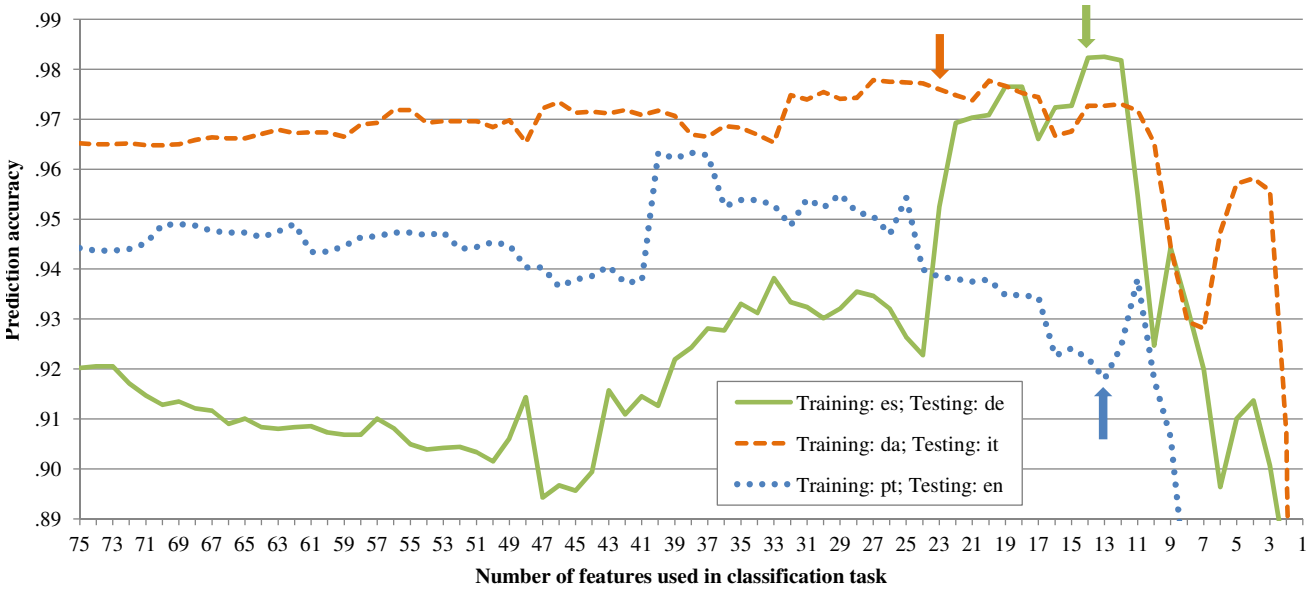


Figure 3: Prediction accuracies for the classifier trained on Spanish texts and tested on German texts (green, continuous line), the classifier trained on Danish texts and tested on Italian texts (orange, dashed line), and the classifier trained on Portuguese texts and tested on English texts (blue, dotted line). For all three classifiers, the accuracy achieved is given as a function of the number of top rank-ordered features used. The arrows denote the automatically determined number of features for these tasks (14, 23, and 13 respectively).

improvement over the baseline. The performance of the pt→en classifier stays mostly within the confidence interval of the baseline, although it clearly outperforms it for feature set sizes 37-40. Accuracy drops and falls below baseline level for fewer than 20 features. Here, the chosen threshold (13) is too low, since this classifier would benefit from additional features. The da→it classifier benefits slightly but significantly from a reduced feature set until accuracy drops sharply for less than 11 features. The threshold (23) is a good choice, although the exact value is less crucial than for the es→de and pt→en classifiers, in that small variations would have little effect on the result.

The majority of positive results in Table 2 suggests that the chosen threshold k is usually suitable to improve the prediction accuracy. In line with that, Figure 3 shows that the algorithm picks a near-optimal value for k for some training/testing combinations. However, the example of the pt→en classifier shows that this is not necessarily the case. On the other hand, it also illustrates that even where feature reduction leads to deteriorating performances, this could be due to a sub-optimal threshold choice. This is clearly the case for the pt→en classifier, where a set of 37-40 features would have improved baseline performance significantly. Optimizing the computation of this threshold, possibly by exploiting the unlabeled data in the target language, would be an interesting problem for future work.

In order to get an idea of the types of features which are typically selected, we ranked them by their information gain using a combined set that included all eight languages. The top 15 features are listed below. Note that the information gain of a certain feature varies depending on the exact set of languages used. However, the ranking in our experiments was fairly stable and the top 15 features rarely differed from the ones below.

1. Single sentence paragraph count
2. Single sentence paragraph/sentence ratio
3. Paragraph length mean
4. Closing parenthesis frequency
5. Opening parenthesis frequency
6. Number frequency
7. Forward slash frequency
8. Single sentence distribution value
9. Colon frequency
10. Sentence length mean
11. Top 10 tf-idf average precision
12. Type/token ratio
13. Document length
14. Paragraph length standard deviation
15. Hyphen frequency

As expected, none of the 25 common-word frequency features was ranked among the top 15. This finding reinforces our intuition that common-word frequencies are useful in mono-lingual genre classification tasks, but harmful to cross-lingual models. While feature 11 above seems to have discriminative power, none of the other tf-idf based features is in the above list. This is likely due to the fact that these features have informative value only in combination with each other. However, information gain ranking evaluates only single features, not sets. A subset based selection approach might be more suitable to identify their strengths (cf. Guyon and Elisseeff (2003)).

Another observation is that features based on paragraph length dominate the ranking. This is likely due to the way texts of the three different genres are structured. Legal texts tend to have very short paragraphs, sometimes consisting of a single token (Example 1 below). Newswire paragraphs are mostly only one or two sentences long, but typically contain more than one token each (Example 2). In transcribed speech (Example 3), paragraphs tend to be longer.

1. **Legal text:**

```
<p>Commission Regulation (EC) No
1135/2006</p>
<p>of 25 July 2006</p>
<p>amending the import duties in the
cereals sector applicable from 26 July
2006</p>
<p>THE COMMISSION OF THE EUROPEAN
COMMUNITIES,</p>
<p>Having regard to the Treaty establishing
the European Community,</p>
```

2. **Newswire text:**

```
<p>The KFX top-20 index lost 0.20 point to
close at 126.29 in overall bourse turnover
of 1.944 billion crowns.  The KFX December
future rose 0.65 point to 126.40 with
10 contracts each worth 100,000 crowns
traded.</p>
<p>Novo Nordisk attracted a good deal of
attention following its announcement of
400 million crown rationalisation cuts for
1997 and 1998, finishing the day a solid 21
crowns up at 954.</p>
```

3. **Transcribed speech text:**

```
<p>Naturally I understand the honourable
Member's concern.  As far as the Commission
is concerned, we have never supported
financially the production or distribution
of school textbooks nor the preparation
of school curricula.  Assistance to the
educational system is focused mainly on
infrastructure, equipment for schools and
direct assistance for school expenses, for
example, salaries.  No request has ever
been made by the Palestinian Authority to
the Commission to finance school curricula
and textbooks.</p>
```

6. Conclusion

Our experiments with eight European languages show that cross-lingual genre classification (at least within these languages) is possible with a minimum of knowledge about the target language. Some features, which are easily extracted from plain texts, can be considered stable predictors of genre across languages. Applications exploiting such features may reduce the need for resources such as parallel corpora or supervised parsers in the target language. We

demonstrate that comparable corpora can be used to automatically identify stable features from a set of candidates. These can help to improve prediction accuracy, even when used in tasks with separate training and target languages.

We also show that using more than one language in the training set can prevent a cross-lingual genre classification model from over fitting the differences between genres in one language and thus improve its robustness. Exploiting a comparable corpus by either identifying stable features or using multi-lingual training sets significantly beats the baseline performances in our experiments.

Finally, we propose a method to construct a comparable corpus including legal texts, newswire texts, and transcribed speeches in eight European languages by remodeling three publicly available corpora. This can be used by researchers to compare cross-lingual genre classification methods.

7. References

Shlomo Argamon, Moshe Koppel, and Galit Avneri. 1998. Routing documents according to style. In *Proceedings of First International Workshop on Innovative Information Systems*.

Nuria Bel, Cornelis Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In Traugott Koch and Ingeborg Slvberg, editors, *Research and Advanced Technology for Digital Libraries*, volume 2769 of *Lecture Notes in Computer Science*, pages 126–139. Springer Berlin / Heidelberg.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

O. de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55–64.

S. Feldman, M. A. Marin, M. Ostendorf, and M. R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4781–4784, Washington, DC, USA. IEEE Computer Society.

Aidan Finn and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *J. Am. Soc. Inf. Sci. Technol.*, 57(11):1506–1518.

Luanne Freund, Charles L. A. Clarke, and Elaine G. Toms. 2006. Towards genre classification for IR in the workplace. In *Proceedings of the 1st international conference on Information interaction in context*, pages 30–36, New York, NY, USA. ACM.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 553–560, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jade Goldstein, Gary M. Ciany, and Jaime G. Carbonell. 2007. Genre identification and goal-focused summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 889–892, New York, NY, USA. ACM.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, London, UK. Springer-Verlag.

Ioannis Kanaris and Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings of the 19th IEEE International Conference on Tools with AI*, pages 3–10, Washington, DC.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075, Morristown, NJ, USA. Association for Computational Linguistics.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38, Morristown, NJ, USA. Association for Computational Linguistics.

Yunhyong Kim and Seamus Ross. 2008. Examining variations of prominent features in genre classification. In *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, HICSS '08, pages 132–, Washington, DC, USA. IEEE Computer Society.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32:485–525, December.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 627–634, Morristown, NJ, USA. Association for Computational Linguistics.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 751–760, New York, NY, USA. ACM.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Petrenz and Bonnie Webber. 2011. Stable clas-

sification of text genres. *Computational Linguistics*, 37(2):385–393.

Philipp Petrenz. 2012. Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.

Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid — when and why did it appear? *Journalism Studies*, 4(4):501–511.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1118–1127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An em based training algorithm for cross-language text categorization. In *Proceedings of the Web Intelligence Conference*, pages 529–535.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In Jude W. Shavlik, editor, *Proceedings of the Third Internaional Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria.

Evan Sandhaus. 2008. New York Times corpus: Corpus overview. LDC catalogue entry LDC2008T19.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47, March.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: Evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 3063–3070, Valletta, Malta, may. European Language Resources Association (ELRA).

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2000a. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, pages 808–814, Morristown, NJ, USA. Association for Computational Linguistics.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. 2000b. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. September.

Elizabeth A. Thomson, Peter R. White, and Philip Kitley. 2008. objectivity and hard news reporting across cultures. *Journalism Studies*, 9(2):212–228.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 235–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682.

Revisiting sentence alignment algorithms for alignment visualization and evaluation

Qian Yu, Aurélien Max, François Yvon

LIMSI-CNRS and Univ. Paris Sud
rue John Von Neuman, F-91 403 Orsay
yu@limsi.fr, amax@limsi.fr, yvon@limsi.fr

Abstract

In this paper, we revisit the well-known problem of sentence alignment, in a context where the entire bitext has to be aligned and where alignment confidence measures have to be computed. Following much recent work, we study here a multi-pass approach: we first compute sure alignments that are used to train a discriminative model; then we use this model to fill in the gaps between sure links. Experimental results on several corpora show the effectiveness of this method as compared to alternative, state-of-the-art, proposals.

1. Introduction

The alignment of *bitexts*, *i.e.* of pairs of texts assumed to be mutual translations, consists in finding correspondences between logical units in parallel texts. The set of such correspondences is called an *alignment*. Depending on the logical units that are considered, various levels of granularity for the alignment are obtained. It is for usual to compute alignments at the level of paragraphs, sentences, phrases or words (see (Wu, 2010; Tiedemann, 2011) for two recent reviews). Alignments are widely used in many fields, especially in multilingual text processing (multilingual Information Retrieval, multilingual terminology extraction and Machine Translation). For all these applications, alignments between sentences must be computed.

Sentence alignment is generally considered an easy task and many sentence alignment algorithms have been proposed in the literature. From a bird's eye view, two main families of approaches can be isolated, which both rely on the assumption that the relative order of sentences is the same on the two sides of the bitext. On the one hand, *length-based approaches* (Gale and Church, 1991; Brown et al., 1991) use the fact that the translation of a short (resp. long) sentence is short (resp. long). On the other hand, *lexical matching approaches* (Kay and Röscheisen, 1993; Simard et al., 1993) identify sure anchor points for the alignment using bilingual dictionaries or surface similarities of word forms. Length-based approaches are fast but error-prone, while lexical matching approaches seem to deliver more reliable results. Most recent, state-of-the art approaches to the problem (Langlais, 1998; Simard and Plamondon, 1998; Moore, 2002; Braune and Fraser, 2010) try to combine both types of information.

In most applications, notably for training Machine Translation systems, only high-confidence, one-to-one, sentence alignments are kept. Indeed, when the objective is to build subsentential (phrase or word) alignments, the other types of mappings between sentences are deemed to be either insufficiently reliable or inappropriate. As it were, the one-to-one constraint is viewed as a proxy to literalness/compositionality of the translation, and warrants the search for finer-grained alignments. However, for certain types of bitexts, for instance literary texts, translation often departs from a straight sentence-by-sentence mode and using such a constraint discards a significant portion of the bitext. For Machine Translation, this is just a regrettable waste of potential training material. For other applications, however, notably applications which imply to visualize or read the actual translations in their context, as is, for instance, the case for second language learning, for training translators, or for automatic translation checking (Macklovitch, 1994), the entire bitext has to be aligned. Furthermore, areas where the translation is only partial or approximative may have to be identified precisely.

Following much recent work, we explore here a multiple-pass approach to sentence alignment. In a nutshell, our approach relies on sure one-to-one mappings detected in a first pass to train a discriminative sentence alignment system, which is then used to align the regions which remain problematic. Our experiments on the BAF corpus (Simard, 1998) show that this approach produces very high quality alignments, and also allows to identify the most problematic passages.

The rest of this paper is organized as follows: we first briefly review existing alignment methods in Section 2. In Section 3., we evaluate these methods and analyze the main sentence alignment errors. Our algorithm is detailed in Section 4., and evaluated on standard benchmarks in Section 5. We discuss further prospects and conclude in Section 6.

2. Sentence alignment: a review

Sentence alignment is an old task and the first proposals date back more than twenty years ago. These initial attempts can roughly be classified in two main categories: *length-based approaches* and *lexical matching approaches* (Tiedemann, 2011). The former family of approaches are based on the correlation of the length of parallel sentences, as introduced independently by Gale and Church (1991) and by Brown et al. (1991). The main intuition here is that long source sentences align preferably with long target sentences, and short source sentences with short target sentences. The difference between these two proposals is the way length is measured: the former study uses the number of characters, while the latter uses the number of words. The second family of approaches rely on sure or obvious

lexical correspondences, as provided, for instance, by entries of a bilingual dictionary, by so-called orthographical cognates[1] (Simard et al., 1993), or by word pairs having similar distributions of occurrence (Kay and Röscheisen, 1993). In both cases, additional simplifying assumptions are used, notably the fact that the relative order of sentences is preserved, and that sentences mostly align near the "diagonal" of the bitext, thus yielding very efficient algorithms.

Realizing the shortcomings of these initial proposals, several authors have proposed ways to combine the length-based approach and the lexical matching approach for aligning sentences (Chen, 1993; Wu, 1994; Moore, 2002; Braune and Fraser, 2010). For instance, the method proposed by Moore (2002) uses a three-step process for aligning sentences. First, a coarse alignment of the corpus is computed using a modified version of Brown et al.'s length-based model where search pruning techniques are used to speed up the discovery of reliable sentence pairs. In a second stage, the sentence pairs having the highest alignment probability are collected to train a modified version of IBM Translation Model 1 (Brown et al., 1993). Finally, the entire corpus is realigned using the IBM Model 1 score as an additional measure of parallelism. This method achieves high accuracy at a modest computational cost and does not require any knowledge of the languages or the corpus except how to break up the text into words and sentences. A very similar multi-pass approach is proposed in (Braune and Fraser, 2010), which basically aims at improving the unsatisfactory recall of Moore's algorithm, which misses many matchings when the bitext are not completely parallel.

Recent years have witnessed very few new proposals for this task and the problem seems to be basically solved. The only notable exceptions are the work of Deng et al. (2007), which tries to go beyond one-to-one sentence alignments, and considers matching large subparts using a divisive segmentation algorithm; the work of Fattah et al. (2007) using supervised learning tools; the robust aligner of Ma (2006), which relies on a statistical weighting scheme to balance the significance of bilingual lexical matches in parallel sentences; and the study of Sennrich and Volk (2010), which considers monolingual sentence alignment techniques after automatically projecting target texts back to the source language with machine translation.

3. A systematic analysis of alignment errors

3.1. Corpus and Baselines

In a first attempt to evaluate existing alignment methods, we selected a French literary work "De la terre à la Lune" by Jules Verne and its English translation "From the earth to the moon". This book is available as part of the BAF corpus (Simard, 1998). The French side of the bitext contains 3,319 sentences, 69,456 running words and 347,691 characters, whereas the English version contains 2,554 sentences, 50,331 words and 245,657 characters. Note the large difference in length between the French and the English side: as previously noted, the translation is only approximative, and it often appears that French paragraphs are summarized, rather than translated, into one or two English sentences. Both texts are shipped with reference sentence segmentations and alignment links.

To make our experimentations easier, we used the Uplug package[2], which provides a unified interface to integrate various sentence alignment methods. The distribution of Uplug ships with several alignment algorithms: the Gale-Church method[3], GMA[4] (Melamed, 1999), hunalign[5] (Varga et al., 2005), and some others. To these, we added the Moore aligner[6], the Gargantua alignment system[7] and BleuAlign[8]. All the input and output files are in the same format, which makes experimentation and inter-system comparison much easier.

Results are given in Table 1, where we display recall, precision and F-measure computed *at the alignment and at the sentence level*[9]. Note that with the latter metric, errors on $0-n$ or $n-0$ alignments are not taken into account. This might be because[10] it is generally considered unimportant to miss such alignments, which are not useful in the perspective of building parallel training material for Machine Translation. As reported in this table, some methods have very good precision, while recall is on average less satisfactory; the most extreme case is Moore's method, which achieves a nearly perfect precision, at the expense of a much worse recall.

3.2. Error analysis

As previously noted by several authors, this corpus is difficult because of the relatively low proportion of 1-to-1 links. This may be due to the use of non-literal translations or to differences in sentence segmentation. As detailed in Table 2, most methods are unable to reproduce the reference link distribution. The main issue is with null links, which, in this corpus, account for approximately 8.6% of the alignments. Only GMA is getting close to the right distribution, at the price, though, of a precision less satisfactory than for other approaches. It should be noted that making errors on such links often cause the desynchronization of entire passages, which has a strong negative impact on performance.

[1] Cognates are words that share a similar spelling in two or more different languages, as a result of their similar meaning and/or common etymological origin, e.g. (English-Spanish): history - historia, harmonious - armonioso. In subsequent references, they are more loosely defined as two words in different languages sharing a common prefix.

[2] http://sourceforge.net/projects/uplug/

[3] Using the implementation of Michael D. Riley.

[4] http://nlp.cs.nyu.edu/GMA/

[5] ftp://ftp.mokk.bme.hu/Hunglish/src/hunalign

[6] http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[7] http://sourceforge.net/projects/gargantua/

[8] https://github.com/rsennrich/bleualign/

[9] Other useful metrics for sentence alignment are based on recall and precision computed at the level of words and characters (see e.g. (Véronis and Langlais, 2000)).

[10] P. Langlais, personal communication.

	Gale	GMA	Hunalign	Moore	Gargantua
Alignment based metrics					
precision	0.30	0.61	0.50	0.85	0.74
recall	0.29	0.65	0.59	0.65	0.71
F-measure	0.29	0.63	0.54	0.74	0.72
Sentence based metrics					
precision	0.34	0.75	0.74	0.98	0.88
recall	0.39	0.77	0.69	0.62	0.77
F-measure	0.36	0.76	0.71	0.76	0.82

Table 1: Performance of various sentence alignment algorithms

Link type	0–1:5	1:5–0	1–1	1–2:5	2–1	2–2:5	others
Reference	0.56	8.05	75.71	4.37	4.60	3.65	3.06
Gale	0	0.41	59.22	3.51	34.63	2.23	0
GMA	0.74	10.54	68.42	4.43	13.02	1.00	1.85
Hunalign	0.20	1.41	61.02	3.93	33.44	0	0
Moore	0	0	100	0	0	0	0
Gargantua	0	0	91.64	3.97	3.85	0	0.54

Table 2: Distribution of predicted alignment types
Column 0-1:5 *gathers all the alignments matching* 0 *source sentence with* $1 \leq n \leq 5$ *target sentences.*

4. A coarse-to-fine approach to sentence alignment

4.1. Overview

We introduce in this section our coarse-to-fine alignment strategy. As with most multi-pass approaches, the first step is meant to provide a computationally cheap way to drastically reduce the alignment search space, by providing us with a first set of very high precision alignment links. All of these sentences that are aligned during this step are used as anchor points for the second step; they are also used to train a classifier aimed at recognizing parallel groups of sentences. The second step of our method uses an exhaustive search to enumerate and evaluate all the possible ways to align the blocks that appear between two anchor points. Based on the previous analysis, these blocks are typically sufficiently small that an exhaustive search is actually feasible. Based on these evaluations, a greedy algorithm is finally used to select the sentence pairs that align with highest probabilities.

For the first step, we simply chose to use the method of Moore (2002) because of its excellent precision. A tighter integration between this first step and the subsequent computations, which require to recompute several statistics that are used in Moore's approach, is certainly desirable. Yet, at this stage, we favored simplicity over computational efficiency. The two other steps are detailed below.

4.2. Detecting parallelism

The second step of our approach consists in training a function for scoring candidate alignments. Following (Munteanu and Marcu, 2005), we used a Maximum Entropy model[11] (Rathnaparkhi, 1998); in principle, many other choices would be possible here. We take the sentence alignments of the first step as positive examples; for negative examples, we randomly chose pairs $(\mathbf{e}, \mathbf{f}')$, where $(\mathbf{e}, \mathbf{f})$ and $(\mathbf{e}', \mathbf{f}')$ are two positive instances and $\mathbf{e}'$ directly follows $\mathbf{e}$. This strategy produced a balanced corpus containing as many negative pairs as positive ones. However, this approach may give too much weight on the length ratio feature and it remains to be seen whether alternative approaches are more suitable.

Our problem is thus to estimate a conditional model for deciding whether two sentences $\mathbf{e}$ and $\mathbf{f}$ should be aligned. Denoting Y the corresponding binary variable, this model has the following form:

$$P(Y = 1|\mathbf{e}, \mathbf{f}) = \frac{1}{1 + exp[-\sum_{i=1}^{k} \theta_k F_k(\mathbf{e}, \mathbf{f})]},$$

where $\{F_k(\mathbf{e}, \mathbf{f}), k = 1 \ldots K\}$ denotes a set of feature functions testing arbitrary properties of $\mathbf{e}$ and $\mathbf{f}$ and $\{\theta_k, k = 1 \ldots K\}$ is the corresponding set of parameter values.

Given a set of training sentence pairs, the optimal values of the parameters are set by optimizing numerically the conditional likelihood; optimization is performed here using L-BFGS (Liu and Nocedal, 1989); a Gaussian prior over the parameters is used to ensure numerical stability of the optimization. In practice, this means that the objective function is the inverse of the conditional log-likelihood, completed with a quadradic term proportional to $\sum_{i=1}^{k} \theta^2$.

In this study, we used the following set of feature functions:

- **lexical features**: for each pair of words[12] (e, f) occurring in $V_e \times V_f$, there is a corresponding feature $F_{e,f}$ which fires whenever $e \in \mathbf{e}$ and $f \in \mathbf{f}$.
- **length features**: denoting $l_{\mathbf{e}}$ (resp. $l_{\mathbf{f}}$) the length of the source (resp. target) sentence, measured in num-

[11] We use the implementation available from http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[12] A word is an alphabetic string of characters, excluding punction marks.

ber of characters, we include features related to length ratio, defined as $F_r(\mathbf{e}, \mathbf{f}) = \frac{|l_\mathbf{e} - l_\mathbf{f}|}{max(l_\mathbf{e}, l_\mathbf{f})}$. Rather than taking the numerical value, we use a simple discretization scheme based on 6 bins.

4.3. Filling alignment gaps

The third step uses the posterior alignment probabilities computed in the second step to fill the gaps in the first pass alignments. The algorithm can be glossed as follows. Assume a bitext block comprising the sentences from index i to j in the source side of the bitext, and from k to l in the target side such that sentences $\mathbf{e}_{i-1}$ (resp. $\mathbf{e}_{j+1}$) and $\mathbf{f}_{k-1}$ (resp. $\mathbf{e}_{l+1}$) are aligned[13].

The first case is when $j < i$ or $k > l$, in which case we create a null alignment for $f_{k:l}$ or for $e_{i:j}$. In all other situations, we compute:

$$\forall i', j', k', l', i \leq i' \leq j' \leq j, k \leq k' \leq l' \leq l,$$
$$a_{i',j',k',l'} = P(Y = 1 | \mathbf{e}_{i':j'}, \mathbf{f}_{k':l'}),$$

where $\mathbf{e}_{i':j'}$ is obtained by concatenation of all the sentences in the range $i' : j'$. Note that this implies to compute $O(|j - i|^2 \times |k - l|^2)$ probabilities, which, given the typical size of these blocks (see below), can be performed very quickly.

These values are then iteratively visited by decreasing order in a greedy fashion. The top-scoring block $i' : j', k' : l'$ of the list is retained in the final alignment; all blocks that overlap with this block are deleted from the list and the next best entry is then considered. This process continues until all remaining blocks imply null alignments, in which case these $n - 0$ or $0 - n$ alignments are also included in our solution.

This process is illustrated on Figure 1: assuming that the best matching link is f_2-e_2, we delete all the links that include f_2 or e_2, as well as links that would imply a reordering of sentences, meaning that we also delete links such as f_1-e_3 etc.

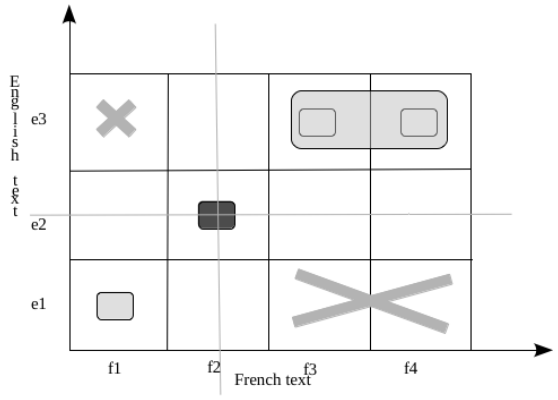


Figure 1: Greedy alignment search

[13]We conventionally enclose the source and target texts between begin and end markers so as to ensure that the first and last sentences are aligned.

5. Experiments

5.1. Results on literary work

In this first round of experiments, we consider the same literary work as in section 3.

The first alignment step, using Moore's algorithm with default parameters, identifies 1936 one-to-one alignments, used as anchor points for the remaining steps of the procedure. These are high-quality alignments which only contain 35 errors. Nonetheless, this first step creates a small number of misalignments: these errors can not be fixed, introduce some noise in the training set of the classifier and will also create more alignment errors in the subsequent steps. Using these anchor points, 447 "paragraphs" need to be further processed, corresponding to 1,383 French and 618 English sentences: the average length of these paragraphs is then respectively 2.9 sentences for the French side and 1.3 sentences for the English side, which makes our search procedure for fine-grained alignments computationally tractable. Note that not all these paragraphs need to be processed: in fact, for 156 of them, the only possible decision is to align 0-to-n or n-to-0.

In order to assess the quality of the Maxent classifier, we split the available training data into 90% for estimating the parameters and 10% for testing, and found that its decisions were correct 75% of the time[14].

A contrast was run using a much larger corpus of parallel sentences extracted from a collection of literary bitext. Here, the total number of training sentences was 133,562. Increasing the number of training sentences increased the precision of the classifier from 75% to 81%.

The third step was to fill the alignment gap using the algorithm presented in previous section. Here again, two strategies were tested: the baseline approach is a faithful implementation of our approach; alternatively, we tried to discard all the alignment links whose probability (for the Maxent model) is less than 0.5. In this condition, the number of null alignments is significantly increased.

The results of our experiments are summarized in Table 3. As reflected by these results, our multi-pass strategy delivers alignment results that significantly improve over the state-of-the-art. Unsurprisingly, we were able to boost the initial recall of Moore's method at the cost of only a small lost in precision. The F-measure is better than all the other alignment techniques, slightly surpassing the recent proposal of Braune and Fraser (2010). Using a larger training corpus has a small effect on the precision of the Maxent classifier, which does not show on the global alignment performance: our classifier is arguably delivering better performance, but its feature weights are less adapted to the specificities of our data. Likewise, using a prefiltering stage has hardly any impact on the global quality of our results; yet, this filtering is useful for speeding up our algorithm as it enables to discard 93% of the potential alignment links.

Looking at errors by alignment types (Table 4), we see that our method is able to better reproduce the distribution of link types, even though 0-to-n links still account for a substantial number of errors.

A qualitative analysis of alignment errors showed that:

[14]These results were obtained using 10-fold cross validation.

Score	Moore+Maxent	Moore+Maxent (+corpus)	Moore+Maxent (filtering_0.5)
Alignment based metrics			
precision	0.74	0.74	0.72
recall	0.81	0.80	0.80
F-measure	0.77	0.77	0.76
Sentence based metrics			
precision	0.93	0.90	0.94
recall	0.80	0.80	0.78
F-measure	0.86	0.85	0.85

Table 3: Performance at the alignment level and sentence level

	Link type						
	0–1:5	1:5–0	1–1	1–2:5	2–1	2–2:5	others
Reference	0.56	8.05	75.71	4.37	4.60	3.65	3.06
Moore+Maxent	2.64	10.73	79.88	2.13	2.46	0.44	1.72

Table 4: Distribution of predicted link types

- modeling null alignments remains difficult, as these links are only produced as a fall-back decision, for lack of finding better alignments. As a result, these alignments continue to account for a large number of errors.
- the model we train to predict alignment probability is a "bag-of-words" model and is only concerned with the cooccurrence of words in the French and English side, no matter how often these words occur. As a result, two adjacent sentences using the same vocabulary tend to confuse our aligners. This also occurs when adjacent sentences contain word pairs that were not seen in training and which play no role in scoring the alignments: the system is then unable to choose between segmenting a block of sentences or keeping them as a group (see examples in Figure 2).

A last question concerns the use of the model's scores as confidence estimation measures for the alignment. To check this, we removed from the final alignment all the blocks whose score is below a given threshold $0 \leq \theta \leq 1$ for varying values of θ; by convention, we assume that Moore's alignment links are sure and are never discarded. As expected, increasing θ from 0 (no filtering) to 1 (filter all but Moore's blocks) increases the precision, but is detrimental to recall. A slightly better F-measure of 0.78 can be obtained for $\theta = 0.4$; the variations are however small and remain to be confirmed for larger scale studies.

5.2. Complementary results on the BAF

In this section, we report on experiments conducted with other documents contained in the BAF corpus. Our goal here is to check that our method, which performs quite well on a "difficult" text, is also able to handle the easier types, such as institutional texts or scientific articles[15]. Our results are summarized in Table 5, where we compare our approach with its main competitors and show that it attains state-of-the-art results on these collections as reflected by the comparison with the Gargantua software.

[15] As is standard pratice, we have not tried to align the technical manuals, which pose specific and difficult alignment problems.

	Moore	Gargantua	Moore+Maxent
Institutional texts			
Alignment based metrics			
precision	0.97	0.96	0.93
recall	0.91	0.96	0.95
F-measure	0.94	0.96	0.94
Sentence based metrics			
precision	0.99	0.98	0.98
recall	0.84	0.93	0.93
F-measure	0.91	0.95	0.95
Scientific articles			
Alignment based metrics			
precision	0.89	0.86	0.85
recall	0.89	0.91	0.93
F-measure	0.89	0.88	0.89
Sentence based metrics			
precision	1.00	0.98	0.95
recall	0.72	0.77	0.81
F-measure	0.84	0.86	0.87

Table 5: Performance at the alignment level and sentence level on other parts of the BAF corpus

6. Conclusions

In this paper, we have presented a novel two-pass approach aimed at improving existing sentence alignment methods in contexts where (i) all sentences need to be aligned and/or (ii) sentence alignment confidence need to be computed. By running experiments with several variants of this approach, we have been able to show that it was slightly better than the state-of-the-art on aligning a novel with its translation, and equivalent to the best approaches on other benchmarks. These results will be complemented by our ongoing experiments with the other benchmarks of Arcade 2 (Chiao et al., 2006) and with other literary corpora.

(src) =”1.2065”	un second tiers voyait mal et n’ entendait pas ;
(src) =”1.2066”	quant au troisième , il ne voyait rien et n’ entendait pas davantage .
(trg) =”1.1555”	a second set saw badly and heard nothing at all ;
(trg) =”1.1556”	and as for the third , it could neither see nor hear anything at all .
(src)=”1.2013”	bonjour , Barbicane .
(src)=”1.2014”	Comment cela va-t-il ?
(trg)=”1.1508”	how d ’ye do , Barbicane ?
(trg)=”1.1509”	how are you getting on ?

Figure 2: Alignment errors. In both cases, two consecutive sentences use similar words, which makes the block alignment look better than a split.

This approach can be improved in many ways: an obvious extension will be to add more features, such as cognates, Part-of-Speech, lemmas, or alignment features as was done in (Munteanu and Marcu, 2005). We plan to provide a much tighter integration with Moore’s algorithm, which already computes such alignments, so as to avoid having to recompute them. Finally, the greedy approach to link selection can easily be replaced with an exact search based on dynamic programming techniques, including dependencies with the left and right alignment links.

Acknowledgments

This work has been partly funded through the “Google Digital Humanities Award” program.

7. References

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89, Beijing, China. Coling 2010 Organizing Committee.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1991, Berkeley, California*, pages 169–176.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Stanley F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL ’93, pages 9–16.

Yun-Chuang Chiao, Olivier Kraif, Dominique Laurent, Thi Minh Huyen Nguyen, Nasredine Semmar, François Stuck, Jean Véronis, and Wajdi Zaghouani. 2006. Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of the 5th international Conference on Language Resources and Evaluation - LREC’06*, Genoa, Italy.

Yonggang Deng, Shankar Kumar, and William Byrne. 2007. Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(03):235–260.

Mohamed Abdel Fattah, David B. Bracewell, Fuji Ren, and Shingo Kuroiwa. 2007. Sentence alignment using P-NNT and GMM. *Computer Speech and Language*, 21:594–608, October.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184, Berkeley, California.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignement. *Computational Linguistics*, 19(1):121–142.

Philippe Langlais. 1998. A System to Align Complex Bilingual Corpora. Technical report, CTT, KTH, Stockholm, Sweden, Sept.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *In Proceedings of LREC-2006*, Genoa, Italy.

Elliot Macklovitch. 1994. Using bi-textual alignment for translation validation: the TransCheck system. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 157–168, Columbia.

I. Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25:107–130.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In Stephen D. Richardson, editor, *Proc. AMTA’02*, Lecture Notes in Computer Science 2499, pages 135–144, Tiburon, CA, USA. Springer Verlag.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Ardwait Rathnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, November.

Michel Simard and Pierre Plamondon. 1998. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80.

Michel Simard, George F. Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora.

In Ann Gawman, Evelyn Kidd, and Per-Åke Larson, editors, *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research, October 24-28, 1993, Toronto, Ontario, Canada, 2 Volume*, pages 1071–1082.

Michel Simard. 1998. The BAF: a corpus of English-French bitext. In *First International Conference on Language Resources and Evaluation*, volume 1, pages 489–494, Grenada, Spain.

Jörg Tiedemann. 2011. *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed). Morgan & Claypool Publishers.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596, Borovets, Bulgaria.

Jean Véronis and Philippe Langlais. 2000. Evaluation of Parallel Text Alignment Systems. In Jean Véronis, editor, *Parallel Text Processing*, Text Speech and Language Technology Series, chapter X, pages 369–388. Kluwer Academic Publishers.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994, Las Cruces, New Mexico Canada*, pages 80–87.

Dekai Wu. 2010. Alignment. In Nitin Indurkhya and Fred Damerau, editors, *CRC Handbook of Natural Language Processing*, number 16, pages 367–408. CRC Press.

Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation

Inguna Skadiņa
Tilde
Vienības gatve 75a, Riga, LATVIA
E-mail: inguna.skadina@tilde.lv

Abstract

This abstract presents the FP7 project ACCURAT that aims to research methods and create tools that find, measure, and use bi/multilingual comparable corpora to improve the quality of machine translation for under-resourced languages and narrow domains. Work on corpora collection, assessment of the comparability of documents pairs in collected corpora, extraction of parallel data for the machine translation (MT) task, and application to the MT task is presented.

Keywords: comparable corpora, under-resourced languages, comparability metric, information extraction, machine translation

1. Introduction

The applicability of current data-driven methods directly depends on the availability of very large quantities of parallel corpus data. For this reason, the translation quality of data-driven MT systems varies dramatically from being quite good for language pairs and domains with large corpora available to being almost unusable for under-resourced languages and domains.

The ACCURAT project (*Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation*) addresses this issue by developing technology for using comparable corpora as resources for machine translation systems (Skadiņa et al., 2010a; Eisele and Xu, 2010). The project aims to research methodology and tools that measure, find, and use comparable corpora to improve the quality of MT for under-resourced languages and narrow domains (e.g., renewable energy and topical news).

The objectives of the ACCURAT project are to:

- Research methods for automatic acquisition of a comparable corpus from the Web which can be used as a source to extract data for MT;
- Create comparability metrics – to develop the methodology and determine criteria to measure the similarity of source and target language documents in comparable corpora;
- Research methods and develop tools for the alignment and extraction of lexical, terminological, and other linguistic data from comparable corpora;
- Measure improvements achieved by applying acquired data against baseline results from statistical and rule-based machine translation systems.

The ACCURAT project particularly targets a number of under-resourced languages: Croatian, Estonian, Greek, Latvian, Lithuanian, and Romanian.

This abstract provides an overview of research results and the tools developed within the project to achieve the above mentioned objectives (more details can be found in Skadiņa et al., 2012 and papers of consortium partners listed in References).

2. Methods for building a comparable corpus from the Web

Several novel approaches how to build a comparable corpus from the Web that are applicable to under-resourced languages have been researched. Tools for the identification of comparable documents in Wikipedia, news documents, and narrow domains have been developed.

For **news texts,** a two-stage method that first gathers documents monolingually and then pairs them across languages to build a comparable corpus has been developed (Aker et al., 2012). In the gathering stage, news texts are downloaded separately in each project language at regular intervals from Google News. The titles are used in further queries for gathering more related articles from Google News. To overcome the relative scarcity of news in non-English languages, titles from the English news articles are parsed for named entities which are then translated into the non-English language and serve as queries for gathering related news texts. Selected RSS feeds from under-resourced languages are also used for the same reasons. Documents are paired across monolingual collections by using a number of features (e.g., date and time of publication and similarity of title length and title content).

For **Wikipedia texts,** we developed a technique to find comparable Wikipedia texts based on the idea that inter-lingually linked Wikipedia text pairs containing significant numbers of shared anchor texts are likely to be quite similar in content (Paramita et al., 2012).

For **narrow domain texts,** a topic definition (specified as a list of topic terms) and a seed URL list are given to a focused monolingual crawler (FMC) that crawls starting from the seed URLs and performs lightweight text classification on pages it encounters to determine if they are relevant to the domain. For a specific topic, all of the returned texts in one language may be paired with all of the texts from another language to form a comparable corpus.

The above described tools are used to gather very large collections of comparable documents for all project language pairs. For news texts comparable corpora for 8 language pairs are collected, with the number of

document pairs ranging from 16,144 to 129,341. The "wikipedia-anchors" method contains corpora in 12 language pairs, with the number of document pairs ranging from 841 to 149,891. The FMC tool is used to collect narrow domain comparable corpora from the Web: 28 comparable corpora in 8 narrow domains for 6 language pairs have been constructed and amount to a total of more than 148M tokens.

3. Criteria of comparability and comparability metrics

In ACCURAT comparability is defined by how useful a pair of documents is for machine translation. Within the project, two different metrics are implemented to identify comparable documents from raw corpora crawled from the Web and to characterise the degree of their similarity (Su and Babych, 2012).

The machine translation based metric first uses the available machine translation API's for document translation and incorporates several useful features into the metric design. These features, including lexical information, keywords, document structure, and named entities, are then combined in an ensemble manner.

The lexical mapping based metric uses automatically generated GIZA++ bilingual dictionaries for lexical mapping. If a word in the source language occurs in the bilingual dictionary, the top 2 translation candidates are retrieved as possible translations in the target language. This metric provides a much faster lexical translation process, although word-for-word lexical mapping results are not as good as automatic translations.

The reliability of the proposed metrics has been tested on semi-manually collected Initial Comparable Corpora (Skadiņa et al., 2010b) used as a gold standard. It turned out that the comparability scores obtained from the comparability metrics reliably reflect comparability levels, as the average scores for higher comparable levels are always significantly larger than those of lower comparable levels. However, for the lexical mapping based metric, the average score for each comparability level drops in comparison to that of the MT based metric. The applicability of the proposed metrics was also measured by its impact on the task of parallel phrase extraction from comparable documents. The results show that a higher comparability level always leads to a significantly higher number of aligned phrases extracted from the comparable documents.

4. Alignment methods and information extraction from comparable corpora

In the ACCURAT project, the term alignment is used in the context of machine translation to describe the pairing of text in one document with its translation in another.

Through studies of existing alignment strategies designed for parallel corpora, comparable corpora, and non-comparable corpora, we showed that the most widely used alignment methods (Giza++ and Moses) are not well suited for use directly on strongly and weakly comparable texts. Therefore, the project consortium proposed new methods and implemented tools that allow the alignment of comparable documents and the extraction of information (paragraphs, phrases, terminology, and named entities) from comparable corpora. All of the important tools that have been developed within the ACCURAT project for the alignment of comparable corpora at different levels and for data extraction from comparable corpora that are useful for machine translation are packed into the ACCURAT Toolkit (ACCURAT D2.6, 2011)[1]. By using the ACCURAT Toolkit, users may expect to obtain:

- **Comparable document (and other textual unit types) alignment**. This will facilitate the task of parallel phrase extraction by massively reducing the search space of such algorithms;
- **Parallel sentence/phrase mapping** from comparable corpora (Ion, 2012). This aims to supply clean parallel data useful for statistical translation model learning;
- **Translation dictionaries** extracted from comparable corpora. These dictionaries are expected to supplement existing translation lexicons which are useful for both statistical and rule-based MT;
- **Translated terminology** extracted (mapped) from comparable corpora (Ştefănescu, 2012). This type of data is presented in a dictionary-like format and is expected to improve domain-dependent translation;
- **Translated named entities** extracted (mapped) from comparable corpora. Also presented in a dictionary-like format, these lexicons are expected to improve parallel phrase extraction algorithms from comparable corpora and be useful by themselves when actually used in translation.

In order to map terms and named entities bilingually, the ACCURAT Toolkit also provides tools for detecting and annotating these types of expressions in a monolingual fashion.

The tools can be applied individually or in the provided workflows: (1) for parallel data mining from comparable corpora and (2) for named entity/terminology extraction and mapping from comparable corpora.

5. Comparable corpora in MT systems

With the ACCURAT toolkit, the consortium is aligning comparable corpora collected from the Web at the document level and extracting MT-related data – parallel phrases/sentences and bilingual lists of named entities and terminology. To evaluate the efficiency and usability of the developed methods for under-resourced languages and narrow domains, data extracted using these methods is being integrated into ACCURAT baseline MT systems. ACCURAT baseline MT systems are built for 17 translation routes using existing SMT techniques on available parallel corpora, e.g., JRC-ACQUIS

[1] http://www.accurat-project.eu/index.php?p=toolkit

Multilingual Parallel corpus and SETimes corpus are available on MT-Serverland software infrastructure[2] and via the Web service.

For narrow domain MT, several successful proof-of-concept experiments were carried out to show that even small amounts of parallel domain specific data will help improve a SMT system.

To test the quality and effect of the data extracted with ACCURAT tools, an experiment with English-German domain-adapted SMT was performed for the automotive industry domain (Ştefănescu et al., 2012). By adding 45,952 sentence pairs extracted from the automotive domain comparable corpus, approximately 6.5 BLEU points over the baseline system were obtained.

The language model adaptation experiment was applied to the renewable energy domain. This led to improvements in terms of BLEU score for the following language pairs: for English->Greek BLEU increased from 15.07 to 15.14, for English->Lithuanian from 18.23 to 23.38, and for English->Croatian from 11.93 to 14.94.

More experiments are in progress, as corpora collection was finished recently and data extraction is still in progress.

6. Conclusion

The project is in its final phase now. The following key methods and tools are developed: crawling methods to identify comparable documents on the Web; comparability metric allowing identify comparable documents and evaluate their similarity; methods for automatic extraction of parallel and quasi-parallel data from any degree of comparable corpora.

7. Acknowledgements

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

8. References

ACCURAT D2.6 (2011) Toolkit for multi-level alignment and information extraction from comparable corpora., 31st August 2011 (http://www.accurat-project.eu/), 123 pages.

Aker, A.; Kanoulas, E. and Gaizauskas, R. (2012) A light way to collect comparable corpora from the Web. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.

Eisele, A. and Xu, J. (2010). Improving machine translation performance using comparable corpora. In: *Proceedings of 3rd Workshop on Building and Using Comparable Corpora* , Malta, pp. 35--41.

Ion, R. (2012) PEXACC: A Parallel Data Mining Algorithm from Comparable Corpora. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.

Skadiņa, I.; Vasiļjevs, A.; Skadiņš, R.; Gaizauskas, R.; Tufiş; D, Gornostay, T. (2010a). Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, European Language Resources Association (ELRA), La Valletta, Malta, May 2010, pp. 6--14.

Skadiņa, I.; Aker, A.; Giouli, V.; Tufis, D.; Gaizauskas, R.; Mieriņa M. and Mastropavlos, N. A. (2010b). Collection of Comparable Corpora for Under-resourced Languages. In *Proceedings of the Fourth International Conference Baltic HLT 2010*, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 219, pp. 161--168.

Skadiņa, I.; Aker, A.; Mastropavlos, N.; Su, F.; Tufis, D.; Verlic, M.; Vasiļjevs, A.; Babych, B.; Clough, P.; Gaizauskas, R.; Glaros, N.; Paramita, M.; Pinnis, M. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.

Ştefănescu, D. (2012). Mining for Term Translations in Comparable Corpora. In *Proceedings of BUCC 2012*, May, 26, Istanbul, Turkey.

Ştefănescu, D.; Ion, R., and Hunsicker, S. (2012).Hybrid Parallel Sentence Mining from Comparable Corpora. In *Proceedings of EAMT 2012*.

[2] http://www.dfki.de/mt-serverland/

LetsMT! - Platform to Drive Development and Application of Statistical Machine Translation

Andrejs Vasiļjevs
Tilde
Vienibas gatve 75a, Riga, LV2101, LATVIA
E-mail: andrejs@tilde.lv

Abstract

This paper presents ICT-PSP project LetsMT! which develops a user-driven machine translation "factory on the cloud". Current mass-market and online MT systems are of general nature, system adaptation for specific needs is prohibitively expensive service not affordable to smaller companies or public institutions. To exploit the huge potential of open statistical machine translation (SMT) technologies LetsMT! has created an innovative online collaborative platform for data sharing and MT building.

Keywords: machine translation, cloud, SMT, parallel corpora, data processing

1. Introduction

In recent years, statistical machine translation (SMT) has become the leading paradigm for machine translation. The quality of SMT systems largely depends on the size of training data. Since the majority of parallel data is in major languages, SMT systems for larger languages are of much better quality compared to systems for smaller languages. This quality gap is further deepened due to the complex linguistic structure of many smaller languages. Languages like Latvian, Lithuanian and Croatian (to name just a few) have a complex morphological structure and free word order.

Another significant challenge is to break down the access barriers to SMT technology by making this platform and process user-friendly. Currently the implementation of SMT solutions, whether proprietary or out-sourced, requires an intensive investment of resources: human (natural language processing experts, system administrators), financial, and linguistic, to create and maintain a custom SMT infrastructure (Varga et al. 2005).

2. Project objectives

LetsMT! is a collaborative platform that thrives on resources contributed by its users. It contributes in a breakthrough regarding the availability of parallel language resources and, consequently, MT services of good and acceptable quality for less-covered languages where the current MT systems perform poorly due to limited availability of training data.

LetsMT! provides a platform that supports the following features:

- Uploading of parallel texts for users that will contribute their own content;
- Automated training of SMT systems from specified collections of training data;
- Custom building of MT engines from a selected pool of training data, for larger donors or paying customers;
- Custom building of MT engines from proprietary non-public data, for paying customers;
- MT evaluation facilities.

LetsMT! platform results from a project with duration of 30 months. It was started on March 1st 2010 and is planned to end by August, 2012. The project consortium consists of 6 partners – Tilde SIA, University of Edinburgh, University of Zagreb, University of Copenhagen, Uppsala University, Zoorobotics BV, Moravia. The project is coordinated by Tilde.

The core objective of the project is to provide innovative online MT services through sharing of parallel corpora provided by users, with emphasis on less-covered languages and specialized domains.

The solution created in the project provides the following core functions:

- A website for uploading parallel corpora and building specific MT solutions;
- A website for translation, where source text can be typed and translated;
- A translation widget provided for inclusion into websites to translate their content;
- Browser plug-ins that will provide the quickest access to translation;
- Integration in CAT tools and other applications.

3. Platform and infrastructure

Work on the LetsMT! platform and infrastructure is the core activity within the project. The LetsMT! platform includes modules for sharing of SMT training data, SMT training and running, use in a news translation scenario, and use in a localisation usage scenario.

The beta versions of all the main modules is completed and deployed. The project Consortium has developed a common platform and supporting software infrastructure that provides the core functions necessary to integrate the modules of the LetsMT! platform. The supporting software infrastructure includes: the LetsMT! website, an API for external systems, User Management and Access Rights Control, Application Logic, an MT web page where users can try trained MT systems, etc.

It is obvious that hosting the LetsMT! platform requires a lot of computing capacity. The Project Consortium,

instead of buying servers, intends to lease capacity. It is economically efficient and will provide flexibility in adding new resources as necessary. During the analysis of detailed requirements, it was discovered that operating the LetsMT! platform on AWS (Amazon Web Services) was the most economically efficient option. It is planned to deploy the LetsMT! platform completely within the AWS, as this is a well-established solution. The AWS cloud provides a reliable and scalable infrastructure for deploying web-scale solutions. Alternative cloud computing suppliers may be selected if AWS fails to meet the requirements of the LetsMT! platform. The LetsMT! platform also can be deployed on a local server infrastructure.

4. SMT resource repository

The backend of the LetsMT! platform includes a modular resource repository. Figure 1 illustrates the general architecture of the software. Its design emphasizes possibilities of running the system in a distributed environment which makes the system suitable for scalable cloud-based solutions.

Communication between the web-frontend and the individual modules is handled by secure web service connections. A central database handles metadata information in a flexible key-value store that supports schema-less expandable information collections. The physical data storage can be distributed over several servers to reduce bottlenecks when transferring large data collections. Data collections can be stored using a version-controlled file system that supports data recovery and history management in a multi-user environment. The repository provides essential features for importing documents to the LetsMT! platform. Documents are converted, and sentences in translated documents are aligned automatically. The software is connected to a high-performance cluster that can execute various jobs with connection to the data stored in the repository, for example, the import and alignment of jobs. A cloud-based cluster enables scalability of the system according to the needs of the platform. The repository software is fully integrated in the current LetsMT! platform and can easily be extended with additional modules.

5. Collecting the training data

A large amount of training data is crucial for statistical machine translation.

The aim of the LetsMT! project is to collect data from both general language and from different subject domains. A special effort is being made by two of the project partners to collect business and finance news and localisation texts, mostly from the IT domain. Other subject domains are interesting for the project, so the partners focus on finding text providers with general language texts, in addition to domain specific texts.

The initial training corpora focused on Croatian, Czech, Danish, Dutch, Latvian, Lithuanian, Polish, Slovak, and Swedish. We still focus on these original languages, though other languages are also collected as part of multilingual corpora (Tiedemann 2009, Steinberger et al. 2006, Koehn 2005).

During the first year, lots of publicly available data was collected and provided on LetsMT! repository. Now Project Consortium concentrates on identifying new text providers and potential future users of the LetsMT! system.

For business and finance news, the Project Consortium uses a list of the largest companies from the involved countries to automatically harvest the newest parallel texts from these companies, and therefore, the collection is steadily growing.

The collection of parallel texts from the general language and from other subject domains is being advanced by making contacts at different levels. At the international level, the Project Consortium is in contact with TAUS (which has one of the largest repositories of parallel corpora) and with various EU institutions and projects, e.g., ACCURAT, TTC and META-NORD. At the national

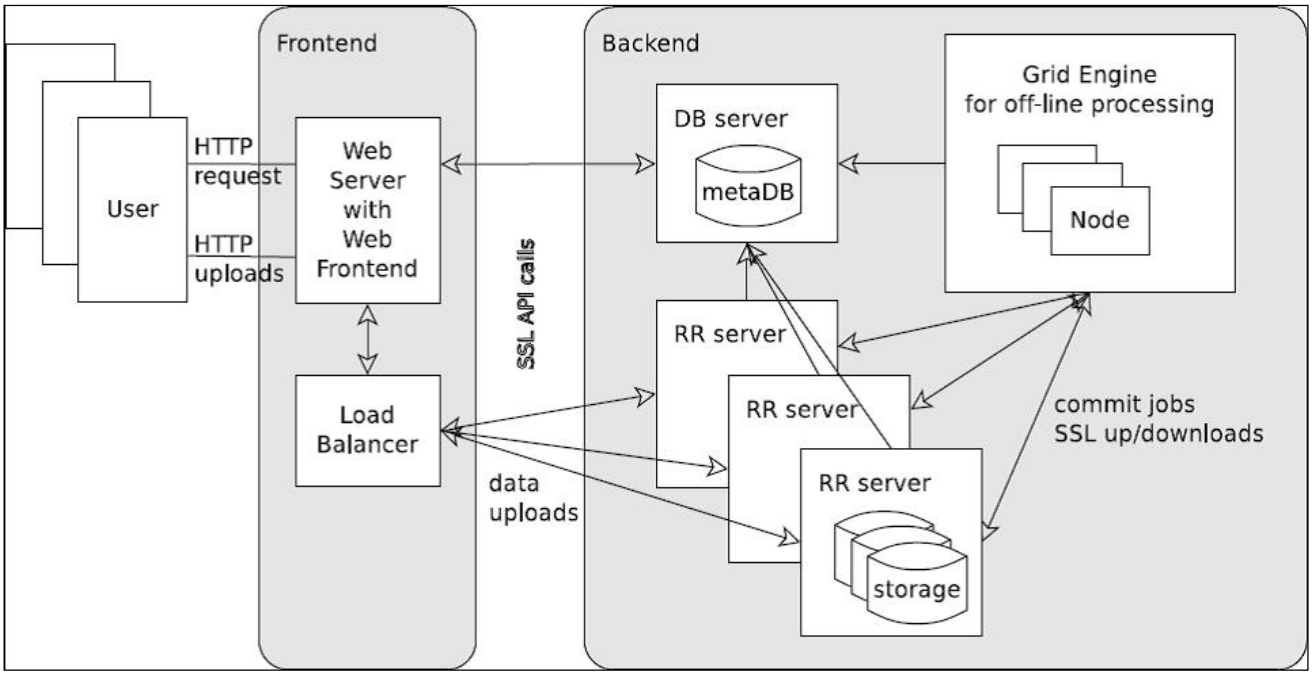


Figure 1: General architecture of a Resource Repository

level, project partner - Tilde has made a cooperation agreement with the National Library of Latvia. The partner University of Zagreb has made contact with several translation and localisation companies that are interested in the project and two of these have committed themselves to become text providers. The project partner-Moravia has made contact with the Slovak national corpus, but due to IPR problems their corpora cannot be used outside the institute. The project partner-Uppsala University contacted two institutes at Stockholm University who might be interested in using LetsMT!. The project partner University of Copenhagen contacted several potential text providers and has received acceptance from a company that write press releases in the EU languages and from at least 12 companies with annual reports. Furthermore, University of Copenhagen has started co-operation with a translation centre connected to University about texts in the domain of university administration.

The LetsMT! project has been presented at various events both at the national and international levels in order to spread knowledge and create awareness of the project in general and of the need for data in particular. Generally, the responses to the presentations are very positive, but IPR constitutes a challenge to the project. It turns out that some of the texts originally identified cannot be used outside the company/institution and thus cannot be uploaded to an external server like LetsMT!. Others can only be uploaded for private use and will not become publicly available on the LetsMT! platform. However, some of the contacts mentioned above are ready to sign a text provider agreement and others will follow soon.

6. SMT Training and running facilities

Users of the LetsMT! platform may select training resources from the SMT Resource Repository and train tailored SMT systems using the selected training resources. SMT training facilities include the following features: a user interface for resource selection and system training, integration with user authentication and access rights module, integration with SMT Resource Repository, simultaneous and effective execution of resources consuming training tasks, an interface providing information about running training tasks, progress, status, etc.

The SMT training facility and web service is built on top of the Moses machine translation toolkit (Koehn et al. 2007). Originally developed at the University of Edinburgh in 2007, Moses has since undergone a great deal of evolution. Many new features have been added, improving translation quality and keeping Moses up to date with the cutting edge of MT research. While of great importance, translation quality, however, is not the only aspect to have been worked on. SMT is extremely computationally demanding. Literally millions of options must be searched through in order to translate a single sentence, and the amount of data required to do so far outstrips the resources of and average desktop computer. Therefore, much research has been conducted on how to speed this process up and reduce the computational resources needed for translation.

Translation is only a part of what the Moses SMT Toolkit can do, though also included with it are the tools to train new translation systems. As with the actual method of translating, huge amounts of work have gone into training systems to yield better translations, as well as making the training process itself less resource intensive. The process of training a translation system is very in depth and intricate, but that too is handled by the toolkit.

Despite all the work that has gone into developing Moses, there are a few features required by the LetsMT! platform that Moses did not have. Having been conceived in academia, the focus of Moses has generally been towards features required by researchers and researchers. However, the environment in which it operates in the LetsMT! platform is very different. Developing Moses to support these new requirements is the main focuses of project activities and the work done in doing so is detailed below.

End users expect a service that delivers translations in a fast, interactive manner. Translating sentences requires a large amount of data, and waiting for this to be loaded each time would make the interactive user experience impossible to deliver. This has been addressed by the implementation of a version of Moses which runs on a background server, can be given sentences to translate interactively, and returns the translations quickly — without having to wait for the whole system to load up.

Users of the LetsMT! platform will also be translating between many different pairs of languages, and therefore, separate background processes for each pair would be impractical. This has been countered by allowing Moses to simultaneously have multiple translation systems in memory and by providing the language to translate into along with the sentence.

Modern computers are increasingly geared towards executing many processes in parallel, instead of doing them sequentially. In order to make the best use of available resources, Moses must be able to translate sentences in parallel. This feature, called 'multi-threading' has been integrated into Moses and enables it to deliver many translations in a fraction of the time compared to doing them one after another.

Other features such as being able to leverage new data without having to retrain the entire system have also been implemented and are in the process of being integrated with the rest of the platform. Methods for improving the fluency of translations using many billions of words of text are also in active development.

The LetsMT! platform is a great example of an EU project putting cutting edge technology to great use for the wider public, and as it does so, feeding back improvements to the academic community from where its ideas originate.

7. Usage scenarios

In particular, two specialized usage scenarios are supported by the LetsMT! platform: 1) machine translation of financial news, and 2) translation process in localization industry companies.

4.1 MT usage in news translation

Project Consortium has implemented the widget and browser (Mozilla, Internet Explorer) plug-ins of the LetsMT! platform.

The business scenario was developed in which the use of the widget is described. The aim for the business scenario is to provide translated business and financial information

through several facilities. There are two scenarios which are currently being investigated, a free and a paid, professional service. Free services will attract a broad audience of users with an interest in business related news and financial background information. The content will be information with a high latency, background information of local stock markets, local listed company information and comments. For low latency and emerging news, users can subscribe to a paid service. The targeted users are professionals and individuals that are interested in local and international breaking news and financial information. At the moment, the LetsMT! widget is integrated into SemLab's (Zoorobotics) business and financial news website www.newssentiment.eu for trial and evaluation purposes. The system is being tested on the website to ensure positive results in dissemination and exploitation activities through other (financial news) websites.

4.2 MT usage in localization

Professional users need MT services integrated in their working environment. Translators use CAT (Computer Aided Translation) tools (such as SDL Trados and MemoQ) in everyday activities. One of the prerequisites conditioning successful localisation scenario implementation is, without any doubt, the integration with CAT tools. In order to fulfil these requirements, the Project Consortium has developed a LetsMT! platform plug-in for SDL Trados Studio 2009 which allows for the use of the LetsMT! platform during translation process and experimentation on the evaluation of an English-Latvian SMT system applied to an actual localisation assignment (Vasiļjevs et al. 2011). The paper shows that such an integrated localisation environment can increase the productivity of localisation by 32.9% without critical reduction in quality.

8. Conclusion

Current development of the SMT tools and techniques has reached the level that they can be implemented in practical applications addressing the needs of large user groups in variety of application scenarios. The consortium partners are inviting Beta testers to evaluate the LetsMT! system and the current positive reviews on user experience indicate that the project is developing in a direction that is demanded by potential users.

Successful implementation of the project will democratize access to custom MT and, facilitate diversification of free MT by tailoring to specific domains and user requirements and have a strong impact on reducing the digital information divide in the EU.

9. Acknowledgements

The research within the LetsMT! project leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web, grant agreement no 250456. The research within the project ACCURAT has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

10. References

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand

Koehn, P., Federico, M., Cowan, B., Zens, R., Duer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177-180, Prague

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC'06*

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing* (vol V), John Benjamins, Amsterdam/Philadelphia, 237-248.

Varga, D., Németh, L., Halļcsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pp. 590–596.

Vasiļjevs A., Skadiņš R. and Inguna Skadiņa I. 2011, Towards Application of User-Tailored Machine Translation, *Proceedings of Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators" JEC 2011*, Luxembourg, 2011

Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform

Núria Bel[†], Vassilis Papavasiliou[*], Prokopis Prokopidis[*], Antonio Toral[‡], Victoria Arranz[§]

[†]Universitat Pompeu Fabra, Barcelona, Spain
*Institute for Language and Speech Processing/Athena RIC, Athens, Greece
[‡]Dublin City University, Dublin, Ireland
[§]ELDA, Paris, France

E-mails: {nuria.bel}@upf.edu, {vpapa,prokopis}@ilsp.gr, {atoral}@computing.dcu.ie, {arranz}@elda.org

Abstract

The objective of the PANACEA ICT-2007.2.2 EU project is to build a platform that automates the stages involved in the acquisition, production, updating and maintenance of the large language resources required by, among others, MT systems. The development of a Corpus Acquisition Component (CAC) for extracting monolingual and bilingual data from the web is one of the most innovative building blocks of PANACEA. The CAC, which is the first stage in the PANACEA pipeline for building Language Resources, adopts an efficient and distributed methodology to crawl for web documents with rich textual content in specific languages and predefined domains. The CAC includes modules that can acquire parallel data from sites with in-domain content available in more than one language. In order to extrinsically evaluate the CAC methodology, we have conducted several experiments that used crawled parallel corpora for the identification and extraction of parallel sentences using sentence alignment. The corpora were then successfully used for domain adaptation of Machine Translation Systems.

Keywords: web crawling, boilerplate removal, corpus acquisition, IPR for language resources.

1. Introduction

There is a growing literature on using the Web for constructing large collections of monolingual and parallel collections. Such resources can be used by linguists studying language use and change (Kilgarriff and Grefenstette, 2003), and at the same time be exploited in applied research fields like machine translation, cross-lingual information retrieval, multilingual information extraction, etc. Moreover, these large collections of raw data can be automatically annotated and used to produce, by means of induction tools, a second order or synthesized derivatives: rich lexica (with morphological, syntactic and lexico-semantic information) and massive bilingual dictionaries (word and multiword based) and transfer grammars. The PANACEA LR factory is an interoperability platform of components creating complex workflows that reproduce the step-by-step process of creating such LRs. Facilities for searching, accessing web services as well as detailed documentation for each web service can be found at the PANACEA registry (http://registry.elda.org). Besides, the PANACEA myExperiment (http://myexperiment.elda.org) offers documentation, search and access facilities and a social platform for PANACEA workflows.

This paper focuses on web services and already deployed workflows for acquiring monolingual and bilingual domain-specific data. We also report on how they can be used for domain adaptation of Statistical Machine Translation (SMT) systems. The corpus acquisition procedure is described in Section 2. Further processing of the acquired data (i.e. sentence extraction, sentence alignment, etc) and their exploitation for domain adaptation are presented in Section 3. In Section 4, issues concerning the Intellectual Property Rights of the produced resources are discussed. Conclusions and future work are reported in Section 5.

2. Corpus Acquisition

In order to construct large-scale domain-specific collections, we developed a Corpus Acquisition Component (CAC) which consists of a focused monolingual (FMC) and a focused bilingual crawler (FBC). Both crawlers have been deployed as web services in the PANACEA platform and are available at http://nlp.ilsp.gr/soaplab2-axis/.

2.1 Acquiring monolingual data

The FMC adopts a distributed computing architecture based on Bixo[1] (an open source web mining toolkit that runs on top of Hadoop[2]) and integrates modules for parsing web pages, text normalization, language identification, document clean-up and text classification.

A required input resource from the user is a description of the targeted topic in a specific language. For topic description, we adopted a strategy proposed by Ardö and Golub, (2007) i.e. using triplets (`<term, relevance weight, topic-class>`) as the basic entities of the

[1] http://openbixo.org/
[2] http://hadoop.apache.org/

topic definition[3]. Topic definitions can be constructed manually or by repurposing online resources like the Eurovoc multilingual thesaurus that we used during development. Another required input is a list of seed URLs pointing to a few relevant web pages that are used to initialize the crawler.

Each fetched web page is parsed in order to extract its metadata and content. Then, the content is converted into a unified text encoding (UTF-8) and analyzed by the embedded language identifier. If the document is not in the targeted language, it is discarded. In addition, the language identifier is applied at paragraph level and paragraphs in a language other than the main document language are marked as such.

Next, each crawled, normalized and in-target language web page is compared with the topic definition. Based on the amount of terms' occurrences, their locations (i.e. title, keywords, body), and their weights, a relevance score is estimated. If this value exceeds a predefined threshold, the web page is classified as relevant and stored.

Relevant or not, each web page is parsed and its links are extracted and prioritized according to a) the relevance-to-the-topic score of their surrounding text and b) the relevance-to-the-topic score of the web page they were extracted from. Following the most promising links, FMC visits new pages and continues until a termination criterion is satisfied (i.e. time limit).

In order to provide corpora useful for linguistic purposes, FMC employs the Boilerpipe tool (Kohlschütter et al., 2010) to detect and mark parts of the HTML source that are usually redundant (i.e. advertisements, disclaimers, etc).

The final output of the FMC is a set of XML documents following the Corpus Encoding Standard[4]. An XML file relevant to the Environment domain in French can be found at http://nlp.ilsp.gr/nlp/examples/2547.xml.

4.1 Acquiring bilingual data

The FBC integrates the FMC and a module for detecting pairs of parallel documents. The required input from the user consists of a list of terms that describe a topic in two languages and a URL pointing to a multilingual web site. The FBC starts from this URL and in a spider-like mode extracts links to pages inside the same web site. Extracted links are prioritized according to the probability that they point to a translation of the web page they originated from, and the two criteria mentioned in 2.1. Following the most promising links, FBC keeps visiting new pages from the web site until no more links can be extracted.

After this stage, the pair detection module, inspired by Bitextor (Esplà-Gomis and Forcada, 2010), examines the structure of the downloaded pages to identify pairs of parallel documents. The module performs better on document pools from well-organized web sites, i.e. multilingual sites with pages containing links to translations comparable in structure and length. The final output of the FBC is a list of XML files, each pointing to a pair of files in the targeted languages[5].

3. Alignment

At this stage, we have pairs of documents produced by the FBC (see Section 2.2). In order to take advantage of this data, it should be aligned at a finer level, i.e. sentence alignment and word alignment. PANACEA has developed web services for a set of state-of-the-art sentence and word aligners. Namely, for sentence alignment, we provide web services for Hunalign, BSA and GMA. Regarding word alignment, GIZA++, Berkeley Aligner and Anymalign have been integrated. All these web services are available at http://www.cngl.ie/panacea-soaplab2-axis/. For a more detailed description of alignment web services, their implementation and deployment, please refer to (Toral et al., 2011).

The sentence-aligned data can then be used for a variety of tasks. For example, we have used this kind of data to adapt a Statistical Machine Translation system to given specific domains (environment and labour legislation) and language pairs (English--French and English--Greek) (Pecina et al., 2011). By using the domain-specific crawled and sentence-aligned data, we are able to improve the performance of Machine Translation by up to 48%.

Another use is the production of domain-specific Translation Memories (Poch et al., 2012). In this case, the data received from the FBC is first sentence-aligned and then converted into TMX, the most common format used to encode Translation Memories. This is deemed to be very useful for translators when they start translating documents for a new domain. As at that early stage they still do not have any content in their TM, having the automatically acquired TM can be helpful in order to get familiar with the characteristic bilingual terminology and other aspects of the domain.

4. IPR case study

It is the aim of PANACEA to explore all the issues related to the usability of the produced resources. Thus, work on the exploitation plan of PANACEA has led to an interesting study of the type of assets produced. This project offers a combination of data, software and web services that need to be considered at different stages. Here we focus on the specific work done on the monolingual and bilingual data described in the previous sections with the aim to establishing an appropriate and clear legal framework for its exploitation in all possible scenarios. Given the trend nowadays to crawl and use data from internet, we considered this case study as crucial for

[3] http://nlp.ilsp.gr/panacea/testinput/monolingual/ENV_topics/ENV_EN_topic.txt is an example of a list of English terms for the environment.

[4] http://www.xces.org/

[5] http://nlp.ilsp.gr/panacea/xces-xslt/202_225.xml links a pair of documents in English and Greek

this and other similar data-production approaches in particular seeing current initiatives choose options like leaving IPR issues in-handled, in the hands of the future users themselves or praising for the good nature of the owners who may take them to court. Once the internet data to be used has been listed, the procedure followed to clear out their IPR issues follows these steps: - Locating all sources and contact points. - Studying terms and conditions (use and possible distribution, if any). - Approaching providers (mostly, on a case per case basis). The complexity behind this procedure ranges considerably due to factors such as: source type; access to some institutions and blogs; need to reassure sources of no ownership right infringement; need to explain data use to data users (what is HLT?); data size; allowed negotiation time (generally long but where the needs of the future data users impose some clear restrictions). In the case of PANACEA, a large number of URLs were to be handled, 14,479, which contained 190,540 pages as a whole. However, given the cost and time restrictions imposed by both the task and the project budget, only the most frequent URLs were selected to undergo negotiation. Thresholds were set up as follows:

- For monolingual data: after an initial collection of relatively small corpora (which was performed early in the project and resulted in storing 5,623 pages from 1,175 web sites), web sites, for which under 7 pages were collected, were not examined for IPR issues; after a second experiment, which resulted in a much larger collection (184,917 pages from 13,304 sites), web sites with under 100 pages were not considered
- For bilingual and aligned data: all sources were targeted. IPR clearance was given top priority given their processing effort (aligned). 27 URLs were contained in both batches of the bilingual data, with their respective 1,948 pages. An interesting conclusion of this work was the analysis of negotiation duration and status reached at this stage of the project (with Year 2 already completed). Leaving aside refused negotiations (e.g., already 10 in the case of the monolingual data), which is a fact that should not be neglected for similar approaches, monolingual negotiations have taken between 1 day (for very fast replies) to 339 days, which shows and average duration of 66 days. Bilingual data have taken between 8 to 344 days, with an average of 176 days. This seems to be the hard reality to be faced when aiming to handle IPR for such type of data.

5. Conclusions

PANACEA project is working for the automatic production of language resources that are the critical components for the multilingual, domain tuned applications embodying different language technologies. We have presented the services already available to produce usable domain-specific aligned corpora based on the parallel data found in the internet. These web services can be chained in workflows that implement the project goals: the automatic production of language resources. By the end of the year, PANACEA will also offer web services and workflows for the automatic production of other resources such as bilingual dictionaries, monolingual rich lexica containing verb subcategorization frame information, selectional preferences, multiword extraction, and lexical semantic class of nouns.

6. References

Ardo, A., and Golub, K. 2007. Focused crawler software package. Technical report.

Espla-Gomis, M., and Forcada, M. L. 2010. “Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor”. The Prague Bulletin of Mathemathical Lingustics, 93, pp.77–86.

Kilgarriff, A. and Grefenstette, G. (2003) Web as Corpus: Introduction to the Special Issue. Computational Linguistics 29 (3) 333-347.

Kohlschuetter, C., Fankhauser, P., and Nejdl, W. 2010. “Boilerplate Detection using Shallow Text Features”. The Third ACM International Conference on Web Search and Data Mining.

Pecina, P.; A. Toral, A. Way, P. Prokopidis and V. Papavassiliou. Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. In Proceedings of the 15th Annual Conference of the European Association for Machine Translation. Leuven (Belgium). May 2011.

Poch, M.; Toral A. and Bel, N.. Language Resources Factory: case study on the acquisition of Translation Memories. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (demo track). Avignon, France. April 2012.

Toral, A.; Poch, M.; Pecina, P. and Way; A.. Towards a User-Friendly Webservice Architecture for Statistical Machine Translation in the PANACEA project. In Proceedings of the 15th Annual Conference of the European Association for Machine Translation. Leuven (Belgium). May 2011.

The PRESEMT Project

Adam Kilgarriff, George Tambouratzis
Lexical Computing Ltd., UK; ILSP, Athens, Greece
E-mail: adam@lexmasterclass.com, giorg_t@ilsp.gr

Abstract

Within the PRESEMT project, we have explored a hybrid approach to machine translation in which a small parallel corpus is used to learn mapping rules between grammatical constructions in the two languages, and large target-language corpora are used for refining translations. We have also taken forward methods for 'corpus measurement', including an implemented framework for measuring the distance between any two corpora of the same language. We briefly describe developments in both these areas.

Keywords: hybrid machine translation, machine learning, corpus distance measures, comparing corpora

1. Introduction

PRESEMT (Pattern-Recognition-based Statistically Enhanced Machine Translation) is an EU FP7 Project running from January 2010 to December 2012. It is developing a language-independent methodology for the creation of a flexible and adaptable system which can be ported to new language pairs and specific user requirements with relative ease. Unlike most statistical system, it does not assume that large parallel corpora are available for a given language pair, as they often are not. It uses a small parallel corpus to learn automatically how the syntactic constructions of the source language map to those of the target language, a bilingual dictionary for lexical transfer, and a large monolingual corpus for target-language modelling. As of April 2012, a prototype system is available on the web for the directed language pairs English to German, German to English, and Czech, Greek and Norwegian to German and to English. In the final year of the project the Consortium will port the methodology to new language pairs, involving translating from any of the aforementioned languages to Italian.

Language technology based on machine-learning from corpora will always depend on the nature and quality of the corpus or corpora used for training. With this in mind, the project has also undertaken foundational work in this area. In this extended abstract, we first outline the system and then briefly describe the work performed within the project on corpus comparison.

2. The PRESEMT MT system

This article focuses on the PRESEMT project (www.presemt.eu), which aims to develop a language-independent methodology for creating MT systems. This method overcomes well-known problems of other MT approaches, such as bilingual corpora compilation or creation of new language-specific rules.

Most recent MT approaches adopt the Statistical Machine translation paradigm (Koehn, 2010), where a statistical model is extracted probabilistically from a large parallel corpus to represent the transition from source (SL) to target language (TL). In Statistical Machine Translation, an important bottleneck is the need for extensive bilingual corpora between SL and TL. Though such corpora may exist between widely-used languages, they rarely exist for less widely-used languages, while their construction would require substantial resources.

PRESEMT builds on experience accumulated within the METIS (Dologlou et al., 2003) and METIS-2 (Markantonatou et al., 2006), projects, where the theme was the implementation of MT using solely data from TL monolingual corpora via pattern recognition techniques. Analysing the behaviour of METIS-2, a potential improvement in translation quality was identified. This involved supplementing the monolingual TL corpus with a small bilingual corpus (of typically a few hundred sentences), to provide the basis for the translation output.

The PRESEMT translation process is based on phrases, as that improves the translation quality. Translation is split into two phases, each of which focuses on processing a single type of corpus to resolve specific types of information in the output sentence. Phase 1 (Structure selection) utilises the small bilingual corpus to determine the appropriate TL phrasal structure for input sentences, establishing the order and type of TL phrases. The structure selection output is a sequence of TL structures that contain phrase and tag information and sets of TL lemmas as retrieved from the bilingual dictionary.

Phase 2 (Translation Equivalent Selection) accesses the monolingual corpus to specify the word order within each phrase and to determine whether function words need to be inserted or deleted as compared to the SL. In addition, in Phase 2 cases of lexical ambiguity are resolved by selecting one lemma from each set of possible translations. That way, the best combination of lemmas is found for a given context. Finally, a token generator transforms TL lemmas into tokens.

A major objective of the PRESEMT project is to develop an MT system that can be easily extended to new language pairs. To this end the PRESEMT project uses readily available linguistic resources as far as possible and avoids the costly development of specialised linguistic resources and tools. Such tools include statistical taggers and chunkers that provide shallow linguistic structures.

3. Corpus comparison

As argued in Kilgarriff (2001), so long as we lack a systematic account of how one corpus relates to another, both corpus linguistics and corpus-based computational

linguistics fall short of scientific standards. While that was as true when that work was done, in the 1990s, as it is now, it was perhaps forgivable then, since there were few corpora available so, in practice, scientists found themselves obliged to use whatever corpus (of the right language and, to some approximation, the right text type) was available. Now we can build corpora to order, automatically, from the web, so the question "how does this corpus relates to others I might use (of the same language) becomes critical. In PRESEMT we are following three strategies for addressing this question: Quantitative comparison, qualitative comparison, and evaluation (which we shall be reporting on later).

3.1 Qualitative comparison

Given two corpora, it has long been acknowledged that one way to get a sense of the differences between them is to look at the keywords of each *vs.* the other (see e.g. Hofland and Johanssen 1982). There has been debate on what statistics are most suitable for identifying keywords, and in Kilgarriff (2009) we make the case for:

- Normalising the frequency of each word in each corpus to a per-million figure
- Adding a parameter *k* to all normalised frequencies
- For each word, finding the ratio between the adjusted normalised frequencies in the two corpora.

The words with the highest ratio are then the keywords of corpus 1 *vs.* corpus 2, and those with the lowest are the keywords of corpus 2 *vs.* corpus 1. There are two advantages to adding *k* before taking the ratio: firstly, it allows us to take a ratio even when a word is absent in one of the corpora; and secondly, it allows us to vary *k* according to the focus of our research. A low value of *k* will tend to give lexical keywords, a higher value give more higher-frequency keywords, usually including grammatical words.

Then we can compare two corpora qualitatively by looking at the keywords of each *vs.* the other. It is usually possible to make some general statements about how the text type of each corpus differs from the text type of the other, by looking at the two lists of 100, or 200, keywords.

3.2 Quantitative comparison

Kilgarriff (2001) shows that a corpus distance measure based on frequency differences of the 500 commonest corpora work well to distinguish more, and less, similar text types. Within PRESEMT we have implemented a version of the 2001 measure within the Sketch Engine http://www.sketchengine.co.uk) so making it possible for researchers to classify which, of a set of three or more corpora for a language, are more similar and which are less so. Whereas the earlier work used a measure based on the chi-square statistic, we now use a variant of the same measure we use for keywords (with *k*=100, and taking the ratio by always dividing the higher number by the lower). We found this variant to be as precise as the one reported on before, and it is convenient to use a method consistent with keyword lists. The display we get for five well-known corpora of English is shown in Table 1.

	BASE	BAWE	BNC	Brown	BrownF
BASE		3.28	2.77	3.11	2.82
BAWE			2.15	2.21	2.09
BNC				1.59	1.32
Brown					1.47
BrownF					

Table 1: Distances between five well-known corpora of English: British Academic Spoken English (BASE), British Academic Written English (BAWE), the British National Corpus (BNC), the Brown corpus, and six 'Brown Family' corpora: Brown, LOB, FROWN, FLOB, BLOB, BE06.

The scores are 'average ratios', always guaranteed to be one (representing identical text types) or more. We can immediately see a cluster of the three corpora aiming at representativeness (BNC, Brown, Brown-Family), with the BASE, comprising spoken material, being the further-out outlier, and BAWE still an outlier but less different. We also note that Brown-family is slightly more similar to the BNC than it is to Brown, even though Brown is one of its component parts. This is perhaps because two thirds of Brown-family is British English, like the BNC, whereas Brown is entirely American.

Any user of the Sketch Engine can use the interface to find how a corpus of their own is situated in relation to other corpora. The interface does not as time of writing give a heterogeneity score for each corpus (which is needed, in order to interpret distance scores correctly) but will shortly be upgraded to provide this information.

References

Dologlou, Y., Markantonatou, S., Tambouratzis, G., Yannoutsou, O., Fourla, S., Ioannou, N. (2003) Using Monolingual Corpora for Statistical Machine Translation: The METIS System. EAMT-CLAW'03 Workshop Proceedings, Dublin, 15-17 May, pp. 61-68.

Hofland, K. & S. Johansson. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Kilgarriff, A. 2001. Comparing Corpora. *Int Jnl Corpus Linguistics,* 6 (1), pp 97-133

Kilgarriff, A. 2009. Simple Maths for Keywords. *Proc. Int Conf Corpus Linguistics*, Liverpool, UK.

Koehn, P. (2010) Statistical Machine Translation. Cambridge University Press.

Markantonatou S., S. Sofianopoulos, O. Giannoutsou & M. Vassiliou 2009: Hybrid Machine Translation for Low- and Middle- Density Languages. Language Engineering for Lesser-Studied Languages,, pp. 243-274. IOS Press.

Building bilingual terminologies from comparable corpora: The TTC TermSuite

Béatrice Daille

University of Nantes
LINA, 2 Rue de la Houssinière
BP 92208, 44322 Nantes, France
beatrice.daille@univ-nantes.fr

Abstract

In this paper, we exploit domain-specific comparable corpora to build bilingual terminologies. We present the monolingual term extraction and the bilingual alignment that will allow us to identify and translate high specialised terminology. We stress the huge importance of taking into account both simple and complex terms in a multilingual environment. Such linguistic diversity implies to combine several methods to perfect accurately both monolingual and bilingual terminology extraction tasks. The methods are implemented in TTC TermSuite based on a UIMA framework.

Keywords: comparable corpora, terminology extraction, alignment, language for special purpose

1. Introduction

The need for lexicons and terminologies is overwhelming in translation applications because of cross-linguistic divergences and mismatches that arise from the perspective of the lexicon. For scientific domains, terminological resources are often not available or up-to-date, especially for emerging domain; moreover, the languages that are covered are often limited to 2 or 3 languages of which one is English. Previously translated texts could be used to create such linguistic resources such as the GIZA++ statistical machine translation toolkit (Och and Ney, 2003). But, there is no parallel corpora for most specialized domain and most pairs of languages. To tackle the drawbacks of term alignment from parallel corpora, comparable corpora seem to be the right solution to solve textual scarcity: as monolingual productions, they are authentic texts out of translations, and the babel web ensures the availability of large amounts of multilingual documents. The TTC project relies and this hypothesis and its aim is to perform terminology extraction from comparable corpora and to demonstrate the operational benefits on MT systems.

To build high-specialized terminologies, terms are extracted monolingually from the comparable corpus. To collect close candidate terms across languages, it is necessary to use a term extraction program that is able to handle both simple and complex terms (Kageura, 2002) and able to deal with terminology variation. Once monolingual candidate terms are extracted from the two parts of the bilingual comparable corpora, the alignment program which task is to propose for a given source term, several candidate translations should be able to handle both simple and complex terms. Within this context, we present TTC TermSuite, a terminology mining chain that performs both monolingual and bilingual terminology extraction from comparable corpora for seven languages.

2. Monolingual terminology extraction

To build high-specialized terminologies, terms are extracted monolingually from the comparable corpus. To collect close candidate terms across languages, it is necessary to use a term extraction program that applies the same method in the source and in the target languages. To work at the multilingual level, we have to reconsider the rough distinction between simple and complex terms to take into account morphological compounds. Morphological compounds are identified by tokenisation programs as single-word terms but for some languages such as German, they look quite similar to multi-word terms. The translation of MWTs is the most need as they constitute around 80% of the domain-specific terms, see for example Nakagawa and Mori (2003) for Japanese language. For German language, morphological compounds appear to be much more frequent than MWTs: 52% of nouns were reported by Weller et al. (2011) on the renewable energy TTC corpus[1].

Compound consists of the concatenation of two or more lexemes to form another lexeme. We distinguish 2 types of compounds: neoclassical compounds and native compounds. The first one are built with at least one neoclassical element such as *patho*, *bio-*, *-logy* (Bauer, 1983); the second are built with words of the native language such as *windmill*. Neoclassical compound could be identified thanks to a list of combining forms and dictionary look-up (Harastani et al., 2012) and native compounds by a splitting algorithm which is combined with a dictionary look-up (Weller and Heid, 2012).

The terminological occurrences that are extracted are SWTs and MWTs whose syntactic patterns correspond either to a canonical or a variation structure. The patterns are expressed using MULTEXT part-of-speech tags and are provided for all TTC languages. The main patterns whatever is the language are N and A for SWTs. For French and Spanish, the main patterns of MWTs are N N, N S:p N and N A. The variants handled are graphical, morphological, and syntactic. The three types of terms face up variants even if some are more likely to concern one main type

[1] http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html

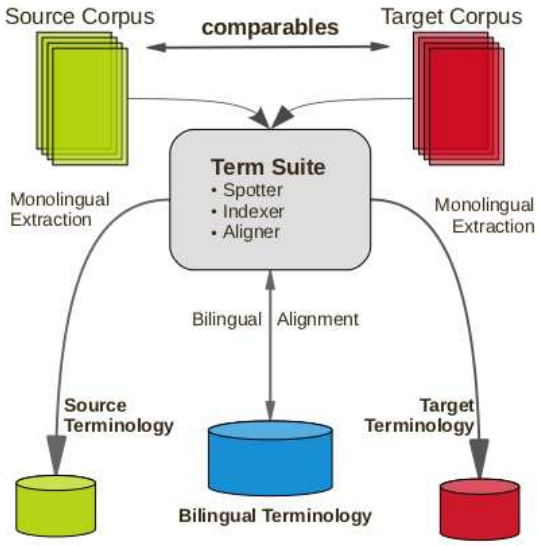


Figure 1: TTC TermSuite Architecture

such syntactic variants for MWTs. Monolingual terminology extraction and variant detection for multi-word terms were evaluated by Gojun and Heid (2012) for German language on the security domain. They gave a recall of 65% and were able to increase the existing terminology of the domain with new terms by 25%.

3. Bilingual terminology alignment

Once source and target terminologies are extracted form monolingual corpora, the alignment step could be set up. The output is a bilingual domain-specific terminology lexicon where for one source term you need to translate, you will obtain several candidate translations ranked form the most likely to the less. The method to align a source term with a target term relies on the hypothesis that a word and its translation tend to occur in similar contexts within a comparable corpora. The context of a word is expressed thanks to co-occurrences appearing in a context window. The co-occurrences are translated using a general bilingual language dictionary in the target language and compare to existing contexts of target words. The context-based projection approach proposed by (Rapp, 1995) for aligning words from bilingual comparable corpora is the gold standard. Using this approach, a precision of 60% is obtained for the translation of SWTs by examining the first 20 candidates translations using specialized language corpora of small size (0.1 million-word English-German corpus in (Déjean et al., 2002) and 1.5 million-word French-Japanese corpus in (Morin et al., 2010). But results drop significantly for MWTs, a precision of 42% of the 20 first candidates in a 0.84 million-word French-Japanese specialized language corpus (Morin et al., 2010). It is thus necessary to use another method.

For MWTs, it is possible to exploit the compositional property that characterizes half of MWTs - 48.7% have been reported by (Baldwin and Tanaka, 2004) for English/Japanese N N compounds. A compositional translation approach will translate each word of the MWT individually using a bilingual dictionary, and then appropriately piecing together the translate parts. It is possible to implement the composition approach at the morpheme or at the word level (Baldwin and Tanaka, 2004). For neoclassical compounds, we apply the compositional approach at the morpheme level making the assumption that most neoclassical compounds in a source language translate compositionally to neoclassical compounds in a target language. For example, the translation of the English noun *hydrology* in French is *hydrologie*, which can be interpreted by the combination of the translation of the composing elements, *hydro* (water): Fr *hydro* and *logy* (study): Fr *logie*. For MWTs, we apply the compositional approach at the word level. For example, the translation of the French MWT *fatigue chronique* is obtained by translating both *fatigue* and *chronique* into *fatigue* and *chronic* using a bilingual dictionary look-up.

4. TTC TermSuite

TTC TermSuite[2] is designed to perform bilingual term extraction from comparable corpora in five European languages: English, French, German, Spanish and one under-resourced language, Latvian, as well as in Chinese and Russian. The general architecture is presented in Figure 1. TTC TermSuite is based on the UIMA framework which supports applications that analyze large volumes of unstructured information. UIMA was developed initially by IBM (Ferrucci and Lally, 2004) but is now an Apache project[3].

4.1. General architecture

The architecture could be described from the point of view of the hierarchy of treatments or from the point of view of the data workflow. TTC TermSuite is a 3-step functional architecture that is driven by the required inputs and provided outputs of each tool. The bilingual term alignment (step 3 ALIGNER) requires processes of monolingual term extraction (step 2 INDEXER), itself requiring text processing (step 1 SPOTTER). The spotter applies a shallow pre-processing of the monolingual corpora, performing tokenization, part-of-speech tagging, stemming and lemmatization. The workflow is summarized in Figure 2: at the first step, we treat one document by one. If we get n documents, we will obtained n documents linguistically analyzed through the spotter; From this set of documents, we perform monolingual term extraction using the indexer which output is a terminology file; The last step is the alignment that requires one source and one target terminology files and proposes as output a bilingual terminology file.

4.2. Monolingual term extraction

Monolingual term extraction consists in processing a monolingual corpus document by document and in providing its terminology. It involves:

1. the recognition and the indexing of both single-word and multi-word terms;
2. the computing of their relative frequency and their domain specificity;

[2] http://code.google.com/p/ttc-project
[3] http://uima.apache.org

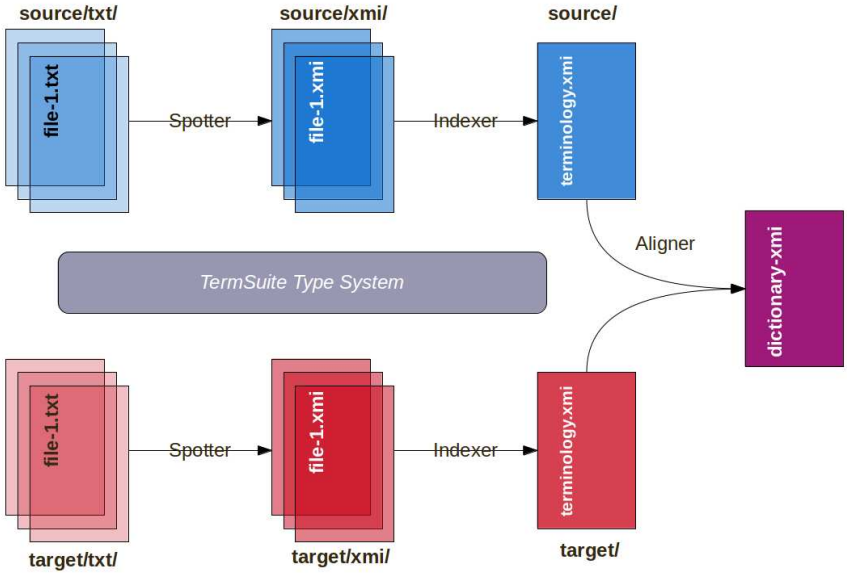


Figure 2: TTC TermSuite Workflow

3. the detection of neoclassical compounds above the set of single-word terms;
4. the grouping of term variants;
5. the filtering of some candidates using thresholds that could be expressed on the relative frequency or the domain specificity.

The term variant grouping functionality takes place once terms have been annotated as single-words or multi-words, and once single-word terms have been flagged as thus or as neoclassical compounds. After the collecting of term-like units, the TTC TermSuite organizes them that result in clusters of candidate terms. The clustering adopts different strategies that depend on the nature of the variation: graphical term variants are detected using edit distances, morphological variants using monolingual lists of affixes, and syntactical term variants using pattern rules over feature structures.

4.3. Bilingual term alignment

Bilingual term alignment adopts different strategies with regards to the nature of terms: for a SWT, it is the context-based projection approach; for neoclassical compounds and MWT compositionality-based method approaches are launched. The alignment of neo-classical compounds were evaluated on the En-Fr, En-De and En-Es pairs of languages on the TTC renewable energy corpus and showed a high precision for all pairs of languages (Harastani et al., 2012). For example, 100 aligned terms were obtained for the En-Fr pair with a precision of 98%. SWTs and MWTs have not been yet evaluated but as state of art methods have been implemented, we foresee for SWTs to reach a precision of around 60% on the first 20 translations, and for MWTs a precision of 68% for a recall of 40% (Morin and Daille, 2009). However, the combination of the two main strategies: context and compositionality-based methods should increase the overall performance. The coming evaluation of TTC TermSuite will hopefully confirm these numbers.

5. Conclusion

TTC term extraction techniques rely on low-level annotated corpora where sentence boundaries, word classes and lemmas are annotated. Patterns are used to extract term candidates: simple and complex terms are handled as well as their variants. Several statistics are computed that could be used to filter the list of monolingual candidate terms. The alignment combined compositional and context-based methods to treat both simple and complex terms. The bilingual terminology building is implemented in TTC TermSuite based on the UIMA framework for English, French, German, Spanish, Latvian, Chinese, and Russian.

6. Acknowledgement

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (*/*FP7/2007-2013*/*) under Grant Agreement no 248005.

7. References

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31.

Laurie Bauer. 1983. *English word-formation*. Cambridge university press.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10:327–348, September.

Anita Gojun, Ulrich Heid, Bernd Weissbach, Carola Loth, and Insa Mingers. 2012. Adaptating and evaluating a generic term extraction tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istambul, Turkey, May. ELDA.

Rima Harastani, Béatrice Daille, and Emmanuel Morin. 2012. Neoclassical compound alignments from comparable corpora. In Alexander F. Gelbukh, editor, *CICLing (2)*, volume 7182 of *Lecture Notes in Computer Science*, pages 72–82. Springer.

K. Kageura. 2002. *The dynamics of terminology: a descriptive theory of term formation and terminological growth*. Terminology and lexicography research and practice. J. Benjamins Pub.

E. Morin and B. Daille. 2009. Compositionality and lexical alignment of multi-word terms. In *Language Resources and Evaluation (LRE)*, volume 44 of *Multiword expression: hard going or plain sailing*, pages 79–95. P. Rayson, S. Piao, S. Sharoff, S. Evert, B. Villada Moirón, springer netherlands edition.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2010. Brains, not brawn: The use of "smart" comparable corpora in bilingual terminology mining. *TSLP*, 7(1).

Hirosi Nakagawa and Tatsunori Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Reinhard Rapp. 1995. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

Marion Weller and Ulrich Heid. 2012. Simple methods for dealing with term variation and term alignment. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istambul, Turkey, May. ELDA.

Marion Weller, Anita Gojun, Ulrich Heid, Béatrice Daille, and Rima Harastani. 2011. Simple methods for dealing with term variation and term alignment. In Kyo Kageura and Pierre Zweigenbaum, editors, *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pages 87–93, Paris, France, November. INALCO.

A viewing and processing tool for the analysis of a comparable corpus of Kiranti mythology

Aimée Lahaussois*, Séverine Guillaume**
*CNRS, UMR 7597, HTL, Univ Paris Diderot, Sorbonne Paris Cité, F-75013 Paris
**CNRS, UMR 7107, LACITO
E-mail: aimee.lahaussois@linguist.jussieu.fr, guillaume@vjf.cnrs.fr

Abstract

This presentation describes a trilingual corpus of three endangered languages of the Kiranti group (Tibeto-Burman family) from Eastern Nepal. The languages, which are exclusively oral, share a rich mythology, and it is thus possible to build a corpus of the same native narrative material in the three languages. The segments of similar semantic content are tagged with a "similarity" label to identify correspondences among the three language versions of the story. An interface has been developed to allow these similarities to be viewed together, in order to allow make possible comparison of the different lexical and morphosyntactic features of each language. A concordancer makes it possible to see the various occurrences of words or glosses, and to further compare and contrast the languages.

Keywords: trilingual comparable corpus, Kiranti languages, mythological narrative cycle

1. Introduction

The challenges encountered when using various stimulus materials to generate parallel or similar texts for language comparison are well-known: "Recording free discourse and/or narrations of picture-book stories may lead to multi-lingual corpora which are too diverse both structurally and semantically to allow for direct comparison because one cannot be sure that the data at hand are compatible with one another." (Stolz and Stolz, 2008: 33). In reaction to this, we became interested in the idea of using *native* stories in different languages as the basis for comparative work. Languages of the Kiranti group of Eastern Nepal share a very rich mythology (Ebert and Gaenszle, 2009) which can be used for this purpose. The stories are remarkably similar, both in their content and, in some cases, in their use of idiosyncratic morphosyntax which is otherwise difficult to elicit.

The Kiranti languages of Eastern Nepal are in the Tibeto-Burman family. There are two major subdivisions within the group: Limbu, on the one hand, is the language in the group with the largest number of speakers and a writing system; the 30-odd Rai languages make up the rest of the group. The Rai languages are exclusively oral[1], and spoken by small communities usually numbering several thousand speakers. They are severely endangered, due to the inroads of the national language Nepali.

While there have been a number of descriptions of Rai languages[2], there has been very little comparative work, except on a case by case basis. Ebert (1994, 2003) has written about the shared structure of the Kiranti languages, and Michailovsky (2009) has carried out work on the phonological reconstruction of proto-Kiranti, but on the whole, comparative work is limited, both in number of languages (a sample of six for Ebert's 1994 comparative work) and also in scope.

The body of shared mythology among Rai peoples presents itself as an appealing option for carrying out comparison work. The ubiquity of the mythological cycle as a form of narrative becomes apparent quite quickly to anyone working on the documentation of these languages. Most spontaneously told stories will be drawn from this body, and the stories are remarkably similar across languages.

Our goal in this paper is to describe how we have created a prototype for a Kiranti comparable corpus by aligning the same story, taken from the mythological cycle, in three languages from the group in order to advance and enable comparative work among these languages.

2. The Kiranti comparable corpus

The data presented in this paper is from personal fieldwork on three languages of the Kiranti subgroup, namely Thulung, Koyi and Khaling[3]. The creation of a comparable corpus could also be achieved using materials in existing descriptions of other Kiranti languages. Such grammars contain transcribed oral narrative which almost invariably includes elements of the same mythological cycle.

For our prototype comparable corpus, we chose to use a single story, with the goal of building up the corpus to include new stories as we collect and align them. The basic storyline for the story which we selected is the following:

[1] Any references to Rai "texts" in this paper are to transcriptions (by the linguistic researcher) of oral narratives.

[2] One can cite for example the grammars that have come out of the Himalayan Languages Project (Sunwar, Wambule, Jero, Yamphu, Bantawa, Dumi,...).

[3] The Khaling data comes from fieldwork done in collaboration with Dhana Bahadur Khaling, Guillaume Jacques, Boyd Michailovsky, Martine Mazaudon, Marie-Caroline Pons.

> Kakcilip and his two olders sisters are orphaned and must learn to fend for themselves. Kakcilip, being the youngest, is not able to contribute much, and his sisters take on the bulk of the work. One day, while they are out in the forest, Kakcilip falls asleep. The sisters, thinking he is dead, leave him behind and decide to separate, each flying in a different direction. One of the sisters encounters an owl, who eats her. The other sister comes looking for her, and manages to get her bones back from the owl. With an enchantment, she rewakens her sister and explains what has happened. Later on, the sisters encounter a series of animals--a louse, a flea, a goat and finally a cock which calls out "khakcilipa" when it comes near them. They realize this is a sign from their brother, as the cock is calling his name, and follow him back to a place where they are reunited. Kakcilip has in the meantime had an adventure of his own in which he, while fishing, caught a stone which turned out to be a female figure he eventually marries.

In some cases, the story was narrated as an independent story: this is the case in Thulung and Khaling. In Koyi, it is woven into a very long origin myth. Thus our stories are all of different lengths[4] (Thulung: 12 minutes, Khaling: 13 minutes, Koyi: 63 minutes), with the Koyi version contained a large amount of additional material which is not in the other two stories.

3. Building the corpus

The corpus was built from preexisting interlinearized XML "annotation" files of the Kakcilip story in three languages. These files were in a format which is used by the LACITO Archive (http://lacito.vjf.cnrs.fr/archivage/index_en.htm) and contain three tiers of data (transcription into IPA, glosses, and free translation), as is typical of analyzed field data used in the description of oral and endangered languages. In the case of all three languages in our corpus, this three-tiered structure was generated using interlinearization software called ITE (Interlinear Text Editor) developed specifically for the LACITO Archive by Michel Jacobson.

Because each language's XML annotation files for the Kakcilip story are archived, we decided, in compiling our corpus, to preserve the original format of the files, rather than modify them to include alignment data. We therefore decided to create a distinct alignment file, in which we defined similar segments, which we call "similarities", across the different texts making up the corpus. A similarity is defined here as a segment, represented by one or more sentences, containing material of similar narrative content or function. Our definition is thus based on narrative and not lexical or morphosyntactic criteria. While we would have prefered a configuration where the basis for similarity alignments was more linguistically-oriented criteria, this was not possible considering the spontaneously produced narrative data we had to work with. Unsurprisingly though, passages of similar narrative content often contain lexical material and structures that are close and sometimes even identical, so that in effect our narratively-based alignment proves useful for linguistic analysis.

The similarities were identified manually by reading through each of the texts in language pairs (Thulung-Koyi, Koyi-Khaling, Thulung-Khaling) and recording into a spreadsheet which sentence numbers of each text corresponded, in semantic content, to which others, and assigning to each correspondence a similarity label.

The spreadsheet was then converted into XML using a perl script, as illustrated in Figure 1.[5]

The annotation files called up by the alignment file contain information about the content of each of viewing levels (users can chose to look at the data in Text, Word or Morpheme views) generated by the ITE software. The text (<TEXT>) breaks down into sentences (<S>) which in turn break down into words (<W>) and morphemes (<M>). Each unit can contain a transcription (<FORM>) and a translation or gloss (<TRANSL>. This is illustrated in Figure 2.

The comparable corpus is thus made up of four files: the three languages' annotation files, which contain the entire version of the story in each language, and an alignment file in XML which contains the information laying out the correspondences between the language versions.

We then defined a graphic interface making it possible to view the alignments of sentences. Considering that a priority for endangered language documentation is often the widespread diffusion of data, we decided to use web-related technology. PHP and XSL style sheets were created to view the corpus.

The first viewing option is of the individual texts in their entirety, with one language per column. We call this the "integral text view". Similarities are identified by a color scheme, so that they can be identified across languages at a glance. This was important because, owing to the great differences in length between the Koyi version of the story and the other two language versions, and the different ordering of narrative events, the similarities rarely occur on the same page in all three languages. The "integral text view" is illustrated in Figure 3.

The second viewing option allows the user to select one of the similarities, and see all the content which corresponds to it in the different languages. We call this the "similarity view", and it is obtained by clicking on any similarity label in any of the three stories. The similarity view is illustrated in Figure 4.

Each of these viewing options has a related XSL style sheet and uses a PHP program to switch from one view to the other.

[4] The standard for comparison is the durations of the language versions, as the transcriptions which make up the comparable corpus are of oral recordings.

[5] Figure 1 and all subsequent figures are found at the end of the article.

We have also developed a concordancer which makes it possible to search for any word or gloss found within the transcription and glossing tiers respectively. Figure 5 shows the results of a concordance on the gloss "sister". The results show the transcription tier, with the word corresponding to the concordanced gloss highlighted (regardless of whether the search was for a word or a gloss). The sentence and source text for each result are identified (the "text" label in the left-most column identifies the story, starting with its Ethnologue language code, TDH=Thulung, KKT=Koyi, KHA=Khaling), and the left and right context for the term are given. The concordance function in effect generates a trilingual correspondence for any gloss in the corpus, and is a useful way to build up a trilingual glossary. This function will be more useful as the corpus is expanded to include more stories covering a greater narrative (and therefore lexical) range.

Each occurrence can be selected (by clicking on the highlighted word) and opens the similarity view: the sentence, if it is part of a similarity set, is shown together with the corresponding sentences in other languages. This makes it possible to identify the morphosyntactic constructions used to expressing the same narrative content.

A concordance of the gloss "INS" (instrumental marker) leads, among other results, to Similarity 35 (which, in the interest of space, is reproduced not as a screen shot but as the text which makes up the similarity, namely examples (1) and (2) below):

(1) [THU][6]
naŋlo-**nuŋ** kutso-nuŋ
winnowing.basket-COM broom-COM
ʤer-tʰɑk-y kʰrems-ɖa
hold-hide-3SG>3SG.PST cover-3SG.PST
ba-iɖa-m
be-3SG.PST-NMLZ
'He held and hid with the basket and broom and covered himself.'

(2) [KOY]
runtsʰis-**wa** dʰep-nasi-nɔ
winnowing.basket-INS cover-3SG.PST.REFL-SEQ
mɔ tsʰa sul-nasi tsʰa
be.anim.3SG.PST HS hide-3SG.PST.REFL HS
'He covered himself with a basket and stayed there and hid.'

Where Koyi uses an instrumental marker (*-wa*) to encode the semantic role of the instrument (the winnowing basket Kakcilip is using to hide himself), Thulung unexpectedly uses a comitative marker (instead of instrumental marker *-ka*), usually reserved to express accompaniment by a person. This type of example points to the potential usefulness of this corpus in uncovering, through comparison, language-internal variation which would not necessarily be covered in descriptive grammars.

4. Issues encountered

4.1 Methodological issues

A number of issues were encountered during the construction of the comparable corpus, including methodological questions about the necessity for manual alignment of the texts, and the nature of similarities. These are discussed below.

4.1.1 Hand alignment

The identification and definition of similarities in the material must be carried out manually. From our understanding, the tools available for well-described languages with numerous digital resources (dictionaries, POS taggers, etc) cannot be used to automatize the work we have done with the Kiranti corpus. This is precisely one of the significant differences between so-called mainstream languages and little described minority ones. The matter of hand-alignment does not represent a problem in the case of the Kiranti corpus, as we are dealing with very small data sets. Nonetheless it will be necessary as the corpus grows to include other languages to find methods to partially automatize the alignment.

4.1.2. The typology of similarity judgments

As defined in section 3 above, similarity judgments were based on the degree of narrative similarity of textual segments, and were thus inherently subjective. Because the three versions of the story are close, and because of the proximity of these languages, similarities often involve equivalent lexical items and sometimes even the same morphosyntactic constructions, but not always. Some examples will be given of the three basic types of similarities we have found.

Similarities with only narrative function in common

Similarity 5 aligns sentences which share almost nothing but narrative function. There is not a single word which is the same across the languages, and grammatically, the only shared element is the use of a converbal marker (*-saka* in Thulung, *-to* in Khaling), as seen in examples (3) and (4) below.

(3) [THU]
əni meɖɖa-m pətsʰi kolem tsʰipʤi-kam nem
and then-NMLZ after one.day cut.bamboo-GEN house
bɤne-saka mɯ-gunu u-ri kʰaktsilip-lai
make-CVB that-inside 3SG.POSS-sibling Kakcilip-DAT
am-saka
make.sleep-CVB
'Then after they made a house out of pieces of big bamboo, and put their brother Kakcilip to sleep inside it.'

(4) [KHA]

[6] All examples will be preceded by a three-letter abbreviation of the language name: THU for Thulung, KOY for Koyi and KHA for Khaling.

grômmɛ-kolo lasmɛ-su-ʔɛ dhawa mɛ ʤʌkhʌl
Gromme-COM Lasme-DU-ERG quickly that nettle.fiber
kâ:k-tɛsu-lo mɛ lektsêm-ʔɛ
peel-3DU>3SG.PST-TEMP that nettle.core-INS
nek-to nek-to khɵs-tɛ
cover-CVB cover-CVB go-3SG.PST
'Gromme and Lasme quickly peeled the nettle fiber and covered him with the inside of the fiber.'

The bamboo in one version of the story is nettle fiber in the other; Kakcilip is mentioned by name in one verions but not the other; the house which covers Kakcilip in one version is a pile of fiber in the other. And yet, narratively, this is the point at which their brother gets covered--because they think he is dead in the Thulung version, and because he is sleeping and they do not see him in the Khaling version--and at which the sisters and brother begin to live their separate stories. Linguistically, this similarity brings us very little, but it could be useful for, for example, an ethnographic study of the evolution, across Kiranti tribes, of basic household activities (the story makes clear that the bamboo- and nettle-peeling are a fundamental household chore).

Similarities with narrative content and some lexical material in common

Similarity 3 aligns sentences which share narrative content (it refers to the point at which the protagonists become orphans) and also some lexical and grammatical material, as seen in examples (5) and (6).

(5) [THU]
mɯrmim-kam tin ʤana ba-mri tsɤŋɖa tura
3PL-GEN three person be-3PL.PST later orphan
dym-miri-ma ba-mri
become-3PL.PST-SEQ be-3PL.PST
'The three of them were there and later became orphans.'

(6) [KHA]
grômmɛ lasmɛ khaktsalʌp tsɵttsɵ mō:-tnu-lo
Gromme Lasme Kakcalop children be-3PL.PST-TEMP
reskʌp tshɯk-tɛnu
orphan become-3PL.PST
'The children Gromme, Lasme and Kakcalop were there and became orphans.'

These two sentences contain examples of existential predication; both use clause sequencing morphosyntax (*-ma* for Thulung, *-lo* for Khaling) and they share lexical items "orphan" and "become" (the latter with a 3rd plural past conjugation in both languages). Again, this is not earth-shattering, linguistically, but provides interesting information.

Similarities revealing shared grammatical constructions

In other cases, the alignments turn up some shared linguistic constructions.

Similarity 4 (examples (7) and (8)) reveals an identical construction for "to come to a decision, to advise with each other", which we find in both Khaling and Thulung here. In Thulung, it involves a loan word from Nepali (*salla*) but in both cases it involves the verb "to do", and we see that in both languages, the agents are ergative-marked. This is a construction that does not come up naturally in elicitation, and the fact that it emerges from the data suggests that there is something to be gained from an alignment based on narrative content.

(7) [THU]
utsi-walwak-ka ʤau-nuŋ kʰleu-nuŋ-ka
3DU.POSS-sibling-ERG Jau-COM Khleu-COM-ERG
tsʰəhi səlla bet-tsi ʔe
CONTR advice do-3DU>3SG.PST HS
'Jau and Khleu came to a decision.'

(8) [KHA]
tunɵl didi bahini grômmɛ
one.day older.sister younger.sister Gromme
lasmɛ-su-ʔɛ mɵl mɯ-ssu
Lasme-DU-ERG counsel do-3DU>3.PST
'One day, Gromme and Lasme had a discussion.'

Similarity 7 (examples (9) and (10)) brings up two elements of interest: the lexical items "hunger" and also the construction "to fall asleep" which, in both languages, contains an additional aspect-bearing element (the auxiliary verbs *sutʂ-* in KOY and *dɵk-* in KHA) which, again, does not come up unless in an appropriate context. An additional element of interest here is that *soʔwa* (in example (9)), elicited in Koyi as a single word, appears to be a mistake: looking at the Khaling cognate and at how the word is used in Khaling suggests that the Koyi equivalent should probably have been analyzed as *soʔ-wa* (hunger-INS). This remains to be verified with a native speaker, but would point to a potential additional benefit of the multilingual alignment if it helps refine transcription and analysis.

(9) [KOY]
ʤimu a-dʰoʔd-u ne soʔwa
food NEG-find-3SG>3SG.PST TOP hunger
dʰal-ʤa soʔwa dʰal-ʤa-lɔ
sway-DUR.3SG.PST hunger sway-DUR.3SG.PST-TEMP
ne ipʰ-a-sutʂ-a tsʰa
TOP sleep-copy-AUX-3SG.PST HS
'When he could not find food, he swayed from hunger, when he swayed from hunger, he fell asleep.'

(10) [KHA]
sô:-ʔɛ mʌt-tɛ-na kɯmîn-ʔɛ
hunger-INS have.to-3SG.PST-SEQ thirst-INS
mʌt-tɛ-na ʔip-dɵk-tɛ-m
have.to-3SG.PST-SEQ sleep-AUX-3SG.PST-NMLZ
'He was hungry and thirsty and had fallen asleep.'

4.1.3. Minor issues

A number of other minor issues were identified, which are part and parcel of the alignment of any material across languages.

-It is important that the glosses used across the languages of the corpus be consistent, in order to simplify concordancing. Even though the three versions of the story were analyzed and glossed by the same person, there are some inconsistencies that must be corrected.

-The similar content for one segment is only found in two of the languages and not the third: this was of course a minor problem, and inevitable given the different narrative structures of the three versions of the story. The alignment file records sentence number information as long as at least two languages share any one similarity.

-The chosen unit for identification of similarities is the sentence, yet only part of the sentence contains similar material across languages. Some similarities thus look like they contain very different material. It was nonetheless felt to be important that any similarities be identified, even if they only involved a small part of a sentence, as any similarity could be relevant for linguistic comparison.

-The order in which the similarities occur within each narrative is different across languages. We resolved this issue by using different colors for each similarity, in order to be able to identify them visually at a glance, and by making it possible to call up a specific similarity's content in the three languages by clicking on the similarity label. (The result is what we see in Figure 4).

4.2. Comparable vs parallel corpus?

One interesting consideration is whether we are dealing with materials for a comparable or a parallel corpus in this instance. If we take the basic definitions laid out in the EAGLES report on corpora, "a parallel corpus is a collection of texts, each of which is translated into one or more other languages than the original." This definition is opposed to that for a comparable corpus, "which selects similar texts in more than one language or variety, [with] as yet no agreement on the nature of the similarity." (Sinclair, 1996). On the one hand, the texts are not identical, something demonstrated very quickly when trying to align the segments. This would suggest that these materials make up a comparable corpus. As a general rule, though, languages have quite different ways of encoding information, resulting in different lengths for a same text, suggesting that no two texts, even when they result from translation, can ever be truly parallel. Note Stolz's (2007: 105) comments about the Petit Prince multilingual corpus: "identical length can only be achieved by cutting off the text at a pre-determined mark because the languages differ widely as to the number of pages, words, or sentences they use." One of the main issues in determining whether we must consider this a comparable or a parallel corpus is that the bulk of theoretical work on corpora seems to involve written materials. In the case of oral materials, which can contain all manner of production errors and self-corrections, it is difficult to imagine that two narratives could ever be "parallel", even if they are by the same speaker. And yet the material, at a metalinguistic level, is the same. To cite Maia (2003) "comparability is in the eye of the beholder."

One of the reasons the questions is even relevant is that there is some debate about whether the Kiranti languages constitute a genetic grouping or instead a cluster of languages that have been in contact for a very long time within a cultural area. "It has never been shown that Kiranti [..] is a valid genetic unit. [...] Hansson assumes in an unpublished report of the Survey Project [Linguistic Survey of Nepal] that the cluster of Kiranti languages results from several migration waves of Tibeto-Burman groups that have influenced each other for a longer period." (Ebert, 2003: 516). Is the material making up our corpus an ancestral proto-Kiranti mythological cycle which has been transmitted through time into successive generations of daughter languages (in which case it is originally the *same text*) or have these stories been transmitted through cultural borrowing among languages which look close but are perhaps not genetically related (despite what looks like a fair amount of shared vocabulary--see Michailovsky (2009) for proto-Kiranti reconstructions), in which case our different language versions of the story constitute *translations* of the original? These questions of genetic grouping and inheritance may well be what this corpus enables to get closer to resolving: lexically, the languages look quite close, but structurally, much more analysis is needed.

5. Avenues opened by such a comparable corpus

The methodology proposed in this paper should in principle be applicable to other languages and subgroups, as long as narratives can be found which are common to the languages to be compared. The main goal, as we conceive it, is essentially linguistic: we aim to find narrative materials that can reveal significant aspects of the (morpho)syntax of the language studied in its own terms.

One such project is currently underway using the Kiranti comparable corpus: a study of the scope of dual and comitative marking, of their combination with other case markers, and coocurrence with numerals and classifiers. The corpus seems well adapted to such a study, and the data so far gives evidence of considerable variation. One appealing aspect of the multilingual corpus is that the similarities reveal unexpected uses, such as seen in examples (1) and (2), where a concordance for comitative markers revealed the use of an instrumental marker in one of the languages.

The Kiranti corpus fits into a larger project, in collaboration with Guillaume Jacques and Alexis Michaud, of building comparable corpora for three subgroups of Tibeto-Burman languages from the greater Himalayan region (Kiranti, rGyalrong and Na). While only Kiranti languages have shared native mythology, the rGyalrong and Na languages have folklore (borrowed from Tibetan in the case of rGyalrongic languages) which

would provide rich materials for the building of such a comparable corpus.

One other angle that we would like to explore is the extension of this concept of comparable corpus to different configurations[7]:

1) multiple versions of a same story by a single speaker (intra-speaker variation)

2) multiple speakers of a same dialect

3) multiple speakers of different dialects of the same language

In addition to the possibilities the comparable corpus opens for linguistic analysis and comparison, there is a strong potential for use by ethnographers documenting oral reports of different customs across a number of communities.

6. Conclusion

While work on endangered languages has embraced the possibilities of corpus linguistics for some time, we feel that our multilingual comparable corpus, which has the crucial distinction of being built of native narrative materials, represents a new tool in the arsenal of the linguist wishing to do comparative work on underdescribed languages.

The size of the comparable corpus presented here is very small (as is natural considering the labor-intensive nature of data collection, transcription, glossing, translation and sound-synchronization, usually involving a single linguist), but will be expanded with additional matching texts and additional languages in the group. This type of comparable corpus will make a larger-scale comparison of the Kiranti languages, which has been limited, more feasible.

The small size of the corpus, the necessity of manual alignment (of a sometimes subjective nature), may be countered by the fact that it does not suffer from most of the biases of larger parallel corpora of more mainstream languages. Wälchli (2007: 133) cites the following biases for the use of parallel text corpora for typological research: "(a) written language bias [...], (b) bias toward planned (conscious) language use (including purism) [...], (c) bias toward religious and legalese registers, (d) narrative register bias, (e) bias toward large languages (in spread zones), (f) bias toward standardized (simplified?) language varieties, (g) bias toward non-native use of languages, (h) bias toward translated language (rather than original language use)."

The only one of these biases which can be leveled against the Kiranti comparable corpus is (d), namely "narrative register bias", as all the material is from a single narrative register. The Kiranti corpus is based exclusively on transcribed oral narrative material; it is made up of foundational mythological texts which cannot be claimed to be religious (or legal) in nature. The languages are spoken by at most several thousand people in a mountainous region, and the corpus is thus made up of truly "minority" material for which there is no standardized language variety (standardization seems to be the domain of written languages, and endangered languages show "an additional layer of variation" even among oral tradition languages (Grinevald, 2007: 45)). As the corpus does not involve translation (the free translation in the data is associated to each sentence by the linguist after data collection) and therefore represents native language use.[8]

The avoidance of so many of the biases against parallel corpora is very strongly in the favor of a comparable corpus such as we have produced. There seems to be enough evidence of the potential usefulness of the corpus and viewing and analysis tool that we feel it to be worthwhile to continue to build the corpus, initially with additional texts already in our possession, and later on by including data from other languages. We feel that the Kiranti comparable corpus may ultimately provide a means of getting a better sense of the linguistic variation (both internal and cross-linguistic) in Kiranti languages, and perhaps offer evidence towards deciding whether or not this is genetic grouping.

7. References

Davies, A., (2003). *The native speaker: myth and reality*. Clevedon: Multilingual Matters

Ebert, K., (1994). *The Structure of the Kiranti Languages*. Arbeiten des Seminar fur Allgemeine Sprachwissenschaft Nr. 13. Zürich: Universitat Zürich Seminar fur Allgemeine Sprachwissenschaft.

Ebert, K., (2003). Kiranti languages: an overview. In: Thurgood G. and R. LaPolla (eds.), *The Sino-Tibetan Languages*. London and New York: Routledge, pp. 505-517.

Ebert, K. and Gaenszle, M., (2009). *Rai mythology: Kiranti Oral Texts* (Harvard Oriental Series, 69). Cambridge: Harvard university press.

Grinevald, C., (2007). Encounters at the brink: linguistic fieldwork among speakers of endangered languages. In: Miyaoka, O., Sakiyama, O. & Krauss, M. (eds), *The Vanishing Languages of the Pacific Rim*. Oxford, Oxford University Press

Jacobson, M., Interlinear Text Editor: http://michel.jacobson.free.fr/ITE/index_en.html

Lewis, P.M., (ed.), (2009). *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version**:** http://www.ethnologue.com/**.**

Maia, B., (2003). What are comparable corpora? Electronic resource, found at

[7] This idea comes from Guillaume Jacques and Alexis Michaud (pc)

[8] The question of who deserves to be called a "native" speaker is a fairly complex issue (Davies 2003), all the more so in situations of extreme endangerement and intense contact.

http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

Michailovsky, B., (1975). Notes on the Kiranti Verb <East Nepal>. *Linguistics of the Tibeto-Burman Area* 2.2. pp.183-218.

Michailovsky, B., (2009). Preliminaries to the comparative study of the Kiranti subgroup of Tibeto-Burman. *Proceedings of the International Symposium on Sino-Tibetan Comparative Studies in the 21st Century, June 24-25, 2010*. Academia Sinica, Taipei, Taiwan. pp. 145-70.

Sinclair, J., (1996). Preliminary recommendations on Corpus Typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

Stolz, T., (2007). Harry Potter meets Le Petit Prince: On the usefulness of parallel corpora in cross-linguistic investigations. In: Cysouw, M & Wälchli, B. (eds.), Parallel Texts: Using Translational Equivalents in Linguistic Typology. Theme issue of Sprachtypologie und Universalienforschung STUF 60.2: 100-117.

Stolz, C. and Stolz, T., (2008). Functional-typological Approaches to Parallel and Comparable Corpora: the Bremen Mixed Corpus. LREC 2008: Workshop on building and using comparable corpora. 33-38.

Cysouw, M. and Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. In: Cysouw, M. & Wälchli, B. (eds.), Parallel Texts: Using Translational Equivalents in Linguistic Typology. Theme issue of Sprachtypologie und Universalienforschung STUF 60.2: 95-99.

Wälchli, B., (2007). Advantages and disadvantages of using parallel texts in typological investigations. In: Cysouw, M. & Wälchli, B. (eds.), Parallel Texts: Using Translational Equivalents in Linguistic Typology. Theme issue of Sprachtypologie und Universalienforschung STUF 60.2: 118-134.

8. Acknowledments, abbreviations

We would like to thank various institutions and agencies which provided funding for field research on the three languages in our corpus: the Fulbright Foundation, the Hans Rausing Endangered Language Documentation Program, and LACITO-CNRS.

Abbreviations: The Leipzig Glossing Rules have been applied, along with including additional abbreviations where needed:
AUX, auxiliary; COM, comitative; CONTR, contrastive; CVB, converb; DAT, dative; DU, dual; DUR, durative; ERG, ergative; GEN, genitive; HS, hearsay; INS, instrumental; NEG, negative; NPST, non-past; PL, plural; POSS, possessive; PST, past; REFL, reflexive; SEQ, sequencer; SG, singular; TEMP, temporal; TOP, topic; X>Y, indicates agent X acting on patient Y

```
<similarities>
	<files>
		<file xml="TDH_KAKCILIP_test.xml" lang="thulung" sound="../audio/Kakcilip.wav"/>
		<file xml="KKT_ORIGIN_test.xml" lang="koyi" sound="../audio/Origin.wav"/>
		<file xml="KHA_KHAKTSALOP_test.xml" lang="khaling" sound="../audio/Khaktsalop.wav"/>
	</files>
	<similarity id="1">
		<color>aliceblue</color>
		<file id="TDH_KAKCILIP_test.xml">
			<sentence id="s1"/>
		</file>
		<file id="KHA_KHAKTSALOP_test.xml">
			<sentence id="s1"/>
		</file>
	</similarity>
	<similarity id="2">
		<color>antiquewhite</color>
		<file id="TDH_KAKCILIP_test.xml">
			<sentence id="s2"/>
		</file>
		<file id="KKT_ORIGIN_test.xml">
			<sentence id="s191"/>
		</file>
		<file id="KHA_KHAKTSALOP_test.xml">
			<sentence id="s2"/>
			<sentence id="s3"/>
			<sentence id="s4"/>
		</file>
	</similarity>
</similarities>
```

Figure 1. Alignment file, generated from a similarity alignment spreadshee

```
<TEXT  xml:lang="x-sil-tdh" id="crdo-TDH_KAKCILIP">
<S id="s1">
   <AUDIO start="1.1704" end="12.0457"/>
     <FORM kindOf="phono">make o dilimʤuŋ u-mam patsoksi u-pap-kam tsɯ-mim</FORM>
   <TRANSL xml:lang="en">Long ago, there were children with a mother, Dilimjung, and a father,
Pachoksi.</TRANSL>
   <W><M><FORM kindOf="phono">make</FORM>
     <TRANSL xml:lang="en">long.ago</TRANSL>
     </M>
   </W>
   <W><M><FORM kindOf="phono">o</FORM>
     <TRANSL xml:lang="en">this</TRANSL>
     </M>
   </W>
   <W><M><FORM kindOf="phono">dilimʤuŋ</FORM>
     <TRANSL xml:lang="en">[name]</TRANSL>
     </M>
   </W> ……..
```

Figure 2. Contents of part of an annotation file

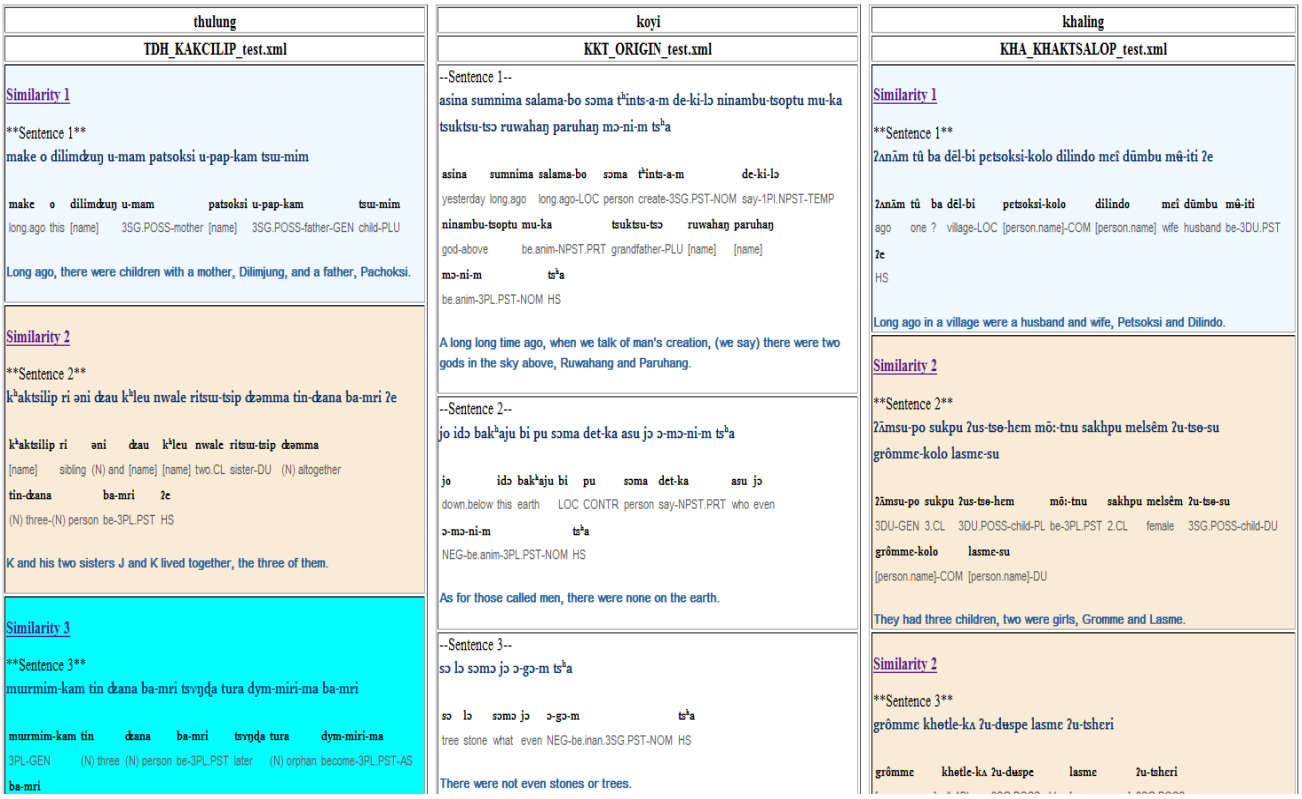
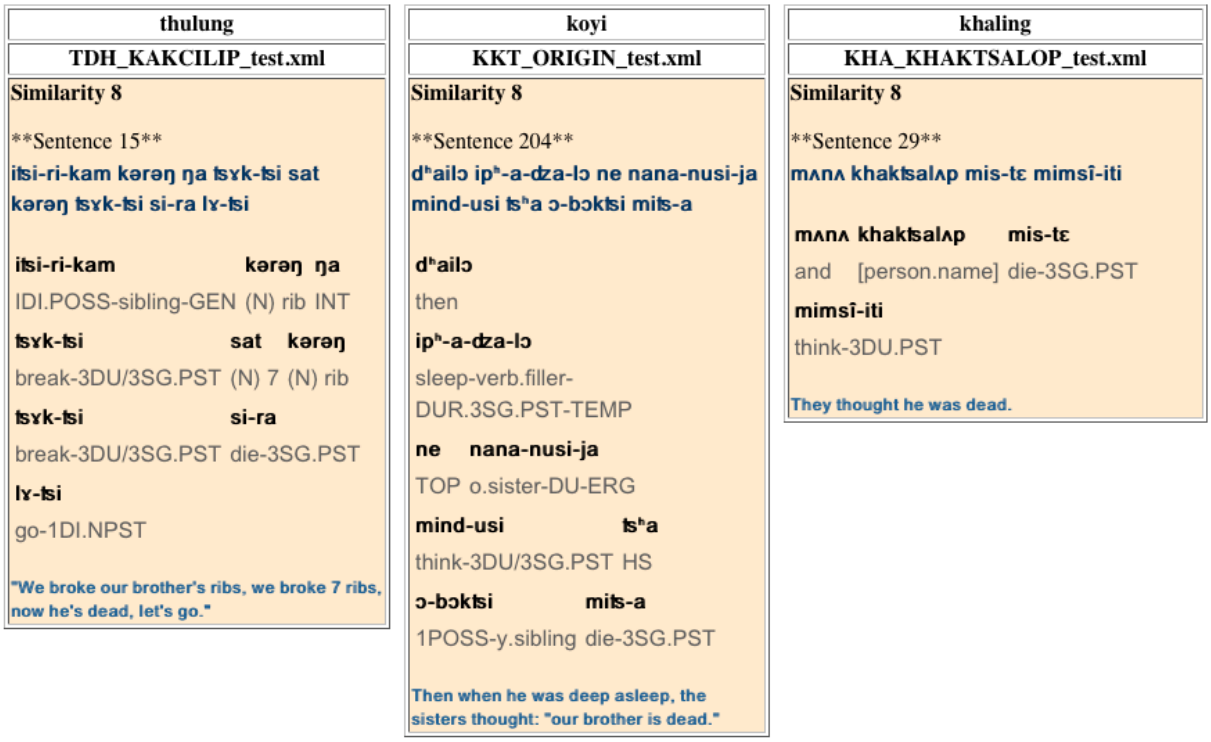


Figure 3. The "integral text view", with each language version of the story in its own column.



thulung	koyi	khaling
TDH_KAKCILIP_test.xml	KKT_ORIGIN_test.xml	KHA_KHAKTSALOP_test.xml
Similarity 8 **Sentence 15** **itsi-ri-kam kərəŋ ŋa tsɤk-tsi sat kərəŋ tsɤk-tsi si-ra lɤ-tsi** **itsi-ri-kam** **kərəŋ ŋa** IDI.POSS-sibling-GEN (N) rib INT **tsɤk-tsi** **sat kərəŋ** break-3DU/3SG.PST (N) 7 (N) rib **tsɤk-tsi** **si-ra** break-3DU/3SG.PST die-3SG.PST **lɤ-tsi** go-1DI.NPST "We broke our brother's ribs, we broke 7 ribs, now he's dead, let's go."	**Similarity 8** **Sentence 204** **dʰailɔ ipʰ-a-ʤa-lɔ ne nana-nusi-ja mind-usi tsʰa ɔ-bɔktsi mits-a** **dʰailɔ** then **ipʰ-a-ʤa-lɔ** sleep-verb.filler-DUR.3SG.PST-TEMP **ne** **nana-nusi-ja** TOP o.sister-DU-ERG **mind-usi** **tsʰa** think-3DU/3SG.PST HS **ɔ-bɔktsi** **mits-a** 1POSS-y.sibling die-3SG.PST Then when he was deep asleep, the sisters thought: "our brother is dead."	**Similarity 8** **Sentence 29** **mʌnʌ khaktsalʌp mis-tɛ mimsî-iti** **mʌnʌ khaktsalʌp** **mis-tɛ** and [person.name] die-3SG.PST **mimsî-iti** think-3DU.PST They thought he was dead.

Figure 4. View of one of the similarities across the three languages, the "similarity view".

Rechercher un terme :

OK

Texte	Phrase	Contexte gauche	Mot	Contexte droit	Gloses
TDH_KAKCILIP_test.xml	s2	ʦɯ mim kʰakʦilip ri əni ʤau kʰleu nwale	ritsɯ	ʦip ʤəmma tin ʤana ba mri ʔe mɯrmim	sister
TDH_KAKCILIP_test.xml	s60	m pəʦʰi meɖɖamma gana reʦa rak ta mɯ	ritsɯ	ʤeɖ ɖy kole ritsɯ bai ra m mɯ	sister
TDH_KAKCILIP_test.xml	s60	reʦa rak ta mɯ ritsɯ ʤeɖ ɖy kole	ritsɯ	bai ra m mɯ ʦɤrʦi lam al pa	sister
TDH_KAKCILIP_test.xml	s61	m mɯ ʦɤrʦi lam al pa luŋ ʦʰəhi	ritsɯ	dys ta memsaka by ry ʔe əni mɯram	sister
TDH_KAKCILIP_test.xml	s61	dys ta memsaka by ry ʔe əni mɯram	ritsɯ	ʦʰəhi ʤeɖ ɖy lo go make ŋa pəhila	sister
KKT_ORIGIN_test.xml	s127	munmuri kʰokʦ a kʰɔnʦ asi dʰanɔ sɔ pu	ʦʰekuma	nusi ʦʰoʔ si umnusi pu kʰur bi buwa	sister
KKT_ORIGIN_test.xml	s147	nɔ mɔ si lɔ ne aʤi ne ana	ʦʰekuma	nusi wɔ oko nɔ kʰur bi dʰoʔo kɔ	sister
KKT_ORIGIN_test.xml	s187	maʔa kim bi ne bɔkʦi ʦɔ mo ni	bigja	nɔ ʦi ʦɔ mo ni kʰɔlɔ mo ni	sister
KKT_ORIGIN_test.xml	s191	luʔ nɔ pʰiŋ u ʦʰa aʤi sɔma ʦɔ	bigjame	ʦɔ made mɔ ni m ʦʰɔ kim bi	sister
KKT_ORIGIN_test.xml	s278	lu ʦʰa naga ne hɔ inʦi sɔlɔ inʦi	ʦʰekuma	inʦi ʦi sɔlɔme ʦʰɔʔ ɔ kɔpɔ sɔma ʦɔ	sister
KKT_ORIGIN_test.xml	s334	nɔ ne heʔnɔ mɔ ʤa si ʦʰa ɔbɔ	ʦʰekuma	hoʔle dʰaiʔnɔ ŋɔ umbika sɔksu pu ana nuwa	sister
KKT_ORIGIN_test.xml	s345	ase papa luʔ si heʔŋɔ dʰam bi ne	ʦʰekuma	ʦi ʦʰɔm ka ʦʰuʔ mu ʦʰɔ ʦʰɔm ka	sister
KKT_ORIGIN_test.xml	s346	ʦʰɔm ka ʦʰuʔ mu ʦʰɔ ʦʰɔm ka ʦʰekumɔ	ʦʰekuma	ʦʰuʔ mu ʦʰɔʔ lɔ ne kɔpa mo a	sister
KKT_ORIGIN_test.xml	s347	ʦʰuʔ mu ʦʰɔʔ lɔ ne kɔpa mo a	ʦʰekuma	ʦi dja nɔ ʦjuri mu di pʰiŋ usi	sister
KKT_ORIGIN_test.xml	s350	se si ʦʰa ʦʰɔmdam bi ʦʰɔm ka ʦʰe	ʦʰekuma	ʦi ʦʰaŋgara pʰiŋ usi ʦʰa ʦʰaŋgara jɔ ja	sister
KKT_ORIGIN_test.xml	s353	a nɔ aʤi pi wa pu dja ʦʰa	ʦʰekuma	nɔ bo hɔbaʦ a nɔ kʰɔkʦulupa dja ʦʰa	sister
KKT_ORIGIN_test.xml	s362	si m ʦelbu nɔ kim hoʔle si dʰaiʔlɔ	bigjame	nusi ʦʰɔm ka bi pu risi pik uni	sister
KHA_KHAKTSALOP_test.xml	s125	ʔɛˆn ʤɵ jo ŋa mɛˆm ʦʉnʉ hɛm ʔɛ	bèŋmɛ	ʦɵ hɛm ʔɛ was ʦhoɔ¯m nu lo ʔathā:	sister

Figure 5. Concordance of the gloss "sister"

MultiMASC: An Open Linguistic Infrastructure for Language Research

Nancy Ide

Department of Computer Science
Vassar College, USA
ide@cs.vassar.edu

Abstract

This paper describes MultiMASC, which builds upon the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008; Ide et al., 2010) project, a community-based collaborative effort to create, annotate, and validate linguistic data and annotations on a broad-genre open language data. MultiMASC will extend MASC to include comparable corpora in other languages that not only represent the same genres and styles, but also include similar types and number of annotations represented in a common format. Like MASC, MultiMASC will contain only completely open data, and will rely on a collaborative community-based effort for its development. We describe the possible ways in which additional corpora for MultiMASC can be collected and annotated and consider the dimensions along which "comparability" for MultiMASC corpora can be determined. Because it is unlikely that all language-specific MultiMASC corpora can be comparable along every dimension, we also outline the measures that can be used to gauge comparability for a number of different criteria.

Keywords: Comparable corpora, Corpus construction, Multi-lingual resources

1. Introduction

In an ideal universe, computational linguistics researchers would have open access to very large language corpora spanning the full range of genres, registers, and languages, all of which would be accompanied by high quality annotations for linguistic phenomena at all levels that can be used to support machine learning and computational linguistics research in general. Parallel data would exist for all languages, and common lexical, semantic, and discourse-level phenomena would be linked across data of all genres and languages. Annotations would come with detailed information about provenance as well as evaluation metrics in order to ensure quality, and researchers could easily request specific data and annotations to be delivered as needed over the web, in a physical format and using "annotation semantics" that can be integrated without modification into their own tools and resources. Unfortunately, this scenario is a long way off, and the greatest obstacle is the high cost of high-quality resource production and maintenance. Another obstacle is the difficulty of obtaining language data representing a variety of genres that is unfettered by licensing constraints so that it may be used for any purpose community-wide.

The Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008; Ide et al., 2010) project attempts to overcome these obstacles to high-quality resource creation through a community-based collaborative effort to create, annotate, and validate linguistic data and annotations on broad-genre open language data. MASC is a half million word corpus of contemporary American English language data drawn from the 15 million word Open American National Corpus (OANC)[1] that includes manually produced or validated annotations for a wide range of linguistic phenomena at all linguistic levels. The corpus includes a balanced set of nineteen genres of spoken and written language data that is completely open for any use. The corpus is freely downloable from the MASC website, as well as through the Linguistic Data Consortium (LDC)[2]. All MASC annotations are represented in a common format so that they may be used collectively to study intra-level interactions, which are important for the deeper analyses that are increasingly the focus in the field.

This paper describes MultiMASC, which builds upon the MASC project by extending MASC to include comparable corpora in other languages. Here, "comparable" means not only representing the same genres and styles, but also include similar types and number of annotations represented in a common format. Like MASC, MultiMASC will contain only completely open data and rely on a collaborative community-based effort for its development.

We first describe MASC as it currently exists, as well as plans for its future development. The remainder of the paper describes the possible ways in which additional corpora for MultiMASC can be collected and annotated. We then consider the dimensions along which "comparability" for MultiMASC corpora can be determined, and, because it is unlikely that all language-specific MultiMASC corpora can be comparable along every dimension, we outline the measures that can be used to gauge comparability for a number of different criteria.

2. MASC

MASC is the only corpus with multiple layers of annotations in a common format that can be used either individually or together, and (unlike, for example, OntoNotes) to which others can add annotations. MASC will be soon increased in size to a million words, although there are currently no resources for further in-house validation; we will depend on the community to validate and contribute annotations to fill in the gap.

[1] http://www.anc.org/OANC

[2] http://www.ldc.upenn.edu

MASC currently contains nineteen genres of spoken and written language data in roughly equal amounts, shown in Table 1. Approximately 15% of the corpus consists of spoken transcripts, both formal (court and debate transcripts) and informal (face-to-face, telephone conversation, etc.); the remaining 85% covers a wide range of written genres, including emerging social media genres (tweets, blogs). Because it is drawn from the OANC, all MASC data represents contemporary American English produced since 1990. The entire MASC is annotated for logical structure, token and sentence boundaries, part of speech and lemma, shallow parse (noun and verb chunks), named entities (person, location, organization, date), and Penn Treebank syntax. Portions of MASC are also annotated for additional phenomena, including 40K of full-text FrameNet frame element annotations and PropBank, TimeML, and opinion annotations over a roughly 50K subset of the data. As the name of the corpus implies, all annotations have either been manually produced or automatically produced and hand-validated. The list of annotation types and coverage is given in Table 2.

MASC also includes sense-tags for 1000 occurrences of each of 100 words chosen by the WordNet and FrameNet teams (100,000 annotated occurrences), described in (Passonneau et al., 2012). The sense-tagged data are distributed as a separate *sentence corpus* with links to the original documents in which they appear. Where MASC does not contain 1000 occurrences of a given word, additional sentences were drawn from the OANC. Several inter-annotator agreement studies and resulting statistics have been published (Passonneau et al., 2009; Passonneau et al., 2010), many of which are distributed with the corpus.

Genre	No. files	No. words	Pct corpus
Court transcript	2	30052	6%
Debate transcript	2	32325	6%
Email	78	27642	6%
Essay	7	25590	5%
Fiction	5	31518	6%
Gov't documents	5	24578	5%
Journal	10	25635	5%
Letters	40	23325	5%
Newspaper	41	23545	5%
Non-fiction	4	25182	5%
Spoken	11	25783	5%
Technical	8	27895	6%
Travel guides	7	26708	5%
Twitter	2	24180	5%
Blog	21	28199	6%
Ficlets	5	26299	5%
Movie script	2	28240	6%
Spam	110	23490	5%
Jokes	16	26582	5%
TOTAL	376	506768	

Table 1: Genre distribution in MASC

All MASC annotations are represented in the ISO TC37 SC4 Linguistic Annotation Framework (LAF) GrAF format (Ide and Suderman, 2007; Ide and Suderman, Submitted), with the objective to make the annotations as flexible for use with common tools and frameworks as possible. The ANC project provides a web application, called ANC2Go[3] that enables a user to choose any portion or all of MASC and the OANC together with any of their annotations to create a "customized corpus" that can be delivered in any of several widely used formats such as CONLL IOB, RDF, inline XML, etc. Modules to transduce GrAF to formats consistent with other tools and frameworks such as UIMA, GATE, and NLTK are also provided.[4] Thus "openness" in MASC applies to not only acquisition and use, but also interoperability with diverse software and systems for searching, processing, and enhancing the corpus.

Annotation type	No. words
Logical	506659
Token	506659
Sentence	506659
POS/lemma (GATE)	506659
POS (Penn)	506659
Noun chunks	506659
Verb chunks	506659
Named Entities	506659
FrameNet	39160
Penn Treebank	*506659
PropBank	55599
Opinion	51243
TimeBank	*55599
Committed Belief	4614
Event	4614
Dependency treebank	5434

* under development

Table 2: Summary of MASC annotations

3. MultiMASC

MultiMASC will both expand MASC and the collaboration effort upon which it depends and exploit the infrastructure and experience that the development of MASC has provided. The eventual result will be a massive, multilingual, multi-genre corpus with comparable multilayered annotations that are inter-linked via reference to the original MASC, as shown in Figure 1.

We see the development of MultiMASC as an incremental process, involving the following steps for any given language:

1. Create and make available a corpus of open language data, comparable in size and genre distribution to MASC.
2. Collect and make available annotations for linguistic phenomena comparable to, and possibly extending beyond, those available for MASC, either automatically or manually produced, in any format.
3. Validate the automatically-produced annotations.
4. Provide the annotations in a format compatible with MASC and other MultiMASC annotations.

[3] http://www.anc.org:8080/ANC2Go/

[4] http://www.anc.org/tools/

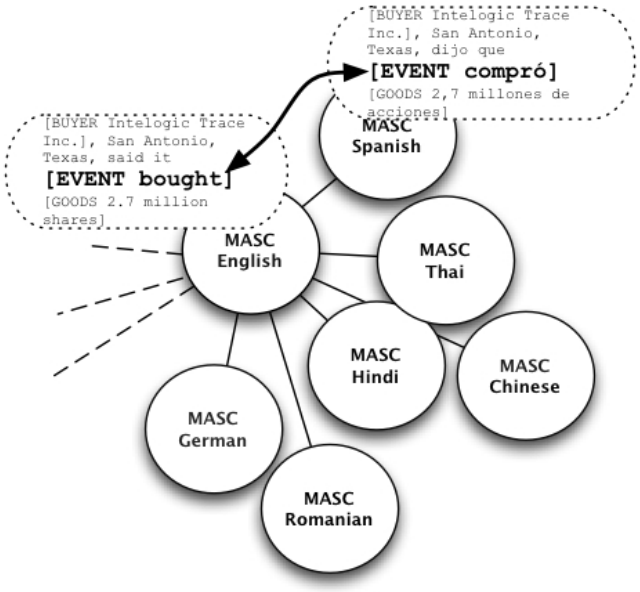


Figure 1: Overview of MultiMASC

5. Provide linkage among annotations in the language-specific data and MASC annotations, and/or annotations in other MultiMASC corpora as appropriate.

Given the expected constraints of funding and resources, we anticipate that for some languages, interim results will be all that is available at any given point in development, or, possibly, that interim results are all that ever becomes available. Even if this is the case, the comparable MultiMASC corpora created in step 1 will provide a resource for computational linguistics research and development that is unmatched at present.

4. Step one: Data gathering

The first step in the creation of MultiMASC is to produce a massive multi-lingual corpus of language-specific data with comparable genre distribution that is open and freely available for community use. "Open" in OANC/MASC terms means that data is either in the public domain or under a license that does not restrict redistribution of the data or its use for any purpose, including commercial use (e.g., the Creative Commons Attribution license[5]). Data under licenses such as GNU General Public License[6] or Creative Commons Attribution-ShareAlike[7] are avoided because of the potential obstacles to use for commercial purposes imposed by the requirement to redistribute under the same terms.

Comparable MultiMASC component corpora will need approximately 25,000 words of open data for each of the nineteen MASC genres, produced by native speakers of the language in question (no translations) after 1989. Fortunately, experience shows us that obtaining and preparing samples of this size is considerably easier than for larger amounts of data, which will hopefully make the prospect of constructing a language-specific portion of MultiMASC less daunting for potential contributors.

4.1. Obtaining open data

The OANC/MASC project has long been identifying and gathering open data for inclusion in both the OANC and MASC. The following are some of the sources and strategies we have utilized:

1. Contributions from publishers who are willing to provide data under a non-restrictive license, as is the case for the OANC/MASC non-fiction materials donated by Oxford University Press and Cambridge University Press, and SLATE magazine articles from Microsoft. To protect their interests, the publishers sometimes provided only a subset of a complete book or collection.

2. Web search for materials in the public domain. Government documents and debate and court transcripts, as well as technical articles in collections such as Biomed and PLOS, are typically in the public domain, for example.

3. Web search for data licensed under non-viral licenses such as CC-BY. Blogs, fiction, and other writing such as essays are very often distributed over the web under these terms.

4. Contributions from college students of class essays and other writing. College students produce considerable volumes of prose during their academic careers, and very often this data is discarded or forgotten once handed in to satisfy an assignment. The OANC site provides a web interface for contributions of this kind

[5] http://creativecommons.org/licenses/by/3.0/
[6] http://www.gnu.org/licenses/gpl.html
[7] http://creativecommons.org/licenses/by-sa/2.5/

that includes a grant of permission to use the contributed materials. We regularly solicit these contributions from students in our own and other institutions.

5. Contributions of data from colleagues in the field. We have received data contributions, including significant amounts of spoken data, from several NLP and linguistics projects. As awareness of the need for open data increases, such contributions should become easier to obtain.

6. Direct solicitation for use of web materials. We have on occasion identified a web site containing interesting or substantial materials and contacted the relevant parties directly to explain our use of the data and ask for permission to use it. We have also contacted providers whose data are freely available for access to the materials in a form more manageable for processing purposes. So far, none of our requests has been turned down.

Different languages, as well as different countries and therefore different copyright laws, will affect the ease with which MultiMASC data can be acquired in any given case. To the extent that it applies, the experience of the MASC project can be relied upon as a resource to support the acquisition of MultiMASC data.

4.2. Identifying comparable data

The definition of "comparable" as it applies to genre is, of course, not exact. The best guideline to determine comparability may be to consider the primary uses to which MultiMASC will be put, including the extraction and/or linkage of parallel segments and paraphrases; semantic frame elements; translations of single words, multi-word expressions, proper names, and named entities; etc., in order to facilitate inter-linguistic discoveries and comparisons. To address this, we can identify several dimensions along which to measure cross-lingual comparability, including structural complexity; lexical richness and specificity; vocabulary register; temporal organization (tense and aspect); referential cohesion; interactiveness; and others (see, for example, the measures outlined in (Biber, 1995)).

Statistics characterizing these dimensions (e.g., simple measures such as type/token ratio, word and sentence length, together with metrics indicating the degree of use of linguistic features such as private verbs, suasive verbs, time and place adverbials, subordination, third person pronouns, proper nouns, and many more), which are available for MASC data, may provide a point of departure for determining comparability. However, more research into this possibility will be required to determine exactly what the best among such measures may be, and, more critically, how the measures may or may not apply depending on the language in question.

Beyond comparability on the basis of metrics like these, we may also consider comparability in terms of topic, that is, data that treats the same or a closely related topic as the original MASC document. One possibility is to consider a continuum of comparability, starting with the most general: same domain (e.g. finance), same topic (e.g., investment), same sub-topic (e.g., 401K accounts), same subject (e.g., report or description of same event, etc.).

4.3. Preparing the data

The ANC project has extensive experience in preparing data that is obtained in any of several formats for use by annotation tools. This experience can be exploited by developers of MultiMASC component corpora in order to make the data preparation process easier, if not entirely trivial. For example, we have an automatic pipeline for processing documents originally in Microsoft Word, Open Office (odt), or Rich Text Format (rtf) that generates a UTF-8 file containing the text content together with standoff annotations for logical structure down to the level of paragraph. The annotations can be automatically rendered in any of several possible output formats, including GrAF.

The ANC project has also developed several modules for the General Architecture for Text Engineering (GATE)[8] to import from and export to GrAF, so that annotations generated within GATE can be immediately rendered in the MASC common format. GATE includes annotation modules for a fairly extensive range of languages, which means that in some cases, generating automatically-produced annotations for MultiMASC in GrAF will be trivial. We have also developed similar GrAF import/export modules for the UIMA annotation framework.

5. Step two: Annotation

Getting the MultiMASC data in place for as many languages as possible provides the base for a community effort to annotate the data. For major languages, it should be relatively easy to obtain automatically-produced annotations comparable to the basic MASC annotations: sentence and token boundaries, at least one part-of-speech/lemma analysis, shallow parse (noun and verb chunks), syntactic phrase structure (trees), and basic named entities (person, organization, location, date).

Validation of the annotations is a much more costly and time-intensive venture. MASC validation has so far been done in-house by trained validators; however, this may not always be feasible, and it is therefore expected that for MultiMASC, considerably more community-based collaboration may be required. The range of possibilities include, at one end, simply publishing the data and unvalidated annotations for community use, with the request that those who use the data contribute any correction or additional annotation they perform.[9] At the other extreme, a sophisticated web-based interface could be provided so that others can directly validate the data, which would track and evaluate annotations as they are produced, use active learning to suggest possible corrections, etc. Crowdsourcing, with or without a sophisticated interface, provides another alternative.

Beyond the types of annotation included (e.g., part-of-speech, named entities, etc.), annotations will ideally be comparable in terms of *syntactic interoperability*, i.e.,the physical format in which they are represented e.g., inline

[8] http://gate.ac.uk

[9] This is the strategy used for the OANC.

vs.standoff annotation, XML, Penn Treebank-like bracketing, etc.). To ensure that all annotations on all language data are usable together and/or with the same tools, annotations can be rendered in the common format used by MASC (LAF/GrAF), or in a format that is trivially mapped to GrAF.

Semantic interoperability among annotations, which involves the actual categories and features used to describe the various linguistic phenomena, is far more difficult to achieve. Clearly, the use of common annotation categories among MultiMASC corpora is not feasible, given that most annotations will first be produced using existing software, and re-tooling existing software to accommodate specific annotation categories (even if it were possible to specify a definitive set that would accommodate all languages and linguistic theories) is unrealistic. Efforts such as ISOCat[10], which attempt to provide ways to map semantic categories and, where this is not possible, specify their differences, are underway. This may enable a greater degree of semantic interoperability among MultiMASC corpora, but such efforts are not expected to be well enough along in the next few years to provide a comprehensive solution. The best measure of comparability that may be possible in the near term might be an indication of the "mappability" between two schemes on a rough scale of difficulty (trivial, medium, hard, unmappable). Ideally, where possible, mappings between schemes for like annotation types among languages would be developed and distributed from the MultiMASC home website.

6. Step three: Creating the inter-linked MultiMASC

The final step in creating MultiMASC will be to link like annotations across languages. We envision linkage among linguistic phenomena at many levels, e.g., part-of-speech categories, syntactic structures, paraphrases, semantic roles, named entities, events, etc.

Linkage among the MultiMASC corpora can be accomplished in at least two ways. First, MASC can be used as a "hub", as depicted in Figure 1, to which annotations of the same phenomenon (a "buy" event in the figure) are directly linked.[11] We anticipate that MultiMASC corpora will be represented in GrAF or a format that is trivially mappable to GrAF. Inter-linkage is then straightforward: an attribute can be added to the XML element for an annotation in a MultiMASC corpus that refers to a corresponding annotation in the American English MASC.

A more elegant and workable solution for inter-linkage among MultiMASC corpora would utilize a reference set of categories, possibly represented in RDF/OWL (for example, resources included in the Linguistic Linked Open Data cloud[12]) and/or residing in a data category registry such as ISOCat[13]. In this scenario, annotations in both MASC and other MultiMASC corpora are linked to an independent entity on the web that provides information about the annotation content, as depicted in Figure 2. For example, a "noun plural" part-of-speech annotation in MultiMASC corpora could include a reference to the PID (persistent identifier) in the ISOCat registry that defines this category. In GrAF, such a reference could look like this:[14]

```
<a label="Token" ref="ann-n3" as="xces">
 <fs>
  <f name="msd" value="...DC-3581"/>
  ...
```

Linkage of this nature will enable cross-linguistic and inter-layer studies on a scale that is currently impossible. Available multi-lingual data from sources such as Wikipedia does not include the layers of annotation we envision for MultiMASC, and Wikipedia data is not completely open due to the restriction to "share-alike". The recently launched Language Library effort[15] includes multiple annotations, but it includes only a handful of materials, most also under "share-alike" constraints, and there is no effort to provide annotations in compatible formats or to inter-link them.

7. Comparability Index

We seek to identify measures of comparability along the several dimensions outlined above that can be used both as a guidelines for the construction of MultiMASC corpora in other languages and as a gauge of comparability for these corpora once they become a part of MultiMASC. The latter is important because we cannot expect that it will be possible in all or even most cases to conform to a strict set of comparability guidelines; with these measures, users will have information that can inform cross-lingual studies that use the MultiMASC data.

Table 3 shows the various dimensions of comparability and an overview of the measures that will be defined to classify them. Note that in principle, all measures apply to the entire language-specific corpus except for DOMAIN/TOPIC/SUBJECT, which will in most cases apply to individual documents or groups of documents within a specific genre. We can envision ultimately providing a very large matrix giving pair-wise comparability indexes for all languages in MultiMASC.

8. Conclusion

A community-wide, collaborative effort to produce high quality annotated corpora is one of the very few possible ways to address the high costs of resource production, and to ensure that the entire community, including large teams as well as individual researchers, has access and means to use these resources in their work. The OANC and MASC already lay the groundwork for such an effort for English, and extending it to other languages seems a logical next step.

[10] http://www.isocat.org

[11] Note that the use of MASC as a hub does not preclude linkage among other language pairs.

[12] http://linguistics.okfn.org/resources/llod/

[13] http://www.isocat.org

[14] Due to space limitations the ISOcat URI prefix http://www.isocat.org/datcat has been replaced by ellipses.

[15] http://www.languagelibrary.eu

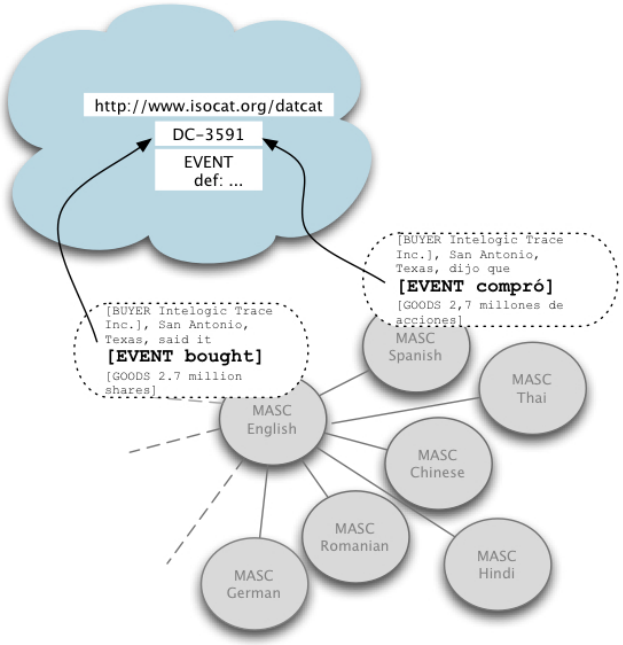


Figure 2: Linked annotations in MultiMASC

Dimension	Information
GENRE	Category among MASC genres
	Comparison measures for each genre, including broad dimensions such as structural complexity, lexical richness and specificity, vocabulary register, etc., with relevant statistics for specific measures (type/token ratio, subordination, use of specific verb types, etc.)
~~DOMAIN/TOPIC/SUBJECT~~*	Continuum along comparability of domain, topic, (one or more) sub-topics, subject
ANNOTATIONS	Comparison with original MASC annotations in terms of the annotation types included, categories provided for each annotation type
	Comparison with annotations included in other language corpora in MultiMASC
	Format, in terms of mappability to a common format or format directly usable with other language corpora in MultiMASC
	Semantics, in terms of conformance or mappability to those in other language corpora in MultiMASC
INTER-LINKAGE	Number and type of inter-linked phenomena

* Applies to individual documents

Table 3: Comparability measures for MultiMASC

The vision of a MultiMASC for a large number and wide variety of languages is to some extent "pie-in-the-sky", as it is certain to take many years to accomplish. Therefore, in order to keep the project within realistic bounds, the plan is to develop MultiMASC opportunistically, incorporating language-specific corpora as they become available and adding annotations and linkages later, if necessary. This way, the community can use and enhance data and annotations as they become available in an extended effort that will hopefully build momentum as the possibilities MultiMASC offers for research become increasingly apparent.

Acknowledgments

This work was supported by National Science Foundation grants CRI-0708952 and CRI-1059312.

9. References

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the First Linguistic Annotation Workshop*, pages 1–8, Prague.

Nancy Ide and Keith Suderman. Submitted. The Linguistic Annotation Framework: A Standard for Annotation In-

terchange and Merging. *Language Resources and Evaluation*.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Paris. European Language Resources Association.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus : A community resource for and by the people. In *Proceedings of the The 48th Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ, USA, July. Association for Computational Linguistics.

Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *SEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 2–9, Morristown, NJ, USA. Association for Computational Linguistics.

Rebecca Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC)*, Paris. European Language Resources Association.

Rebecca J. Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The masc word sense sentence corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Paris. European Language Resources Association.

Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for English-Romanian language pair

Elena Irimia
Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie, No 13
elena@racai.ro

Abstract

The paper describes a tool developed in the context of the ACCURAT project (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation). The purpose of the tool is to extract bilingual lexical dictionaries (word-to-word) from comparable corpora which do not have to be aligned at any level (document, paragraph, etc.) The method implemented in this tool is introduced by (Rapp, 1999). The application basically counts word co-occurrences between unknown words in the comparable corpora and known words from a Moses extracted general domain translation table (the base lexicon). We adapted the algorithm to work with polysemous entries in the translation table, a very frequent situation which is not treated in the standard approach. We introduced other heuristics, like 1. filtration of the context vectors according to a log likelihood threshold, 2. lists of verbs (specific to each language) that can be main verbs but also auxiliary or modal verbs; 3) a cognate heuristic based on the Levenshtein Distance. The implementation can also run in multithreading mode, if the user's machine has the capacity to enable parallel execution.

1. Introduction

The task of extracting translation equivalents from bilingual corpora has been approached in different manners, according to the degree of parallelism between the source and target parts of the corpora involved. For a well sentence aligned parallel corpora one can benefit from reducing the search space for a candidate translation to the sentence dimension and external dictionaries are not required. In the case of comparable corpora, the lack of aligned segments can be compensated by external dictionaries (Rapp, 1999) or by finding meaningful bilingual anchors within the corpus based on lexico-syntactic information previously extracted from small parallel texts (Gamallo, 2007).

The word alignment of parallel corpora has been received significant scientific interest and effort starting with the seminal paper of Brown et al. (1990) and continuing with important contributions like Gale & Church (1993), Kay & Roscheisen (1993), Och, F.J. et al. (1999), etc. and many more recent approaches. They are already various free software aligners used in the industry and research, from which we mention only the famous GIZA++ (Och and Ney, 2003). Moreover, the error rate goes down to 9% in experiments made with some of these approaches (Och and Ney, 2003). By comparison, the efforts and results in extracting bilingual dictionaries from comparable corpora are much poorer. Most of the experiments are usually done on small test sets, containing words with high frequency in the corpora (>99) and the accuracy percentages are not rising above 65%.

The most popular method to extract word translations from comparable corpora, on which we based the construction of our tool, is described and used by Fung & McKeown (1997), Rapp, (1999), Chiao & Zweigenbaum, (2002). It relies on external dictionaries and is based on the following hypothesis:

word target1 is a candidate translation of word source1 if the words with which target1 co-occur within a particular window in the target corpus are translations of the words with which source1 co-occurs within the same window in the source corpus.

The translation correspondences between the words in the window are extracted from external dictionaries, being seen as *seed* word pairs. In the following table, we present, in the context of the corpus we worked on (see section 4.1), the words with which "level" tend to co-occure in the English part with their specific log-likelyhoods (ex. left column, "high level" with LL 335.0537) and the words with which a possible translation of "level", e.g. "nivelul" tend to co-occure in the Romanian part. The words in the columns are ordered so as the word in the right column on a specific line it is a possible translation of the word in the left column on the same line. (e.g.: said = anunțat, low = scăzut, mic, etc.)

level	**nivelul**
high*335.0537	ridicat*108.0321
said*111.74	anunțat*10.0774
low*110.9197	scăzut*29.3577, mic*20.6037
years*86.9735	an*16.5761
fell*83.3033	scăzut*29.3577
current*77.2435	actual*48.8756
rate*63.3928	rata*12.5533

Table 1. The words with which "nivelul" co-occurs in the Romanian corpus within a certain window (here, of length 5), listed in the right column, are translations of the words with which "level" co-occurs in the English corpus within the same window, listed in the left column.

Gamallo & Pichel (2005) used as seed expressions pairs of bilingual lexico-syntactic templates previously extracted from small samples of parallel corpus. This strategy led to a context-based approach, reducing the searching space from all the target lemmas in the corpus to all the target lemmas that appear in the same seed templates. In the improved version of the approach (Gamallo, 2007), the *precision-1* (the number of times a correct translation candidate of the test word is ranked first, divided by the number of test words) and *precision-10* (the number of correct candidates appearing in the top 10, divided by the number of test words) scores go up to 0.73 and 0.87 respectively.

In the following we will describe the algorithm implemented by our tool as introduced by Rapp (1999) and we will highlight the modifications and the adaptations we made, based on the experimental work we conducted. In Section 2 we present the original approach of Rapp, Section 3 describes our contribution to the improvement of the algorithm in the tool creation's process and Section 4 introduces the results of the experiments done on 3 types of comparable corpora.

2. Short presentation of the original approach

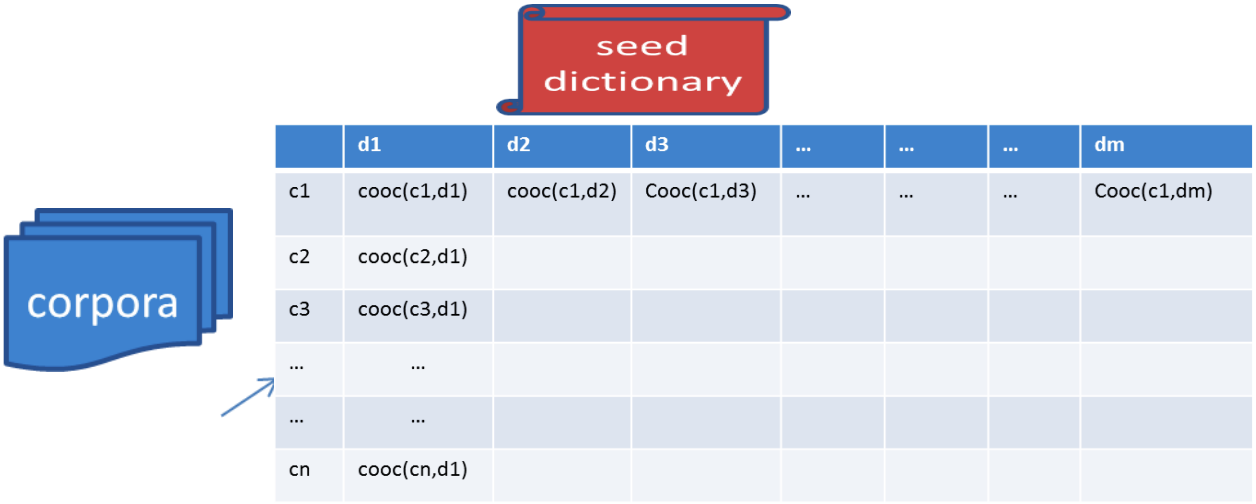
In a previous study, Rapp (1995) had already proposed a new criterion (the co-occurrence clue) for word alignment appropriate for non-parallel corpora. The assumption was that "there is a correlation between co-occurrence patterns in different languages" and he demonstrated by a study that this assumption is valid even for unrelated texts in the case of English-German language pair.

Starting from a more or less small seed dictionary and with the purpose of extending it based on a comparable corpus, a co-occurrence matrix is computed both for the source corpus and for the target corpus. Every row in the matrix corresponds to a type word in the corpus and every column corresponds to a type word in the base lexicon. For example, the intersection of a row i and a column j in the co-occurrence matrix of the source corpus contains a value sourcecooc(i,j) = frequency of common occurrence of word i and word j in a window of pre-defined size (see Figure 1 for a graphic of a generic co-occurence matrix).

The target and source corpus are lemmatized and POS-tagged and function words are not taken in consideration for translation (they are identified by their POS closed class tags: pronouns, prepositions, conjunctions, auxiliary verbs, etc.).
For any row in the source matrix, all the words with which the co-occurence frequency is above 0 are sent for translation to the seed lexicon. The unknown words (absent in the lexicon) are discarded and a vector of co-occurrences for the word correspondent to each row is computed versus the list of the translated words remained.

Experiments conducted to the need of replacing the co-occurence frequency in the co-occurence vectors by measures able to eliminate word-frequency effects and favor significant word pairs. Measures with this purpose were previously based on mutual information (Church & Hanks, 1989), conditional probabilities (Rapp, 1996), or on some standard statistical tests, such as the chi-square test or the log-likelihood ratio (Dunning, 1993). In the approach we based our tool on, the measure chosen was the log-likelihood ratio.

Finally, similarity scores are computed between all the source vectors and all the target vectors computed in the previous step, thus setting translation correspondences between the most similar source and target vectors. Different similarity scores were used in the variants of this approach; see (Gamallo, 2008) for a discussion about the efficiency of several similarity metrics combined with two weighting schemes: simple occurrences and log likelihood. Another related study was made by Laroche & Langlais (2010) which is presenting experiments around more different parameters like context, association measure, similarity measure, seed lexicon.



	d1	d2	d3	...	...	...	dm
c1	cooc(c1,d1)	cooc(c1,d2)	Cooc(c1,d3)	...	...	...	Cooc(c1,dm)
c2	cooc(c2,d1)						
c3	cooc(c3,d1)						
...	...						
...	...						
cn	cooc(cn,d1)						

Figure 1 A generic co-occurrence matrix

3. Our approach

3.1 Adaptations of Rapp's algorithm

With the aim of obtaining a dictionary similar to a translation table of the type a decoder like Moses would need to produce its translation, we decided that the lines and columns of the matrixes will be populated in our approach by word forms and not by lemmas, as in the standard approach. The option for lemma entries in the matrix was assumed also by works like (Gamallo & Pichel, 2005) and (Gamallo, 2008).

As the purpose of this tool (and of all the other tools in the ACCURAT project) was to extract from comparable corpora data that would enrich the information already available from parallel corpora, it seemed reasonable to focus (just like Rapp(1999) did) on the open class (versus closed class) words. Because in many languages, the auxiliary and modal verbs can also be main verbs, frequently basic concepts in the language (see "be" or "have" in English), and most often the POS-taggers don't discriminate correctly between the two roles, we decided to eliminate their main verb occurrences as well. For this purpose, the user is asked to provide a list of all these types with all their forms in the languages of interest (parameters: sourceamverblist, targetamverblist).

We gave the user the possibility to specify the length of the text window in which co-occurrences are counted by modifying a parameter in the configuration file. As our experiment conducted to good results for a text window of length five, this is the default value of the parameter.

Being based on word counting, the method is sensitive to the frequency of the words: the higher the frequency, the better the performance. In previous works, the evaluation protocol was conducted on frequent words, usually on those with the frequency above 100. Even in works like (Gamallo, 2008), where the evaluation was made on a list of nouns whose recall was 90% (those nouns that together come to the 90% of noun tokens in the training corpus), this corresponded to a bilingual lexicon constituted by 1,641 noun lemmas, each lemma having a token frequency>=103, for a bilingual comparable corpus of around 15 million tokens for each part. It doesn't seem too efficient to extract only a small amount of tokens from a big size corpus. Therefore, even if it causes loss of precision, the frequency threshold must be lowered when we are interested in extracting more data. In our tool, this parameter can be set by the user, according to his/her needs, but it should be above 3 (our minimal threshold) and it should take into account the corpus dimension.

As we mentioned in the previous section, the polysemy in the seed lexicon is not discussed in the standard approach. Other approaches either keep for reference only the first translation candidate in the dictionary or give different weights to the possible translations according to their frequencies in the target corpus (Morin et al., 2007).

Our seed lexicon is based on a general domain translation table automatically extracted (with GIZA++) and this is consistent with the idea that we want to improve translation data obtained from parallel corpora. But as a consequence, we deal with high ambiguity and erroneous data in the seed lexicon. In the Table 2 you can see an excerpt from the base lexicon displaying all the possible translation for the word form "creates" with their translation probabilities. Only the first three entries are exact translations of the word form "creates" while 3 of them ("instituie", "stabilește" and, in a lesser extent, „ridică" are acceptable translations in certain contexts). The two bold entries, „naștere" (birth) and „duce" (carries), may seem wrong translations learned from the training data, having a translation probability score similar to some correct translation (like "creând" or „crea"), but they also can be acceptable translations in certain contexts. We think we need to have access to all these possible translations as the semantic content of a linguistic construction is rarely expressed in another language through an identical syntactic or lexical structure. This is true especially in the case of a comparable corpus.

Our solution was to distribute the log-likelihood of a word pair (w1, w2) in the source language to all the possible translations of w2 in the target language as follows:

$$LL(w_1, w_2) = \sum_i LL(w_1, w_2) * p(w_2, t_i)$$

where $p(w_2, t_i)$ is the probability of a word w_2 to be translated with t_i and $\sum_i p(w_2, t_i) = 1$.

Every translation pair *(w_2, t_i)* is identified in the base lexicon by an unique id, making it possible to compute a similarity score across the languages.

id	*word*	*translation*	*transl. prob.*	*LL distribution*
72083	creates	creea	0.0196	LL(man,72083) =12*0.0196078 =0.2352936
72084	creates	creează	0.6862	LL(man,72084) =12*0.686275 =8.2353
72085	creates	creând	0.0196	LL(man,72085) =12*0.0196078 =0.2352936
72086	creates	**duce**	0.0196	LL(man,72086) =12*0.0196078 =0.2352936
72087	creates	instituie	0.1176	LL(man,72086) =12*0.117647 =1.411764
72088	creates	**naștere**	0.0196	LL(man,72086)

				=12*0.0196078 = 0.2352936
72089	creates	ridică	0.0392	LL(man,72089) =12*0.0392157 = 0.4705884
72090	creates	stabileşte	0.0196	LL(man,72086) =12*0.0196078 = 0.2352936

Table 2: An excerpt from the base lexicon with the possible candidate translations of the word „creates" and the distribution of LL(man, creates) = 12 according to the translation probabilities of the candidates

Previous to the LLs distribution, there is a step of LL filtering, in which all the words that occur with an LL smaller than a threshold are eliminated (the threshold is set by the ll parameter in the configuration file). This was motivated by the need to reduce the space and time computational costs and is also justified by the intuition that not all the words that occur at a specific moment together with another word are significant in the general context of our approach and the LL score is a good measure of this significance.
Following the conclusions of Gamallo's (2008) experiments, we used as a vector similarity measure the DiceMin function.

In computing the similarity scores, we did not allowed the cross-POS translation (a noun can be translated only by a noun, etc.); the user can decide if he/she allows the application to cross the boundaries between the parts of speech, through a parameter modifiable in the configuration file. Each choice has its rationales, as we know that a word is not always expressed through the same part of speech when translated in another language. On the other hand, putting all the words in the same bag increases the number of computations and the risk of error.

If the user's machine has multiple processors, the application can call a function that splits the time consuming problem of computing the vector similarities and runs it in parallel. This function is activated by the user through a "multithreading" parameter in the configuration file. To avoid overloading the memory, the application gives the user the opportunity to decide how many of the source/target vectors are loaded in the memory at a specific moment, through the "loading" parameter, activated only for "multithreading: yes"; setting this parameter to a value smaller than the matrix size can cause an important time delay, so it's in user's hands to set properly the parameters and balance advances and disadvantages according to the time constraints and according to the available memory resources.

For the proper nouns, which are more probably to be translated into a similar graphic form from a language to another, we introduced a cognate score, which is used in the computing of the similarity metric to boost the cognate candidates. This is specified in the configuration file by the parameter LD (Levenhstein Distance, the metric we based the cognate score on). This score is taken into account only if decreases under a certain threshold, which we empirically set at 0.3.

In the following, we will reproduce the configuration file we already mentioned and where the default values set for the parameters can be seen:

```
*multithreading:yes/no (default=no)
*loading:int(default=0) if the parameter's value is higher than the number of vectors in the matrix, its use becomes obsolete.
*frequency:int(default=3)
*window:int(default=5)
//5.asking for the loglikelyhood of a co-occurrence to be higher than a certain threshold, the user can reduce the space and time costs
*ll:int(default=3)
*sourceamverblist:string (default=is are be will shall may can etc.)
*targetamverblist:string (default=este sunt suntem sunteţi fi poate pot putem puteţi etc.)
*crossPOS:yes/no(default:no)
// 9.the user has to provide a list of all the open class POS labels (i.e. labels for common nouns, proper nouns, adjective, adverbs and main verbs) of the source language
*sPOSlist:string(default=nc np a r vm)
// 10.the user has to provide a list of all the open class POS labels (i.e. labels for common nouns, proper nouns, adjectives, adverbs and main verbs) of the target language
*tPOSlist:string(default=nc np a r vm)
//11.the user can decide if a cognet score (Levenshtein Distance) will be taken into account in computing the vector similarities for proper nouns
*LD:yes/no(default=no)

multithreading:yes
loading:5000
frequency:10
window:5
ll:3
sourceamverblist:am is are was were been beeing had has have be will would shall should may might must can could need
targetamverblist: este sunt eşti suntem sunteţi vei va voi vor vom veţi era eram erai eraţi fi fost pot poţi poate putem puteţi putea puteai puteam puteaţi puteau ar aţi am aş ai are avem au aveţi aveam avea aveaţi aveai aveau
crossPOS:no
sPOSlist:nc np a r vm
tPOSlist:nc np a r vm
LD:no
```

The tool is implemented in the programming language C#, under the .NET Framework 2.0. It requires the following settings to run: NET Framework 2.0., 2+ GB RAM (4 GB preferred). The application can be run as an executable file both under Windows and Linux platforms. The tool is language independent, providing that the corpus is POS-tagged according to the MULTEXT-East tag set (see http://nl.ijs.si/ME/V3/msd/html/msd.html) and that the user is introducing manually in the configuration file the list of source and target verbs concerning the parameters `sourceamverblist` and `targetamverblist`.

4.1 Experiments and results

4.1.1. Experimental setup

The base lexicon used by this tool is a word-to-word sub-part of a translation table, extracted with GIZA++ from corpora in different registers. Only the content words were kept. The translation table can be loaded as two different dictionaries EN-RO (64,613 polysemous entries) and RO-EN (66,378 polysemous entries).

Tests have been conducted on different sizes and different types/registers of comparable corpora:

1. A comparable corpora of small size representing the civil code of Romania in force until October 2011 (184,081 words) vs. the civil code of Quebec – in English (199,401 words). The corpora were manually downloaded from specific websites and we took into account the necessity to find a version of the document with diacritics for the Romanian part. The structure of the corpora is quite rigid and the noise (comprising dates or the numbers of the articles and paragraphs) was easily removed. Although we will not present detailed results here, we mention that they are not satisfactory. We assume this is due to the small size of the corpus.

2. A corpus of articles extracted at RACAI from Wikipedia: 743,194 words for Romanian, 809,137 words for English. This corpus is a strongly comparable one, with little noise (due to the fairly similar structure of the wiki pages, which facilitated the elimination of the boilerplates).

3. The corpora compiled by USFD in this project is a journalistic corpora downloaded from Google News through a heuristic based on a list of English paper titles, translated into Romanian. After the elimination of the words without content from the titles, they were used as queries into Google News and the results were downloaded for both languages. Before being released, the corpora were been cleaned for boiler plates. (For more details, see *D3.4 Report on methods for collection of comparable corpora* on the internet page of the project: http://www.accurat-project.eu/index.php?p=deliverables)

All corpora were tokenized using a library implemented in our research centre. We than checked for the presence of the diacritics and we noticed that the USFD corpora had Romanian documents which lacked those features. We used DIAC+, a tool developed at RACAI (Tufiș and Ceauşu, 2008) which automatically inserts diacritics in Romanian texts, with an error margin of 0,27% in the character accuracy.
Consequently, we checked the USFD text for repeating sentences/paragraphs and eliminated them. This reduced a lot the dimension of the USFD corpus, especially of the Romanian part.

All corpora were than lemmatized and POS-tagged using the TTL toolkit (Ion, 2007). The POS-tagging is a necessary process for selecting the content words. The output of TTL is in XML format and the annotation is compliant to the MULTEXT-East morpho-lexical specification (MSD tags, which are complex), therefore we recovered the information and put it in a simpler format (ex: *man^Nc*), keeping only the data we needed in our approach.

4.1.2. Some results

The evaluations are in progress, therefore only a small part will be presented here. We manually compiled a gold standard lexicon of around 1,500 words (common nouns, proper nouns, verbs and adjectives) from the Wikipedia corpus. In the conditions described by the default parameters in the configuration file, the precision-1 and precision-10 scores introduced earlier were computed:

POS	Precision-1	Precision-2
common nouns	0.5739	0.7381
proper nouns	0.6956	0.7336
adjectives	0.4943	0.6292
verbs	0.6620	0.8275

Table 3: P-1 and P-10 for the 1,500 test words from Wikipedia corpus

additional^af	**significant^af**
suplimentari^af 0.1268#	**importante^af** 0.0468#
general^af 0.0014#	**semnificativă^af** 0.0427#
financiare^af 0.0011#	mari^af 0.0418#
referitor^af 0.0010#	**principalele^af** 0.03902#
nouă^af 0.0008#	prezente^af 0.0367#
mari^af 0.0008#	**importantă^af** 0.0367#
indian^af 0.0007#	economice^af 0.0346#
comună^af 0.0007#	culturale^af 0.03423#
medie^af 0.0006#	**semnificative^af** 0.0339
nordică^af 0.0006#	singurele^af 0.0315#
francez^af 0.0006#	**semnificativ^af** 0.0309#
religious^af	**modern^af**
religioase^af 0.06583#	considerată^af 0.0457#
culturale^af 0.0448#	veche^af 0.0423#
politice^af 0.0412#	cunoscut^af 0.0403#
religioasă^af 0.0400#	antică^af 0.0390#
umane^af 0.0370#	roman^af 0.03790#

economice^af 0.0369#	engleză^af 0.0377#
diferite^af 0.0369#	vechi^af 0.0372#
administrativ^af 0.03474#	**modern**^af 0.0319#
sociale^af 0.0335#	latină^af 0.0314#
economic^af 0.0330#	importante^af 0.0310#
diverse^af 0.0318#	marea^af 0.0307#

Table 4: Sample of the result file for the adjective translations; the correct translations are bolded.

The experiments with the USFD corpus were very disappointing in the beginning. We realised the need for correcting some POS annotations and also to change the strategy for the LL filtration, because of the big difference in size between the two corpora (7,280,609 English words and 2,170,425 Romanian words). We decided to keep in the co-occurrence vectors only the first *n* words in descending order of their log likelihood scores. The threshold n was set experimentally to 50.

We also used the Levenshtein Distance for all the POS analysed to boost the scores for the translations graphically more similar with the word to be translated. This boost is done after all the similarity scores between a certain source word and all the target words are computed. The threshold to which the words were considered cognates were a LD<0.3 and the boost meant a multiplication with 10 of the similarity score. All the scores that resulted above 1 were reduced to 0.99.

We also felt the need for introducing two different frequency thresholds for the two corpora, to compensate the difference in size. The values of the frequencies established after more experiments were 100 for the source words (English) and 20 for the target words (Romanian).

After all these heuristics, the results become more reasonable, but still not rising to the performances on the Wikipedia corpus. We explain that but the serious difference in the degree of comparability between the corpora.

Because of the time constraints (the final and cleaner version of the USFD corpora was made available shortly before the deadline for this paper) we focused only on three POS: common nouns, adjectives and verbs. We constructed for each POS a gold-standard dictionary with 100 entries and Precision-1 and Precision-10 scores were computed:

POS	Precision-1	Precision-10
common nouns	0.2909	0.5454
adjectives	0.3663	0.5049
verbs	0.24	0.48

Table 5: P-1 and P-10 for the 300 test words from USFD corpora

The effect of introducing the cognate test for all the POS was important for many of the good results, producing more forms of the same lemma as possible translations, which is consistent with the reach morphology of Romanian and is very useful in a dictionary:

ministers^nc|*ministru^nc ministrul^nc miniştrilor^nc*
fund^nc|*fondului^nc fondul^nc fond^nc*
sector^nc|*sector^nc sectorul^nc sectorului^nc*

republican^af|*republican^af republicani^af republicană^af*
national^af|*naţional^af naţională^af naţionale^a*
german^af|*german^af germană^af germane^af germani^af*

considered^vm|*considerat^vm consideră^vm considera^vm consider^vm considerând^vm*
continue^vm|*continua^vm continuă^vm continue^vm continuat^vm*
confirm^vm|*confirmat^vm confirmă^vm confirma^vm confirmată^v*

This phenomenon occurred for around 46% of the correct translated nouns, 39% of the correct translate adjectives and 29% of the correct translated verbs.

For some translations in which the cognate test didn't interfered, multiple solutions could be seen also:

policies^nc|*plan^nc program^nc planul^nc măsurilor^nc măsuri^nc*
debts^nc|*datoriile^nc datoriilor^nc*

former^af|*fostul^af fostului^a*
black^af|*negru^af neagră^af*
last^af|*trecut^af fostul^af recent^af*

played^vm|*juca^vm jucat^vm*
earned^vm|*câştigat^vm obţinut^vm*
die^vm|*murit^vm mor^vm muri^vm moară^vm*

5. Conclusions

We created a tool destined to extract bilingual word-to-word lexicons from comparable corpora. Based on a well-known approach (Rapp, 1999) we intended to extend it to deal with polysemy, so that we can use automatically extracted translation tables as seed dictionaries. We also proposed a filtration of the co-occurrence vectors according to the log likelihood score, starting from the idea that this score is a good measure for the significance of two words occurring together. The tool can be also used in multithreading mode if the user's machine has multiple processors.

From the three types of corpora we experimented with, only one (No.2) showed good and really usable results. This is coming from the strong comparability of the corpora (Wikipedia articles are quite similar, with some in one language being poorer in content than in the other language). We will keep working on the corpus No. 1, by adjusting the parameters in the configuration file and on the corpus No.3 by experimenting with the LL score

filtration. We also need to evaluate how many new words (which are not part of the seed dictionary) are translated through our method.

6. Acknowledgements

Research reported in this paper has been supported by the ACCURAT project within the EU 7th Framework Pogramme (FP7/2007-2013), grant agreement no. 248347.

References

Brown, P.; Cocke, J.; Della Pietra, S. A.; Della Pietra, V. J.; Jelinek, F.; Lafferty, J. D.; Mercer, R. L.; Rossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2), 79-85.

Chiao, Y.-C., Zweigenbaum, P. (2003) Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of Coling 2002*, Taipei, Taiwan, 26-30 August 2002

Church, K. W., Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, 76-83.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Fung, P., McKeown, K. (1997) Finding terminology translations from non-parallel corpora. *Proceedings of the Fifth workshop on Very Large Corpora*. ed. Joe Zhou and Kenneth Church, 18 August 1997, Tsinghua University, Beijing, China, 20 August 1997, Hong Kong University of Science and Technology, Hong Kong; pp.192-202.

Gamallo P., Pichel, J.R. (2005). An Approach to Acquire Word Translations from NonParallel Texts, *Lecture Notes in Computer Science*, vol. 3808. SpringerVerlag.

Gamallo, P. (2007). Learning bilingual lexicons from comparable english and spanish corpora. In *Machine Translation SUMMIT XI*, Copenhagen, Denmark.

Gamallo P. (2008) Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. *Proceedings of LREC 2008 Workshop on Comparable Corpora*, Marrakech, Marroco, pp. 19-26. ISBN: 2-9517408-4-0.

Gale, W. A., Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(3), 75-102.

Kay, M., Roscheisen, M. (1993). Text-Translation Alignment. *Computational Linguistics*, 19(1), 121-142.

Laroche A., Langlais P.: Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of COLING 2010*: 617-625.

Morin E., Daille B., Takeuchi K., Kageura K. (2007). Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (ACL'07), pp. 664-671.

Och, F.J. and Ney, H. (2003) A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003

Och, F. J., Tillmann, C., Ney, H. (1999). Improved alignment models for statistical machine translation. *Joint SIGDAT conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Proceedings ed. Pascale Fung and Joe Zhou*, 21-22 June 1999, University of Maryland, College Park, MD, USA; pp.20-28.

Rapp, R. (1995). Identifying word translations in nonparallel texts. In: *Proceedings of the 33rd Meeting of the Association for Computational Linguistics. Cambridge*, Massachusetts, 320-322.

Rapp, R. (1996). Die Berechnung von Assoziationen. Hildesheim: Olms.

Rapp, R. (1999) Automatic identification of word translations from unrelated English and German corpora. *ACL-1999: 37th Annual Meeting of the Association for Computational Linguistics. Proceedings of the conference*, 20-26 June 1999, University of Maryland, College Park, Maryland, USA; pp.519-526.

Tufiş, D, Ceauşu A. (2008). DIAC+: A Professional Diacritics Recovering System, in *Proceedings of LREC 2008 (Language Resources and Evaluation Conference)*, May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408

Romanian Translational Corpora: Building Comparable Corpora for Translation Studies

Iustina Ilisei[⋆], Diana Inkpen[§], Gloria Corpas[‡], Ruslan Mitkov[⋆]

[⋆]Research Institute in Information and Language Processing, University of Wolverhampton,
Wulfruna Street, Wolverhampton, United Kingdom,
iustina.ilisei, r.mitkov@wlv.ac.uk

[§] School of Information Technology and Engineering, University of Ottawa,
800, King Edward Street, Ottawa, ON, K1N 6N5, Canada
diana@site.uOttawa.ca

[‡]Department of Translation and Interpreting, University of Málaga, Málaga, Spain
gcorpas@uma.es

Abstract

Building comparable corpora for the investigation of translational hypotheses is an important task within the translation studies domain. This paper describes the compilation of a translational comparable corpus for the Romanian language. The resource comprises translated and non-translated news articles and it is designed to be used in the investigation of translational language and translational hypotheses.

1. Introduction

Translational hypotheses proposed in the last two decades require certain resources. Most of these hypotheses (e.g., translation universals, laws or norms) imply the comparison between translated texts produced by professional translators to non-translated texts. As a consequence, there is a need of monolingual comparable corpora specifically designed for the study of translational language. These corpora need to contain two subcorpora: a subcorpus that comprises translated texts, and a comparable one which comprises non-translated, original texts.

This paper is structured as follows: first, several reasons are given as to why it is important to compile comparable corpora for translation studies, and then the definitions required for this study are described. In section 2., some other, similar resources built for other languages are highlighted, and furthermore the paper continues with the main section of the compilation of the current corpus. This main section, 3., comprises various details regarding the data collection, data preparation, and the statistics reported for the corpus. It also provides a short example of investigations which can be undertaken relying on this linguistic resource. Finally, the paper concludes with the highlights of the corpus.

1.1. Motivation

Compiling comparable corpora for the investigation of various hypotheses proposed within the area of translation studies is currently one of the main, time-consuming tasks within the domain. These hypotheses attempt to grasp and analyse certain features of the translational language and the lack of resources proves to be a serious obstacle for further refinement of the scholars' ideas and findings, and consequently for the advancement of translation theory.

The translationese effect, one of the assumptions of the discipline which considers translated language have certain specific, peculiar traits at various linguistic levels (Borin and Prütz, 2001; Hansen, 2003; Baroni and Bernardini, 2006; Puurtinen, 2003), has been a subject of debate for the last fifteen years, bringing together different perspectives on translational language. Translation universals are hypotheses that have also raised various questions among scholars; their validity is a continuous subject of debate (Corpas et al., 2008; Becher, 2011). More rigorous evidence of these claims would lead to a refinement of the theory, would raise awareness among translators about possible effects over translated texts (Laviosa, 2002, p. 77) and would facilitate further methodologies to more accurate translations with more "desired effects and fewer unwanted ones" (Chesterman, 2000). However, the lack of appropriate resources is a significant impediment to this end.

The exploitation of monolingual comparable corpora has been widely sustained among scholars, and the call for more developments of specific tools and resources for professional translators has had an impact on the domain. Even though a few translational corpora have been built (one well-known example is the English Translational Corpus), most languages still lack a proper resource for the investigation of the translational hypotheses. To the best of our knowledge, the Romanian language would be one of these languages. This work bridges this gap and reports on the compilation of the RoTC corpus, a monolingual comparable corpus that comprises newspaper articles.

Nevertheless, the exploitation of this type of resource is not restricted to translation researchers. It can also be used in other fields: for instance, for the improvement of statistical machine translation (SMT) systems. Scholars, such as (Kurokawa et al., 2009; Lembersky et al., 2011), found that making use of translation studies' main hypotheses and findings and training their SMT framework on translational corpora can result in an overall improvement of their system.

1.2. Translational Comparable Corpora

First, an attempt at defining comparable corpora is required. The key attributes of what constitutes comparable corpora are described as follows (McEnery, 2003): two corpora, A and B, are considered to be comparable if both A and B are found to have:

- the same *sampling frame* with *similar balance* and *representativeness*
- the same *proportions* of the same *genres* in the same *domains*
- the same *sampling period*

These requirements are imposed on the current resource and further details follow in section 3.

However, a definition of comparable corpora is not yet agreed by the scholars in the field. There is only a standard provided by EAGLES (1996) in which it is emphasised that a comparable corpus is *a corpus which comprises similar texts in more than one language or variety*. This standard describes the circumstances when a comparable corpus is needed: in a comparative analysis between two or more languages, or between two or more varieties of texts.

To prevent possible misinterpretations introduced by this definition (i.e., no translational corpus can be considered comparable since the resource only has texts in one language), Baker (1995) suggests that the concept of translational corpus to be seen as a new type of comparable corpus. The resource proposed includes two subcorpora in one and the same language: one subcorpus with originally produced texts in a given language, the other one with texts translated into the same language from one or more source languages. Baker (1995) proposes that both subcorpora should be similar in terms of domain, variety of language, time span, and to be of comparable length.

Considering these definitions, it seems to be a matter of how *similar* can be understood or modelled depending on the research question. The degree of comparability is "in the eye of the beholder", strictly depending on the requirements and the objectives of the research study (Maia, 2003). Although several scholars discuss this topic, the vagueness of the concept still continues, mainly because of its fuzzy notions from the definition.

Second, the concept of translational corpus is tackled. A translational corpus contains translated texts written by human translators, and it is usually exploited within the area of translation studies. Therefore, for the investigation of hypotheses which compare assumed features of translated texts to non-translated texts, a translational comparable corpus can be considered an appropriate resource for the given research question. If the translational hypothesis does not imply a comparison between translated and non-translated texts, then a translational corpus, comprising only translated texts, may suffice.

2. Related Work

As translated text is the focal point of the translation studies domain, compiling translational corpora (both comparable and parallel) is the vital resource for various investigations. As a result, several corpus-based approaches exploit monolingual comparable corpora, where comparability is between translated and non-translated texts in the same language. Despite the difficulties which arise in the compilation process, there are linguistic resources available for the following main languages: English (Baker, 1995), Portuguese (Frankenberg-Garcia, 2004), Spanish (Corpas, 2008), Dutch and German (De Sutter and Van de Velde, 2008), Chinese (Xiao et al., 2008).

The Translational English Corpus, TEC, is probably one of the first compiled corpora for translation studies in the mid-nineties (Baker, 1995). The ten-million-word corpus comprises four categories of texts: biography, fiction, newspaper texts and in-flight magazines, with translations into English from both European and non-European languages. The main experiments were employed manually and they show that corpus-based research allowed translation universals to be more clearly defined, to progress to large-scale, target-oriented research, and to consider a wider range of socio-cultural factors (Laviosa, 2002).

For Spanish, the statistical significance of various features proposed to stand for the simplification hypothesis[1] were tested using monolingual comparable corpora on medical and technical domains (Corpas, 2008; Corpas et al., 2008).

3. RoTC : Corpus Compilation

Regarding the comparability of corpora, all the definitions have in common the following parameter: *similarity between texts*. Furthermore, the definition narrows down the concept of similarity and is described in terms of genre, domain, sampling and time-frame, all of which are tackled in the compilation of process of the RoTC corpus.

Beyond the tricky notion of comparable corpora, there are also practical issues when compiling a corpus. Some of them are classical and some of them are specific to translational corpora. Fundamental aspects to consider are the *validity* and *reliability* of the research experiments based on the specific corpus, tailored to meet the intended purpose. *Representativeness* is a challenging aspect for this type of linguistic resource, as it is difficult to assure that the data is representative of a particular language or genre. When considering which texts should be included in the corpus, the decision process can go beyond the text type or genre, text function or scope and how typical or influential the given text can be. Also, regional and temporal factors have to be taken into consideration, being part of the criteria of a corpus. Nationality, age, native language, ethnicity, etc. can all be decisive factors according to the research purpose, and more often than not this type of information cannot be accessed.

Sample size is another relevant consideration and may be the most important feature in achieving representativeness: how many texts should be included in the corpus and what the size of each of them should be. Representativeness depends on whether the sample includes the full range of language variability intended, so the researchers who use the corpus will be able to generalise their findings. In contrast, Kennedy (1998) argues that a bigger corpus is not necessarily more useful than a smaller one, as the data amount under

[1] Simplification hypothesis suggests that translated texts appear to be simpler than the non-translated ones (Baker, 1993).

investigation is always limited (Kennedy, 1998, p. 66-70). Nevertheless, a smaller corpus can be sufficient in some cases, for example, if the research lines have the grammar in focus (Hunston, 2002, p. 26) and, ultimately, the data availability factor of suitable texts should not be dismissed.

3.1. Corpus Design

Some scholars from the domain suggest that the best resource for the investigation of translationese is a monolingual translational comparable corpus (i.e., containing translated and non-translated texts in the same language) (Olohan, 2004), because in this manner the approach would avoid any foreign interference (Pym, 2008) and, consequently, it would fit well in the investigation of the nature of translated text.

The main objective of this resource, the Romanian Translational Corpus, is to allow the investigation of translationese and the related translational hypotheses, such as translation universals. As no study of the Romanian language has been done for translationese, to the best of our knowledge, a dedicated type of resource did not exist. For this reason a comparable corpus has been specially compiled for this task, consisting of newspaper articles published between 2005-2009.

The RoTC corpus comprises two subcorpora: a translated subcorpus and a non-translated subcorpus. The translated one is collected from the South-East European Times[2], a multilingual news portal translated into nine languages of the Balkans, one of them being Romanian. The translated subcorpus comprises 223 articles written between 2005-2009 to keep the same time frame as the non-translated subcorpus. The non-translated subcorpus comprises 416 documents in the same domain, from a well-known newspaper in Romania called 'Ziua'[3].

3.1.1. Data Preparation

The content of the South-East European website is realised as public domain, meaning it can be used and distributed without permission. The process of selecting the articles for the RoTC corpus is described in the following paragraphs.

All the articles were downloaded using various scripts which use the URL structure information. The link allows the selection of the articles to fit various needs, that in the given context are:

- to select articles after the language (i.e., the URL contains the string "www.setimes.com/ .../ro/... "for the Romanian language),
- to select articles after the date (i.e., the date can be easily extracted from the link as it appears in this format "www.setimes.com/ .../yyyy/mm/dd/... ").

The topic of the articles selected was the international news in order to be able to cover the same subjects over the same time-span, and hence obtain a comparable corpus between the texts selected from the South-East European Times website and the Ziua newspapers. Also, the number of texts between non-translated and translated texts have been balanced by randomly selecting 416 non-translations written between 2005-2007 versus the 224 translations written between 2005-2010. A ratio of 2:1 is kept.

The RoTC corpus has in total 341320 tokens (200211 for the translated subcorpus and 141109 tokens for the non-translated subcorpus). The selected articles are written by various translators, so the possibility of a specific style playing a role in the classification task is avoided. The main shortcoming of the translated subcorpus is that the portal, due to confidentiality issues, fails to provide precise information about the source language or the identity of the original author, nor the translator. Nevertheless, some of the articles do mention the source of their news information (e.g., Reuters) and it can thus be assumed the original source language of the given text. In addition, it is often stated that various information sources were used when the given article was produced.

The argument that the articles are translations and not original texts is inferred from two distinct sources: first, this portal was entirely harvested and used in a machine translation task, reporting the resource as having translations into languages of the Balkans, including the Romanian language (Tyers and Alperen, 2010). Second, it is inferred from the following rationale: one text can not be originally produced in ten languages and yet be perfectly aligned from one language to another (i.e., one Romanian article to have its source language Romanian, the corresponding, parallel Turkish article to have its source language Turkish, and at the same time, both the Romanian article and the Turkish one to be perfectly aligned to each other). The fact that all are aligned to each other leads to the assumption that, at least nine out of ten parallel articles are in fact translations. Consequently, it results in a high probability to have mostly translations, if not only translations, in the RoTC translated subcorpus. However, the attempt to clarify this issue from its source failed due to the portal's confidentiality policy.

The non-translated subcorpus does not present the same difficulty in assessing whether the texts are originally produced articles, since the newspaper is a national one having its texts written only in the Romanian language. Additionally, the articles do state their authors, and their full names indicate that they are Romanian natives. Thus, the subcorpus comprises non-translated texts, written by various authors.

3.1.2. Part of Speech Tagger

All the texts were tagged using the part of speech tagger provided as a web service by the Research Institute for Artificial Intelligence[4], the Romanian Academy (Tufiş et al., 2008b; Tufiş et al., 2008a), and its output transformed into XML[5] format to ease the access to the data representation of the document. A sample of the XML format is represented in figure 1. In the following section, a few statistics about the size of the RoTC corpus and its components are reported.

[2] http://www.setimes.com
[3] http://www.ziuaveche.ro
[4] http://www.racai.ro/webservices/
[5] Extensible Markup Language

```
<sentence id="w128">
<token id="w129"><text>Acestea</text>
<lemma>acesta</lemma><tags>
<morpho>Pd3fpr</morpho></tags>
</token>

<token id="w130"><text>au</text>
<lemma>avea</lemma><tags>
<morpho>Va--3p</morpho></tags>
</token>

<token id="w131"><text>fost</text>
<lemma>fi</lemma><tags>
<morpho>Vmp--sm</morpho></tags></token>

<token id="w132"><text>primele</text>
<lemma>prim</lemma><tags>
<morpho>Mofprly</morpho></tags></token>

<token id="w133"><text>alegeri</text>
<lemma>alegere</lemma><tags>
<morpho>Ncfp-n</morpho></tags></token>
... ... ...
</sentence>
```

Figure 1: Sample of the output provided from the POS tagger converted into XML format.

3.2. RoTC Corpus Statistics

Some fundamental statistics are computed for the RoTC corpus. In table 3.2. the size of the corpus is presented as the number of tokens for each subcorpus, and as a whole. It is noted that the RoTC corpus has a slight majority of non-translated texts, comprising 58.6578 % of the total number of articles. This happens as the amount of texts available for the same topic in the comparable translated corpus is slightly lower compared to the number of non-translated articles, and the intention is to obtain as many articles as possible to be able to use the resource in a machine learning framework. Obviously, the comparability aspects are considered, so it is settled to keep a ratio of 2:1 between the translated and non-translated texts to comply with the same sampling frame with similar balance factor.

RoTC Corpus			
Subcorpus	Tokens No.	Texts No.	Percentage
Non-Translated	200211	223	58.6578 %
Translated	141109	416	41.3421 %
Total	341320	639	100%

Table 1: RoTC Corpus Statistics.

To tackle the same proportions of the same genres in the same domain requirement, table 3.2. presents the average value for the number of tokens per text. The figures show that the RoTC corpus has an average number of tokens of 481 for the translated subcorpus, and 632 for the non-translated texts. These values are closely related (as expected since in this corpus there are only newspapers articles), and it remains to be investigated further whether the slight difference is due to some feature assumed to be specific to either translational language or to non-translational one (some hypotheses make references related to the size of translated texts in general). Nevertheless, the RoTC corpus also complies with the same proportion requirement for a comparable corpus.

RoTC Corpus	
Subcorpus	Average
Non-translated	632.7757848
Translated	481.2764423

Table 2: Average tokens per document.

Furthermore, a few details about the applicability of this linguistic resource in the investigation of translational hypotheses (Ilisei and Inkpen, 2011; Ilisei et al., 2011). In (Ilisei et al., 2011) the hypothesis targeted was the explicitation hypothesis, and brief details regarding their findings are summarised in the following subsection.

3.3. RoTC Corpus Applied in the Investigation of the Explicitation Hypothesis

The Explicitation hypothesis, also assumed to be a universal of translational language (Baker, 1996), states that additional background information which is found implicitly within the message of the source text appears explicitly spelled out in the equivalent translated text. Considering the opposite phenomenon resulting from this hypothesis, ellipsis would occur much more often within the non-translated texts than translational language. Therefore, investigating ellipsis within translated or non-translated texts can lead to findings regarding the explicitation hypothesis. A machine learning system was built for this analysis (Ilisei et al., 2011) and the following section provides brief details of these experiments and their results.

Ellipsis constitutes one of the attributes proposed for the investigation of the explicitation hypothesis. The correct understanding of ellipsis is absolutely essential in the translation process, and hence any type of linguistic resource labelled with this information would be highly appreciated within the domain. As the ellipsis of subjects is the most frequent type, the study focuses only on the anaphoric zero pronoun (hereafter noted as AZP). A tool which uses machine learning techniques is used to identify the verbs which have a zero pronoun in the subject position (Mihăilă et al., 2010; Mihăilă et al., 2011). The software used is known to have an accuracy of 74%.

Before presenting the results of the AZP impact on translational language, the notion of anaphoric zero pronoun is defined. As an agreement between scholars has not yet emerged, anaphora is still a controversial topic and there are thus different classifications of ellipsis (Mladin, 2005). The adopted definition is the following: an anaphoric zero pronoun appears when an anaphoric pronoun is omitted but nevertheless understood (Mitkov, 2002), in which case the zero pronoun corefers to one or more overt nouns or noun phrases in the text (entities which provide the information for the correct understanding of the ellipsis).

Their findings on the RoTC corpus show that a machine learning system is able to distinguish between translated and non-translated texts relying only on the anaphoric zero pronoun attribute. The accuracy obtained is between 71% and 75% (Ilisei et al., 2011). Therefore, once more it can be emphasised that the monolingual comparable corpus compiled for the Romanian language appears to be a reliable linguistic resource in the investigation of translational hypotheses, and most likely for other domains, such as translation technology. This linguistic resource will be made available online[6] once its documentation is complete.

4. Conclusion

Building comparable corpora for the investigation of translational hypotheses is an important task within the translation studies domain. This paper describes the compilation of a translational comparable corpus for the Romanian language. The resource comprises translated and non-translated news articles and is designed to be used in the investigation of translational language and translational hypotheses. Moreover, a few details about the applicability of this linguistic resource are mentioned: explicitation hypothesis is investigated by analysing the impact of the anaphoric zero pronouns in translational language compared to non-translational one.

5. References

M. Baker, 1993. *Text and Technology: In Honour of John Sinclair*, chapter Corpus Linguistics and Translation Studies Implications and Applications, pages 233–250. Amsterdam & Philadelphia: John Benjamins.

M. Baker. 1995. Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target*, 7(2):223–43.

M. Baker, 1996. *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, chapter Corpus-based Translation Studies:The Challenges that Lie Ahead, pages 175–186. Amsterdam & Philadelphia: John Benjamins.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21, 3:259–274.

V. Becher. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Ph.D. thesis, University of Hamburg.

L. Borin and K. Prütz. 2001. Through a glass darkly: Part-of-speech distribution in original and translated text. *Language and Computers*, 37(1):30–44.

A. Chesterman, 2000. *Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects*, chapter A Causal Model for Translation Studies, pages 15–27. St. Jerome.

G. Corpas, R. Mitkov, N. Afzal, and V. Pekar. 2008. Translation Universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the AMTA*, Waikiki, Hawaii.

[6] `http://clg.wlv.ac.uk`

G. Corpas. 2008. *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main, Berlin & New York: Peter Lang.

G. De Sutter and M. Van de Velde. 2008. Do the mechanisms that govern syntactic choices differ between original and translated language? A corpus-based translation study of PP extraposition in Dutch and German. In R. Xiao, L. He, and M. Yue, editors, *Proceedings of the international symposium on using corpora in contrastive and translation studies (UCCTS 2008)*.

EAGLES. 1996. Expert Advisory Group on Language Engineering Standards Guidelines.

A. Frankenberg-Garcia. 2004. Are translations longer than source texts? A corpus-based study of explicitation. In *Third International Corpus Use and Learning to Translate Conference, Barcelona, Spain*, January.

S. Hansen. 2003. *The Nature of Translated Text. An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. Ph.D. thesis, Saarbrücken: Saarland University.

S. Hunston. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

I. Ilisei and D. Inkpen. 2011. Translationese Traits in Romanian Newspapers: A Machine Learning Approach. *International Journal of Computational Linguistics and Applications*.

I. Ilisei, C. Mihaila, D. Inkpen, and R. Mitkov. 2011. The Impact of Zero Pronominal Anaphora on Translational Language: A Study on Romanian Newspapers. In *Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT2011, Cluj-Napoca, Romania*, July 46.

G. Kennedy. 1998. *An Introduction to Corpus Linguistics*. Amsterdam: Rodopi.

D. Kurokawa, C. Goutte, and P. Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of the MT-Summit*.

S. Laviosa. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam & New York: Rodopi.

G. Lembersky, N. Ordan, and S. Wintner. 2011. Language Models for Machine Translation: Original vs. Translated Texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

B. Maia. 2003. What are Comparable Corpora? In *Proceedings of the Workshop on Multilingual Corpora: Linguistic Requirements and Technical perspectives, Corpus Linguistics*, pages 27–34, Lancaster, U.K.

A. McEnery, 2003. *Oxford Handbook of Computational Linguistics*, chapter Corpus Linguistics, pages 448–463. Oxford: Oxford University Press.

C. Mihăilă, I. Ilisei, and D. Inkpen. 2010. To Be or Not to Be a Zero Pronoun: A Machine Learning Approach for Romanian. In *Proceedings of the Processing Romanian in Multilingual, Interoperational and Scalable Environments Workshop (PROMISE)*.

C. Mihăilă, I. Ilisei, and D. Inkpen. 2011. Zero Pronom-

inal Anaphora Resolution for the Romanian Language. *Research Journal on Computer Science and Computer Engineering with Applications "POLIBITS"*, 42.

R. Mitkov. 2002. *Anaphora Resolution*. Longman, London.

C. I. Mladin. 2005. Procese şi Structuri Sintactice "Marginalizate" în Sintaxa Românească Actuală. Consideraţii Terminologice Din Perspectivă Diacronică Asupra Contragerii - Construcţiilor - Elipsei. *The Annals of Ovidius University Constanţa - Philology*, 16:219–234.

M. Olohan. 2004. *Introducing Corpora in Translation Studies*. Routledge.

T. Puurtinen, 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, chapter "Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals?", pages 141–154. Amsterdam & New York: Rodopi.

A. Pym, 2008. *Beyond Descriptive Translation Studies*, chapter On Toury's laws of how translators translate, pages 311–328. Benjamins.

D. Tufiş, D. Ştefănescu, R. Ion, and A. Ceauşu, 2008a. *Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007), Lecture Notes in Computer Science*, volume 5152, chapter RACAI's Question Answering System at QA@CLEF 2007, pages 3284–3291. Springer-Verlag, September.

D. Tufiş, R. Ion, A. Ceauşu, and D. Ştefănescu. 2008b. RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco*, number ISBN 2-9517408-4-0. ELRA - European Language Ressources Association, May.

F. M. Tyers and M. Alperen. 2010. South-East European Times: A parallel corpus of the Balkan languages. In *Proceedings of the 7th Language Resources and Evaluation Conference - LREC 2010, Valletta, Malta*.

R. Xiao, L. He, and M. Yue. 2008. In Pursuit of the 'Third Code': Using the ZJU Corpus of Translational Chinese in Translation Studies. In Lianzhen He Richard Xiao and Ming Yue, editors, *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008)*.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova

University of Oslo, Department of Informatics
angelii@ifi.uio.no

Abstract

Machine-readable dictionaries play important role in the research area of computational linguistics. They gained popularity in such fields as machine translation and cross-language information extraction. Wiki-dictionaries differ dramatically from the traditional dictionaries: the recall of the basic terminology on the Mueller's dictionary was 7.42%. Machine translation experiments with the Wiki-dictionary incorporated into the training set resulted in the rather small, but statistically significant drop of the the quality of the translation compared to the experiment without the Wiki-dictionary. We supposed that the main reason was domain difference between the dictionary and the corpus and got some evidence that on the test set collected from Wikipedia articles the model with incorporated dictionary performed better.

Keywords: machine-readable bilingual dictionary, Wiki-dictionary, statistical machine translation

1. Introduction

Machine-readable bilingual dictionaries are employed in fields such as machine translation and cross-language information extraction. Possibilities for automatic generation of high quality resources of this type are being actively investigated by the research community because manual development is expensive and time-consuming. The main challenges for this task are found in achieving a reasonable level of accuracy, excluding noisy data and providing required coverage of terminology. With efficient methods for creation of bilingual dictionaries for different domains, we can, for example, experiment with usage of these dictionaries in the alignment modules of translation systems.

In this article we investigate the quality and the content of an English-Russian dictionary (Wiki-dictionary)[1] created from Wikipedia. In order to perform an in-depth evaluation of the resulting dictionary, we did named entity recognition and classification, computed the recall of the translation pairs on the traditional English-Russian Mueller's dictionary, collected corpus statistics from ÚFAL Multilingual Corpora[2] and incorporated the dictionary into a statistical machine translation system.

Even though it has been repeatedly shown that Wiki-dictionaries have many advantages, our experiments with the Wiki-dictionary show that it is important to clearly understand the domain to which they are applicable, otherwise improper usage may lead to drop of accuracy in the translation task.

2. Related Work

In the last decade the online encyclopedia Wikipedia has gained popularity because it is a multilingual, dynamic and rapidly growing resource with user-generated content. Wikipedia link structure was exploited, for example, for linking ontology concepts to their realizations in text (Reiter et al., 2008), for generating comparable corpora using a link-based bilingual lexicon for identification of similar sentences (Adafre and de Rijke, 2006).

(Erdmann et al., 2008) propose a method for creating a bilingual dictionary from interlanguage links, redirect pages and link texts. The number of backward links of a page is used to estimate the accuracy of a translation candidate because redirect pages with wrong titles or titles that are not related to the target page usually have a small number of backward links. The authors show the advantages of their approach compared to dictionary extraction from parallel corpora and manual crafting. (Rohit Bharadwaj G, 2010) discuss the iterative process of mining dictionaries from Wikipedia for under-resourced languages, though their system is language-independent. In each step near comparable corpora are collected from Wikipedia article titles, infobox information, categories, article text and dictionaries built at previous phases.

(Yu and Tsujii, 2009) automatically extract bilingual dictionary from Chinese-English comparable corpora which is build using Wikipedia inter-language links. Single-noun translation candidates for the dictionary are selected by employing context heterogeneity similarity (a feature that claims that the context heterogeneity of a given domain-specific word is more similar to that of its translation in another language than that of an unrelated word in the other language) and then ranked with respect to dependency heterogeneity similarity (a feature that assumes that a word and its translation share similar modifiers and head).

There has also been research done on the effectiveness of the usage of bilingual dictionaries in machine translation. A bilingual dictionary can be used as an additional knowledge source for training of the alignment models. The parameters of the alignment models can be estimated by applying the EM algorithm. A dictionary is assumed to be a list of word strings (e, f) where e and f can be single words or phrases.

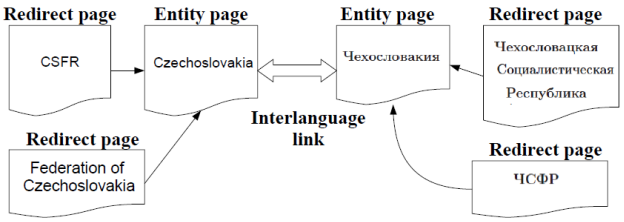
One of such methods of integrating of the dictionary into EM algorithm, described in (Brown et al., 1993), requires adding every dictionary entry (e, f) to the training corpus with an entry-specific count called effective multiplic-

[1] `http://folk.uio.no/angelii/wiki_dic.htm`
[2] `http://ufal.mff.cuni.cz/umc/cer/`

ity $\mu(e, f)$. Results of experiments in (Brown et al., 1993) showed that the dictionary helps to improve the fertility probabilities for rare words.
Another method described in (Och and Ney, 2000) suggests that effective multiplicity of a dictionary entry should be set to a large number if the lexicon entry occurs in at least one of the sentence pairs of the bilingual corpus and to low value if it doesn't occur in the corpus. The approach helps to avoid a deterioration of the alignment as a result of a out-of-domain dictionary entries.

3. Method

We created the Wiki-dictionary using the interlanguage links and redirect pages methods described in (Erdmann et al., 2008) (see Figure 1 for more details). The first assumption is that the titles of the articles connected by the interlanguage link are translations of each other. The second assumption is that the titles of redirect pages are the synonyms of the title of the target page. We collected titles of the articles conjoined by the interlanguage links and redirects from Wikipedia and created the dictionary from them.



Czechoslovakia | Federation of Czechoslovakia | CSFR

Чехословакия | Чехословацкая Социалистическая Республика | ЧСФР

Figure 1: The interlanguage links and redirect pages methods for the Wiki-dictionary development

We included in the dictionary the Russian-English translation pairs that are present in the Russian Wikipedia dump and are absent from the English Wikipedia dump and the English-Russian translation pairs that are present in the English Wikipedia dump and are absent from the Russian Wikipedia dump. We have such data because of two reasons: first, the dumps were made on different dates, during this gap Wikipedia editors made changes to the encyclopedia, second, some articles have only one-way mappings, e.g. there is an interlanguage link from Russian article to English article but there is no interlanguage link from this English article or any of its redirect pages to the given Russian article. For example, Russian article "Случайные знаки" has an interlanguage link to the English article *"Accidental (music)"*. The latter article has a bi-directional interlanguage link with the article "Альтерация (музыка)" which means it is not connected with the article "Случайные знаки" in English-Russian direction.

4. Evaluation

In order to estimate the proportion of named entities in the Wiki-dictionary, we used the heuristics suggested in (Bunescu and Pasca, 2006) and some additional heuristics (e.g. a one-word title is a named entity if it contains at least one capital letter and at least one digit). The numbers show that 88% of the translation pairs are named entities while only 12% are non-named entities (non-NEs). For comparison, only 7.5% of entries in the traditional Mueller's dictionary contain named entities.
Having a large percentage of named entities in the Wiki-dictionary, it was interesting to see the distribution of classes of named entities. We performed named entity recognition and classification in order to learn more about the content of the dictionary. Using Wikipedia's own category system we labeled the Wiki-dictionary with the standard named entity tags (PER, LOC, ORG, MISC) which can be further used by the information extraction tools. We implemented a bootstrapping algorithm for the named entity classification task (Knopp, 2010). Each named entity class is represented as a vector of Wikipedia categories and the algorithm computes similarity between the category vectors of unclassified articles and the named entity class-vectors in each iteration. The class with the highest similarity score is assigned to the corresponding articles and the categories of these new classified articles are added to the vectors of their respective named entity class.
We manually marked-up a random sample of 300 dictionary entries and found out that the results of the automatic named entity recognition had an accuracy rate of 76.67% and the true distribution of the classes on the sample was:

- 24.33% entities of class PER;
- 2.67% entities of class ORG;
- 29.33% entities of class LOC;
- 15.67% entities of class MISC;
- 72% named entities in total.

In order to evaluate the Wiki-dictionary we checked whether Wiki-dictionary covers the vocabulary of the unidirectional English-Russian dictionary by V. K. Mueller. We obtained a machine readable version of the Mueller's dictionary in four plain text files: abbreviations, geographical names, names and base dictionary. The size of the Muller's dictionary is shown in the Table 1 ("Names" is a list of personal names, "Base" is a list of translation pairs that are non-NE). The Wiki-dictionary contains 348,405 entries.
The algorithm works the following way. It searches for the exact match of the lowercased English word from the Mueller's dictionary in the Wiki-dictionary, e.g. we take a record

```
Czechoslovakia
_ист. Чехословакия
Transliteration: _ist. čexoslovakija
```

from the Mueller's dictionary and search for the word *"Czechoslovakia"* in the Wiki-dictionary. If the entry of the Wiki-dictionary with this word is found, we collect all the Russian translations from the Wiki-dictionary. In our example the corresponding Wiki-dictionary record would

Mueller's dictionary file	**Geographical names**	**Names**	**Abbreviations**	**Base**
Number of entries	1,282	630	2,204	50,695

Table 1: Size of Mueller's dictionary files

Mueller's dictionary file	**Geographical names**	**Names**	**Abbreviations**	**Base**
Recall of the Wiki-dictionary	82.18%	75.88%	22.64%	7.42%

Table 2: Recall of the Wiki-dictionary on the Mueller's dictionary

be (the entry is shortened):

Czechoslovakia | Federation of Czechoslovakia | Czechoslowakia | Czechaslavakia | CSFR
Чехословакия | Чехословацкая Социалистическая Республика | Чешско-Словацкая Социалистическая Республика | Чешско-Словацкая Федеративная Республика | ЧСФР
Transliteration: čexoslovakija | čexoslovackaja socialističeskaja respublika | češsko-slovackaja socialističeskaja respublika | češsko-slovackaja federativnaja respublika | čsfr

We concatenate all the lines of the translation part in the Mueller's dictionary in one line and for each translation from the Wiki-dictionary we check if it occurs as a substring in Mueller's dictionary translation.
The reason why we concatenate the translation part in one line and search the Wiki-dictionary translations as substrings, is that the Mueller's dictionary often provides an explanation of a term rather than just a simple translation.
The results of the evaluation are summarized in Table 2. The highest recall we obtain is for the geographical names, 82.18%, while for the names we have 75.88%. Surprisingly, the highest recall we have obtained for the abbreviations, even taking the English expansions of the abbreviations into the account, is only 22.64%. Recall for the base dictionary is only 7.42% which shows the low coverage of non-NEs in the Wiki-dictionary.
Words that are included in Wikipedia but not in the Mueller's dictionary are largely (a) very specific terms (such as *monogeneric*, *apature*, *rem sleep parasomnias*, *tokamak*, *tropidoboa*) that are more likely to be present in field-specific dictionaries rather than in general lexicon and (b) particular named entities (local geographical names such as *Santana do Acaraú*, *Lake Semionovskoye*, *Emelyaianovski district*; names of public people such as *Edvard Speleers*, *Princess Theresa of Bavaria*, *Alberto Medina Briseno*, *William de Lyon*; football teams such as *FC Zauralets Kurgan*; car models such as *Mercedes-Benz W221*; etc.).

5. Machine Translation Experiments

For the machine translation experiments we used sentence-aligned ÚFAL Multilingual Corpora (UMC) and we chose the Moses[3] toolkit which is a complete machine translation system for academic research. UMC is a parallel corpus of texts in Czech, Russian and English languages created for the purpose of machine translation. The source of the content are news articles and commentaries from The Project Syndicate[4].
We were interested in the frequency of dictionary phrases in corpus data and we had a goal to do pre-evaluation of the corpus to find out whether we could use it for machine translation experiments with the dictionary. We therefore collected statistics of occurrences of the translation pairs from the Wiki-dictionary in the UMC. The evaluation was done by word forms (using a tokenized version of the dictionary) and by normal forms (using a tokenized lemmatized version of the dictionary and a normalized version of the corpus data). Results show that translation pairs from the Wiki-dictionary are present in the corpus but not to a large extent. Approximately 28% of the non-normalized sentence pairs from the training set don't contain any translation pairs from the Wiki-dictionary, while approximately 24.7% of the non-normalized training set contains exactly one translation pair from the Wiki-dictionary.
First, we performed several experiments without the Wiki-dictionary and achieved the highest BLEU score of 24.76 using the English monolingual data from Europarl corpus[5] as additional data for training a language model.
We then incorporated the Wiki-dictionary into the training set: the dictionary was split into pairs of synonyms and appended to the end of the UMC training set. The inclusion of a dictionary as an additional parallel corpus data is the standard method. But this resulted in a drop of BLEU score, the best value we got was 20.42.
We used paired bootstrap re-sampling to estimate the statistical significance of the the difference in BLEU score between the model created with and without the Wiki-dictionary. As the difference between BLEU scores of the systems was small, we couldn't be sure if we could trust automatic evaluation results that one system outperformed the other on the test set. Our question was if the difference in test scores was statistically significant.
The approach is described in (Koehn, 2004). We collected 1000 trial sets of the size 300 sentences from the original test set (which had the size of 1000 sentences) by random sampling with replacement. We computed BLEU score for both systems in question on each of the 1000 trial sets and calculated how many times one system outperformed the other.

[3] http://www.statmt.org/moses/

[4] http://www.project-syndicate.org/

[5] http://www.statmt.org/europarl/

We compared the models that were created without an additional corpus for language model training. The results are summarized in the Table 3. According to our evaluation, 3-gram model without the Wiki-dictionary is better than the model trained with the Wiki-dictionary with 98.5% statistical significance, 4-gram model is better with 96% statistical significance and 5-gram model is better with 87.1% statistical significance.

A possible explanation for the drop is the domain difference of the corpus and the Wiki-dictionary. UMC corpus contains texts from the collection of the news articles and commentaries from a single resource The Project Syndicate while Wikipedia is an Internet encyclopedia. Typically, the more data is used for the translation model training the higher translation performance can be achieved. However, the significant amount of out-of-domain data added to the training set cause the drop of the translation quality (Hildebrand et al., 2005). In such a case a general translation model that was trained on in-domain and out-of-domain data does not fit the topic or style of individual texts. For the ambiguous words the translation highly depends on the topic and context they are used in.

The UMC training set contained a significant number of sentences that comprised zero or only one word from the Wiki-dictionary. We believe that might mean that the domains of the Wiki-dictionary and the UMC corpus are quite different. We suppose that was the reason of the lower quality of the translation that we got from the model trained on the train set with the Wiki-dictionary incorporated in it.

Therefore we collected a new test set using the text of three articles from Wikipedia (Wiki-set). The text of the articles needed pre-processing. First, we converted MediaWiki text into plain text using the Java Wikipedia API (Bliki engine)[6] which is a parser library for converting Wikipedia wikitext notation to other formats. The class PlainTextConverter from this library can convert simple Mediawiki texts to plain text. Secondly, we removed that traces of template markup (e. g. {{cite web}}) that still remained after removing Mediawiki markup. Thirdly, we split the text into sentences with the script split-sentences.perl written by Philipp Koehn and Josh Schroeder as part of Europarl v6 Preprocessing Tools suit[7]. The tool uses punctuation and capitalization clues to split paragraphs of sentences into files with one sentence per line. Fourthly, we performed tokenization using the same script as in Chapter 2, the script tokenizer.perl from Europarl v6 Preprocessing Tools suit. Finally, we corrected the automatic tools errors and removed the remaining noise manually.

Both the UMC test set and the Wiki-set consist of 1000 sentences, but there are 22,498 tokens in the Wiki-set while the UMC test set contains 19,019 tokens. Since there is no gold standard, we manually compared the quality of the translations produced by the models trained with and without the Wiki-dictionary on two random samples of 100 sentences collected from the UMC test set and from the Wiki-set. Table 4 presents the results of this manual ranking. In most of the cases one of the systems was ranked higher than the other because of the better representation of the meaning of the original sentence. In many other cases the missing words and grammatical structure played the key role in the final decision. There were several pairs for which one translation was preferred against the other because of the vocabulary, as some synonyms suit particular contexts better than the other synonyms. The model trained without the Wiki-dictionary performs better on the sample from the UMC test set; it is ranked first on 55 sentences. This outcome corresponds to the BLEU evaluation results. The model trained with the Wiki-dictionary is ranked first on 50 sentences of the sample from the Wiki-set while the outputs of the two models are of indistinguishable quality on 6 sentences. This brings some evidence that the Wiki-dictionary can be useful when it is applied to the appropriate domain.

Out-of-vocabulary (OOV) words are the words of the source language that the machine translation system didn't manage to translate into the target language. The total number of OOV words is less for the model trained with the Wiki-dictionary on both test sets. As we expected there are many cases when the model trained without the Wiki-dictionary didn't translate named entities while the model trained with the Wiki-dictionary recognized and translated the named entities correctly. For example,

```
<s1>sociologist at yale university
immanuel валлерстайн believes that
by 2050 , lenin inevitably become a
national hero russia .  </s1>
<s2>marketing sociology at yale
university , immanuel wallerstein
believes that by 2050 , lenin
inevitably will be the national hero
russia .  </s2>
```

The number of OOV words is twice bigger on the Wiki-set while the sizes of the test sets are comparable. The increase in the number of OOV words is most likely caused by the shift of the topic.

6. Conclusions

In this work we evaluated a bilingual bidirectional English-Russian dictionary created from Wikipedia article titles. This dictionary is very different from the traditional Mueller's dictionary, e.g. most of the phrases and words are named entities, the recall of the common terminology is only 7.42% and at least 96% of the basic terminology that the Wiki-dictionary shares with the Mueller's English-Russian dictionary are noun phrases. Evaluation on the parallel ÚFAL Multilingual Corpora revealed that even though the translation pairs from the Wiki-dictionary occur in the corpus, there is a significant number of sentences (about 28%) that don't contain any terms from the Wiki-dictionary. Such statistics indicates that the dictionary doesn't properly cover the domain of this corpus. As a next step, we incorporated the Wiki-dictionary into a translation system. According to the BLEU score, paired bootstrapping, OOV words analysis and manual evaluation, the translation accuracy dropped compared with the models trained without the Wiki-dictionary. The difference in the domain of the corpus and the dictionary could explain this result. We got

[6] http://code.google.com/p/gwtwiki/

[7] https://victorio.uit.no/langtech/trunk/tools/alignment-tools/europarl/

Model 1	Model 2	Statistical significance that model 1 outperforms model 2
3-gram	3-gram + Wiki-dict.	98.5%
4-gram	4-gram + Wiki-dict.	96%
5-gram	5-gram + Wiki-dict.	87.1%

Table 3: The results of the paired boostrap re-sampling show the statistical significance of the fact that the models trained without the Wiki-dictionary outperform the models trained with the Wiki-dictionary

	Model **without Wiki-dict** is ranked first, # of sent.	Model **with Wiki-dict** is ranked first, # of sent.	Translations are equally bad/good, # of sent.
sample of 100 sent. from **UMC** test set	**55**	37	8
sample of 100 sent. from **Wiki-set**	44	**50**	6

Table 4: Manual ranking of the results

some evidence to support this hypothesis in the new experiment on the test set collected from Wikipedia. We found that the model trained with the Wiki-dictionary performed better than the model trained without the Wiki-dictionary according to OOV words analysis and manual evaluation.

7. References

S. F. Adafre and Maarten de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty. 1993. But dictionaries are data too. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 202–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Razvan C. Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. An approach for extracting bilingual terminology from wikipedia. In *Proceedings of the 13th international conference on Database systems for advanced applications*, DASFAA, Berlin, Heidelberg. Springer-Verlag.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of EAMT*, Budapest, Hungary, May.

Johannes Knopp. 2010. Classification of named entities in a large multilingual resource using the Wikipedia category system. Master's thesis, University of Heidelberg.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 1086–1090, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nils Reiter, Matthias Hartung, and Anette Frank. 2008. A resource-poor approach for linking ontology classes to wikipedia articles. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 381–387, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vasudeva Varma Rohit Bharadwaj G, Niket Tandon. 2010. An iterative approach to extract dictionaries from wikipedia for under-resourced languages. ICON 2010, IIT Kharagpur, India.

K. Yu and J. Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *Proceedings of Machine Translation Summit XII*.

A Visualizing Annotation Tool for Semi-Automatically Building a Bilingual Corpus

Quoc Hung-Ngo

Faculty of Computer Science
University of Information Technology
Vietnam National University – HoChiMinh City
hungnq@uit.edu.vn

Werner Winiwarter

University of Vienna
Research Group Data Analytics and Computing
Universitätsstraße 5, 1010 Vienna, Austria
werner.winiwarter@univie.ac.at

Abstract

Bilingual corpora are critical resources for machine translation research and development since parallel corpora contain translation equivalences of various granularities. Manual annotation of word alignments is of significance to provide a gold-standard for developing and evaluating both example-based machine translation models and statistical machine translation models. The annotation process costs a lot of time and effort, especially with a corpus of millions of words. This paper presents research on using visualization for an annotation tool to build an English-Vietnamese parallel corpus, which is constructed for a Vietnamese-English machine translation system. We describe the specification of collecting data for the corpus, linguistic tagging, bilingual annotation, and the tools specifically developed for the manual annotation. An English-Vietnamese bilingual corpus of over 800,000 sentence pairs and 10,000,000 English words as well as Vietnamese words has been collected and aligned at the sentence level; and a part of this corpus containing 200 news articles was aligned manually at the word level.

Keywords: annotation tool, bilingual corpus, word alignment

1. Introduction

In natural language processing, a bilingual corpus is a valuable resource. A huge bilingual corpus is not only used to train natural language processing (NLP) tasks effectively but also to evaluate NLP systems objectively, such as chunking in bilingual text, bilingual comparison, bitext transfer, and machine translation.

In building corpora, developing tools is also as important as collecting data, aligning, and tagging linguistic information. If the corpus is built semi-automatically, it means it is tagged or corrected by annotators and by using annotation tools. Therefore, the visualization ability of an annotation tool helps annotators to review and correct the linguistic information as well as the whole document in the corpus. For this purpose, several tools have been researched and developed, such as the Yawat tool of Ulrich Germann (2008), the Cairo tool of Smith and co-authors (2000), annotation tools for parallel treebanks by Yvonne S. and Martin V. (2007), or tools for a Japanese-Chinese parallel corpus by Yujie Zhang and co-authors (2008).

For the English-Vietnamese language pair, there exist several projects for building an English-Vietnamese corpus for special purposes, such as building a bilingual corpus for word sense disambiguation by Dinh Dien(2002), and building a bilingual corpus through web

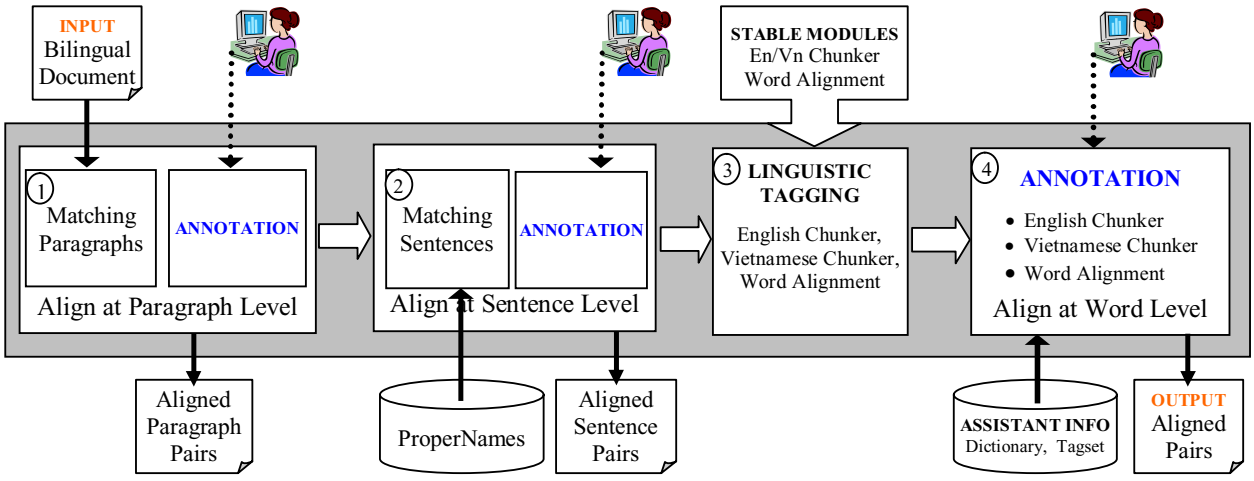


Figure 1: Overview of Building Bilingual Corpus Process

mining by Van D. B. and Bao Quoc H. (2007). However, most of these corpora are not available for download or just at the aligned sentence level.

In this paper, we describe the design of an annotation tool for building an English-Vietnamese Bilingual Corpus (EVBCorpus). More specifically, the goal is to build and annotate a large bilingual corpus which is tagged with linguistic information, such as part-of-speech, chunks, bitext alignment at the word level, and more. This bilingual corpus can then be used for the automatic training of machine translation systems.

In this work, we use three main stages. Firstly, we collect the data from the Internet and classify it based on the type of text as well as categories. Collected data is also normalized to reduce errors and to create a unique format between two languages. Secondly, we use NLP toolkits to tag linguistic information. Finally, a tool for annotation is built to annotate and correct linguistic tags, which have been assigned before.

Figure 1 shows the process of bilingual corpus building, including three main modules: pre-processing, linguistic tagging, and bilingual annotation. In particular, the pre-processing steps include (1) matching paragraphs and (2) matching sentences. These steps also need annotation to ensure that the result of these steps are English-Vietnamese sentence pairs. These bilingual pairs are tagged linguistically by the tagging modules (3), including English chunking, Vietnamese chunking, and English-Vietnamese word alignment. The aligned source and target chunks can be corrected as chunking result, alignment result as well as Vietnamese word segmentation result at the bilingual annotation stage (4). The Vietnamese word segmentation result can be corrected at this stage because the Vietnamese chunking module includes a word segmentation module.

2. Data

The EVBCorpus consists of both original English text and its Vietnamese translations, and original Vietnamese text and its English translations. The original data is from books, fictions or short stories, law documents, and newspaper articles. The original articles were translated by skilled translators or by contribution authors, and were checked again by skilled translators. Parallel documents are also chosen and classified into categories, such as economy, entertainment, health, science, social and politics, and technology.

Each article was translated one to one at the whole article level, so we first need to align paragraph to paragraph and then sentence to sentence. At the paragraph stage, aligning is simply moving the sentences up or down and detecting the separator position between paragraphs for both articles. At the sentence stage, however, aligning is more complex and depends on the translated articles which are translated by one-by-one method or a literal meaning-based method. In many cases (as common in literature text), several sentences are merged into one sentence to create the one-by-one alignment of sentences. The details of the corpus are listed in Table 1.

Source	Document	Paragraph	Sentence	Word
En-Vn Books	15	13,980	80,323	1,375,492
En-Vn Fictions	100	192,723	590,520	6,403,511
En-Vn Laws	250	86,803	98,102	1,912,055
En-Vn News	1,000	24,523	45,531	740,534
Total	**1,365**	**318,029**	**814,476**	**10,431,592**

Table 1: Details of data sources of EVBCorpus

An important feature of the corpus is that it has been pre-processed at the basic linguistic level, namely that of words. Especially, in Vietnamese, tokens are not words, and a word can be a token or a group of tokens. Therefore, the first important step in pre-processing is a Vietnamese word segmentation which is just done to evaluate the corpus, whereas this step used for later processing is included in the Vietnamese chunking module. In our project, we use vnTokenizer of Le H. Phuong et al (2008) to segment words in Vietnamese text.

There are 10,431,592 English words and 10,298,531 Vietnamese words (containing 13,143,290 Vietnamese tokens) in our bilingual corpus (see Table 2). Vietnamese words are counted based on the result of using the vnTokenizer module on the Vietnamese text.

Based on the results shown in Table 2, it can be seen that the length of most sentences in the corpus is from 10 to 25 words, and books are the bitext type with the longest average sentences. An interesting characteristic is that there are over 4% quite long sentences which have more than 50 words per sentence, even one hundred words in several cases. Moreover, the average paragraph length is just under 5 sentences per paragraph. Books also have the

Sentence Length	~10	~20	~30	~40	~50	~60	~70	~80	~90	~100	~110	~120
En-Vn Books	9,719	14,265	10,772	5,990	3,058	1,398	657	294	183	92	54	28
En-Vn Fictions	248,699	157,588	63,117	22,587	7,828	2,608	976	400	161	86	52	34
En-Vn Laws	38,071	17,789	12,513	7,776	4,360	2,154	1,073	545	266	139	83	67
En-Vn News	9,065	12,660	7,168	2,360	686	184	34	20	9	6	3	2

Table 2: Number of English sentences for each length

highest number of sentences. We carry out these statistics to look for a sensible way of building an annotation tool at a later stage.

3. Design of Annotation Tool

To add the linguistic information to the corpus and reduce the amount of effort for annotating, we integrate the NLP modules into the annotation tool. For linguistic tagging, we tag chunks for both English and Vietnamese text. English-Vietnamese sentence pairs are also aligned word-by-word to create the connections between the two languages. The data of the corpus is stored in the HTML and SGML standard.

3.1. Standard for Data Storage

We use both the HTML and SGML standard to store and process the data. For visualization, our tool stores files of the bilingual corpus based on the HTML format (see following example). Web browsers can open and render the representation of the corpus file easily with this format. It is also easy to store and review pairs in the corpus as parallel text (see Figure 5 in Sect. 3.4). In the HTML source, tag *span* is used to define POS tags, tag *sub* is used to define chunks, and tag *sub* with class *sentence* is used to define S tags (for whole sentences).

Besides HTML format, our tool also supports to store and export the corpus files to the SGML format based on Ide's guidelines (Ide N., 1998). Moreover, as another phrase corpus, English-Vietnamese bilingual corpus files are stored in column format by our annotation tool.

An example of the visualization of the chunk result and its HTML source is shown in Figure 2.

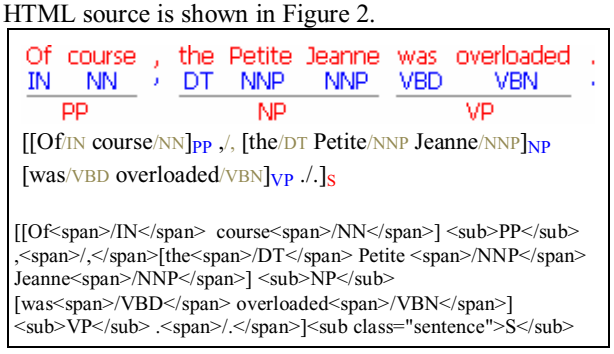


Figure 2: An example of chunking result and its HTML source

For the SGML format, the entire sentence is bracketed by tag *sentence*. Phrase structures are represented with tag *chunk*. The attribute *cat* represents the phrase symbol of a phrase. For example, the noun phrase "the Petite Jeanne" is represented as "<chunk cat="NP">the Petite Jeanne</chunk>". The next element is tag *wrd*, which is used to present words. The attribute *pos* represents the part-of-speech of a word. This is also similar to tokens in English text, however, it can be a group of tokens in Vietnamese text. The smallest element tag is *tok*. Each word in English and token in Vietnamese text is bracketed by *tok* tag.

```
<sentence id="s0"><chunk id="c0" cat="PP">
<wrd id="w0" pos="IN"><tok id="t0">Of</tok></wrd>
<wrd id="w1" pos="NN"><tok id="t1">course</tok></wrd>
</chunk><tok id="t2">,</tok><chunk id="c1" cat="NP">
<wrd id="w2" pos="DT"><tok id="t3">the</tok></wrd>
<wrd id="w3" pos="NNP"><tok id="t4">Petite</tok></wrd>
<wrd id="w4" pos="NNP"><tok id="t5">Jeanne</tok></wrd>
</chunk> <chunk id="c2" cat="VP">
<wrd id="w5" pos="VBD"><tok id="t6">was</tok></wrd>
<wrd id="w6" pos="VBN"><tok id="t7">overloaded</tok></wrd>
</chunk><tok id="t8">.</tok></sentence>
```

The encoding indicates that the translation text and its chunk tagging result is "[[Tất_nhiên/Np]$_{PP}$ [chiếc/Nc Petite_Jeanne/Np]$_{NP}$ [đã/R chở/V]$_{VP}$ [quá/T nặng/A]$_{AP}$ ./.]$_{S}$". The word alignment result in HTML format is "[1,2-1,2];[4-3];[5,6-4,5];[7,8-6,7,8,9]". It is stored in the SGML format as:

```
<links id=ls0 Xtarget="c0:c0">
   <linkw id=lw0 type=n:n Xtarget="t0,t1:t0,t1"></linkw>
   <linkw id=lw1 type=1:1 Xtarget="t3:t2"></linkw>
   <linkw id=lw2 type=n:n Xtarget="t4,t5:t3,t4"></linkw>
   <linkw id=lw3 type=n:n Xtarget="t6,t7:t5,t6,t7,t8"></linkw>
</links>
```

3.2. Linguistic Tagging

3.2.1 Chunking for English

There are several available chunking systems for English text, however, we focus on parser modules to build an aligned bilingual treebank in future. Based on Rimell's evaluation of five state-of-the-art parsers (Rimell, 2009), the Stanford parser is not the parser with the highest score. However, the Stanford parser supports both parse trees in bracket format and dependencies representation (Dan Klein et al, 2003; Marie-Catherine de Marneffe et al, 2006). We chose the Stanford parser not only for this reason but also because it is updated frequently, and to provide for the ability of our corpus for semantic tagging in future.

In our project, the full parse result of an English sentence is considered to extract phrases as chunking result for the corpus. For example, for the English sentence "*Products permitted for import, export through Vietnam's border-gates or across Vietnam's borders.*", the Stanford parser result is:

```
(S (NP (NNPS Products))
   (VP (VBD permitted)
     (PP (IN for)
        (NP (NP (NN import))
            (, ,)
            (NP (NN export))))
     (PP (PP (IN through)
         (NP (NP (NNP Vietnam) (POS 's))
             (NNS border-gates)))
         (CC or)
         (PP (IN across)
            (NP (NP (NNP Vietnam) (POS 's))
                (NNS borders)))))
     (. .))
```

Extracting chunks based on the Stanford parser result concentrates on noun and verb phrases rather than preposition phrases. The result of the extraction procedure for the example sentence is:

[Products]$_{NP}$ [permitted]$_{VP}$ [for]$_{PP}$ [import]$_{NP}$, [export]$_{NP}$ [through]$_{PP}$ [Vietnam's border-gates]$_{NP}$ [or]$_{PP}$ [across]$_{PP}$ [Vietnam's borders]$_{NP}$.

3.2.2. Chunking for Vietnamese

There are several chunking systems for Vietnamese text, such as noun phrase chunking by Le M. Nguyen et al (2008) or by Nguyen H. T. et al (2009). In our system, we use the full phrase chunker of Le M. Nguyen and Cao T. H. (2009) to chunk Vietnamese sentences. This is module SP8.4 in the VLSP project[1].

The VLSP project is a KC01.01/06-10 national project named Building Basic Resources and Tools for Vietnamese Language and Speech Processing. This project involves active research groups from universities and institutes in Vietnam and Japan, and focuses on building a corpus and toolkit for Vietnamese language processing, including word segmentation, part-of-speech tagger, chunker, and parser.

For example, the chunking result for the sentence "*Các sản phẩm được phép xuất khẩu, nhập khẩu qua cửa khẩu, biên giới Việt Nam.*" is "[Các sản_phẩm]$_{VP}$ [được]$_{VP}$ [phép]$_{NP}$ [xuất_khẩu]$_{VP}$, [nhập_khẩu qua]$_{VP}$ [cửa_khẩu]$_{NP}$, [biên_giới Việt_Nam]$_{NP}$.".

(In English: "[Products]$_{NP}$ [permitted]$_{VP}$ [for]$_{PP}$ [import]$_{NP}$, [export]$_{NP}$ [through]$_{PP}$ [Vietnam's border-gates]$_{NP}$ [or]$_{PP}$ [across]$_{PP}$ [Vietnam's borders]$_{NP}$.")

The chunking result also includes the word segmentation and the part-of-speech tagger result. These results are based on the result of word segmentation by Le H. Phuong, N. T. M. Huyen et al (2008). The tagset of chunking includes 5 tags: NP, VP, ADJP, ADVP, and PP.

3.2.3. Word Alignment in Bilingual Corpus

In a bilingual corpus, word alignment is very important because it demonstrates the connection between two languages. In our corpus, we apply a class-based word alignment approach to align words in the English-Vietnamese pairs. Our approach is based on the result of D. Dien et al (2002), to which we also contributed. This approach originates from the English-Chinese word alignment approach of Ker and Chang (1997). The class-based word alignment approach uses two layers to align words in a bilingual pair, dictionary-based alignment and semantic class-based alignment. The dictionary used for the dictionary-based stage is a general machine-readable bilingual dictionary while the dictionary used for the class-based stage is the Longman Lexicon of Contemporary English (LLOCE) dictionary, which is a type of semantic class dictionary.

Aligning words with a bilingual dictionary is estimating the distance *DTSim(s, t)* by using the meaning sets in the bilingual dictionary (*s* is a word in the source sentence and *t* is a token/word in the target sentence). Based on the collection of dictionary-based alignments, the model calculates the acquisition of pairs of mutually translatable classes (X, Y). Finally, aligning words based on classes is estimating the probability values *Pr(s,t)* based on the conceptual similarity *ClassSim(X, Y)* (*s* is a member of class *X* and *t* is a member of class *Y*) and the distortion probability *dis(i, j)* (*i* is the position of *s* in the source sentence and *j* is the position of *t* in the target sentence) (Dien Dinh et al, 2002; Ker et al, 1997). The result of the word alignment is indexed based on token positions in both sentences. For example:

English: I had rarely seen him so animated .
Vietnamese: Ít khi tôi thấy hắn sôi nổi như thế .

The word alignment result is [1-3], [3-1,2], [4-4], [5-5], [6-8,9], [7-6,7], [8-10] (visualized in Figure 3).

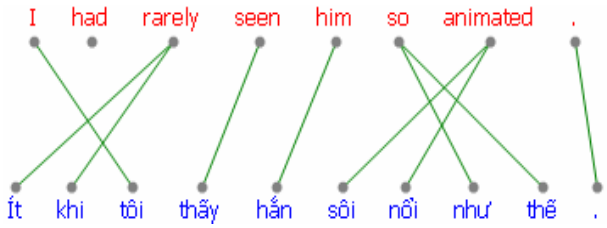


Figure 3: An example of word alignment in bilingual corpus

3.3. Word Alignment Visualization

Because of the huge value of bilingual corpora, numerous tools for the visualization and creation of word alignments have been developed. Most of them employ one of two visualization techniques. The first is to draw lines between associated words (as shown in Figure 3). The second is to use an alignment matrix (as shown in Figure 4), where the rows of the matrix correspond to the words of the sentence in one language and the columns to the words of that sentence's translation into the other language. Marks in the matrix's cells indicate whether the words represented by the row and column of the cell are linked or not.

Basically, with both visualization techniques it is easy to get an overview of the alignments at the word level, however, the drawing line technique has several advantages. For this technique, it is easy to combine the results of chunker modules and the parse trees for both

[1] http://vlsp.vietlp.org:8080/demo/

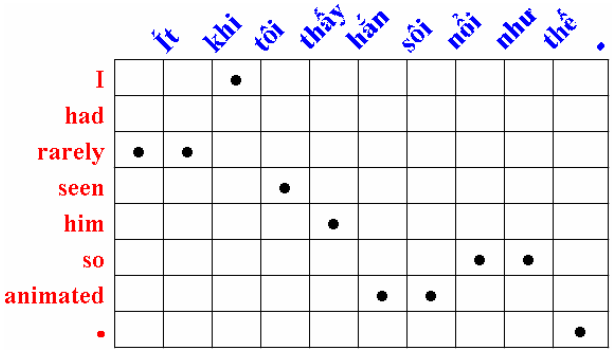


Figure 4: Visualization of word alignments with an alignment matrix

sentences (see Figure 6 in Sect. 3.4.3). It is also less space-consuming in case of lengthy sentence pairs. Because of these advantages, we use this technique in our annotation tool to demonstrate the word alignments of the English-Vietnamese sentence pairs.

3.4. Bilingual Annotation Process

As shown in Figure 1, there are three annotation stages in whole process, including matching paragraphs, matching sentences, and aligning words.

3.4.1. Matching Paragraphs and Sentences

In our system, before annotating for paragraph alignment, we use the Edit Distance algorithm to match sentences and split them into paragraphs by using the endline symbols of paragraphs in source document or target document. The string edit distance algorithm is sometimes known as Levenshtein distance. A very comprehensive and accessible explanation of the Levenshtein algorithm is available on the web at http://www.merriampark.com/ld.htm. The Levenshtein algorithm measures the edit distance where edit distance is defined as the number of insertions, deletions, or substitutions required to make the two strings match. A score of zero represents a perfect match. This algorithm has been applied to match names in English and Arabic by Freeman and co-authors (2006).

For matching paragraphs in both documents, it is essentially the matching of the sequence of sentences in these documents. This process is implemented by matching two strings where each sentence is represented by an element in the string. In our system, these elements are featured by merging a number of proper names and several special signs (such as question marks, exclamation marks, quotation marks, and so on).

With two strings, string s of size m and string t of size n, the algorithm has O(nm) time and space complexity. A matrix is constructed with n rows and m columns. The function $e(s_i,t_j)$ where s_i is a character in the string s, and t_j is a character in string t returns the value 0 if the two characters are equal and the value 1 otherwise. The algorithm extracts matched sub-sequences in both strings and then inserts zero values into the two strings so that they have equal length.

For example, string s is 003100210, representing the source document encoded with 9 sentences and sentence 3, 4, 7, and 8 having 3, 1, 2, and 1 proper names. Similarly, string t is 0030102100, representing 10 sentences in the target document with sentence 3, 5, 7, and 8 having 3, 1, 2, and 1 proper names. Our algorithm based on the Edit Distance algorithm tries to insert the value 0 into both strings and match characters as much as possible. The result in this example is 00301002100 with the length of 11 sentences. This result is decoded with two blank sentences which are inserted into s after sentence 3 and sentence 9.

3.4.2. Annotation for Sentence Alignment

The first stage of building a bilingual corpus is a bitext alignment, which aligns paragraph by paragraph and then sentence by sentence. Firstly, documents are manually segmented into chapters. These chapters are segmented into paragraphs by endline symbols. Basically, paragraphs in both languages are ordered as a sequence and there is rarely a change in order among paragraphs between a document pair. However, the merging and splitting of paragraphs occurs more frequently. In the next stage, paragraphs and sentences in two parallel documents are automatically aligned by the Levenshtein Edit Distance algorithm based on the number of proper names in each sentence. Finally, automatically aligned paragraph pairs are reviewed and corrected by annotators by using our tool.

For visualization, our tool simply shows paragraph pairs in each row (see Figure 5). Therefore, if the alignment of the previous pair is incorrect, the following pairs are incorrect, too. In addition, paragraph pairs with incorrect alignment have usually differences in paragraph length. In contrast, paragraph pairs with correct alignment are quite similar. Therefore, while scrolling through chapters and documents, annotators can identify the differences quickly and concentrate on correcting them. Our tool also supports to drag and drop paragraph items on paragraphs in order to merge paragraphs and to cut a paragraph into smaller paragraphs at the end of a particular position by pressing a hotkey.

3.4.3. Annotation for Word Alignment

Based on the results of the English chunking module, the Vietnamese chunking module, and the word alignment module in step 3 of the process (see Figure 1 with an explanation in the Section 3.2) , the parallel sentence pairs are linked together at the chunk level (see Figure 6).

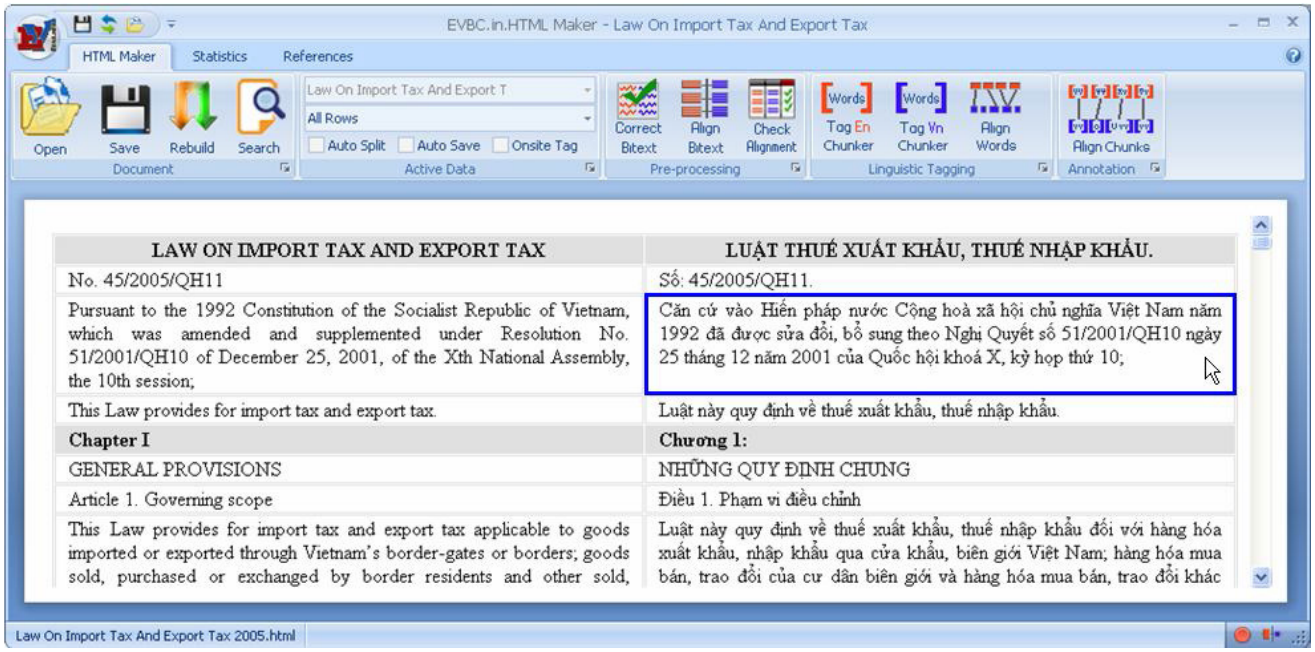


Figure 5: Drag and droppable interface of the tool for manual paragraph alignment annotation

With the visualization provided by our tool, annotators review whole phrase structures of English and Vietnamese sentences. They can compare the English chunking result with the Vietnamese result and correct them in both sentences.

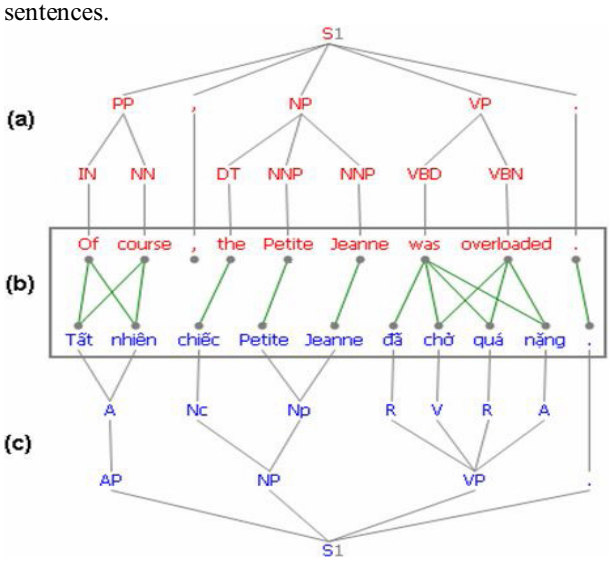


Figure 6: Combine English chunking (a), Vietnamese chunking(c), and word alignment (b)

Moreover, mistakes regarding word segmentation for Vietnamese, POS tagging for English and Vietnamese, and English-Vietnamese word alignment can be detected and corrected by drag, drop, and edit label operations (actions) of our tool. Based on drag and drop on labels and tags, annotators can change the results of the tagging modules visually, quickly, and effectively.

Different from paragraph alignment, which is based on chapter or document level, the word and chunk alignment is based on paragraph level with 2 to less than 5 sentences for each paragraph on average (as shown in Table 2). With the linguistic information including word/token, POS tag, chunking tag and word alignment, each sentence pair can be presented in one screen page. For long and complex sentences, annotators can scroll the horizontal scrollbar to view and correct the hidden part.

3.5. Details of Annotation Tool

In general, annotators have a good knowledge of linguistics, however, they have limitations in understanding formats for NLP corpora, which are normally used to process on computers. Moreover, for building a valuable corpus, the amount of annotation is very huge. Therefore, our goal is to develop a tool for annotating a corpus visually, quickly, and effectively at the alignment level of sentences, words, and chunks.

Drag and drop actions are mainly a convenient feature of the annotation tool. It allows annotators to drag a node (a word), a part of tree (a phrase), or multi-selected parts, and drop the item(s) on another node of the other tree to create alignments. For convenience purposes in annotating lengthy sentences, our tool also supports to grip the whole view and move it horizontally or vertically instead of clicking on the scrollbars. The parse trees can be expanded or collapsed to see the full details of sentences, or just an overview, or a part of long sentence pairs. Aside from mouse control, hotkeys are set up for the annotation tool. These hotkeys help annotators to navigate among

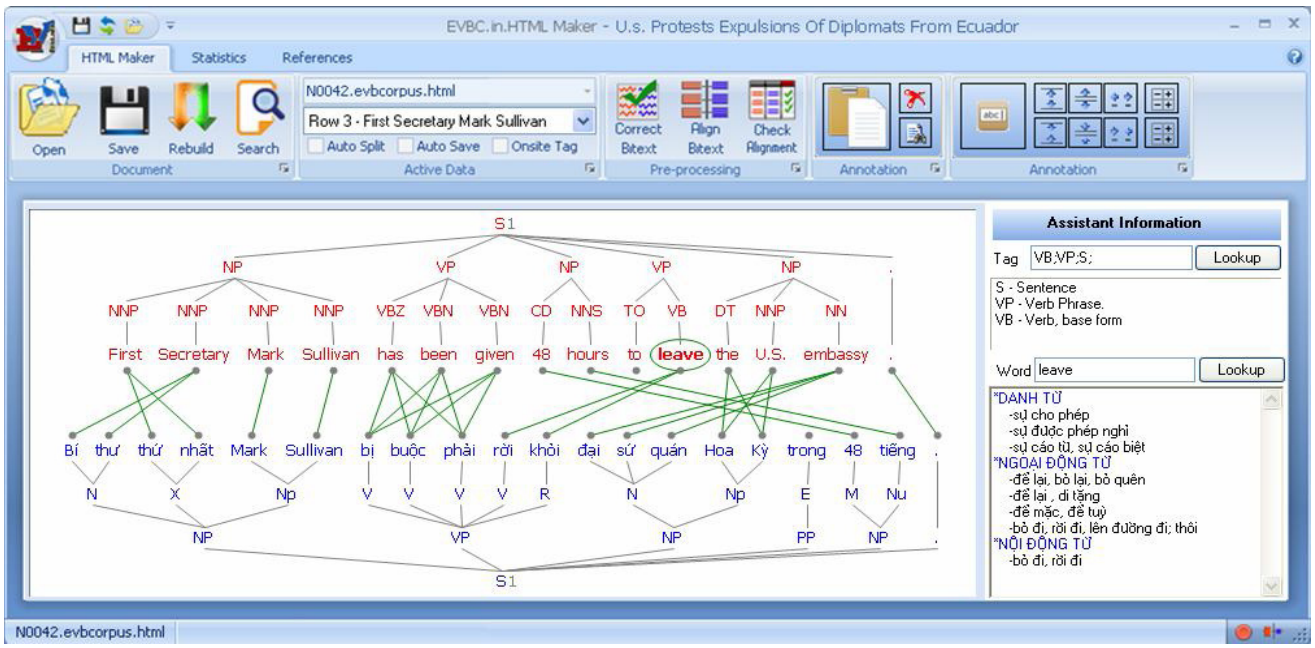


Figure 7: Overview of annotation tool for manual word/chunk alignment annotation

pairs, or to make/remove alignments.

Moreover, linguistic assistant information is shown following the annotator's actions. This assistant system accesses dictionaries to look up and show the meaning of the current word at the cursor (see Figure 7). Our annotation tool also supports both sure alignments and possible alignments which are two types of alignments.

4. Results and Analysis

4.1. Bilingual Corpus

From four resources, we built an English-Vietnamese bilingual corpus with over 800,000 sentence pairs and 10,000,000 words. This corpus is tagged with chunker labels for both English and Vietnamese, and aligned at word level. We also developed an annotation toolkit by integrating NLP modules for tagging, and a drag and droppable interface module for annotating. Our overall process illuminates four main steps of building a parallel corpus: (1) collect data and align bitext at the paragraph level; (2) align bitext at the sentence level, (3) linguistic analysis and tagging; (4) annotate and correct corpus with toolkits.

As a main result of the project, we built an English-Vietnamese bilingual corpus with 1,217 documents, over eight hundred sentences, and over ten million words from four resources: books, literal novels, law documents, and news articles. As mentioned in Section 2.1, all of these documents are collected and aligned as chapter-to-chapter (for books, novels, and laws), or article-to-article (for news articles) at first. Next, they are semi-automatically separated to align at the paragraph level, and at the sentence level at last. However, we still keep the context of paragraphs and sentences, which is very useful for other tasks in several machine translation models, such as document classification before translating or detecting the context of words in documents. A part of this corpus and the annotation tool are published at http://code.google.com/p/evbcorpus/ .

4.2. Annotation Process

The annotation process costs a lot of time and effort, especially with a corpus of over 10 million words for each language. In our evaluation, we annotated 200 news articles with 6,723 sentence pairs, and 116,246 English words (125,762 Vietnamese words and 164,447 Vietnamese tokens), as shown in Table 3.

	English	Vietnamese
Files	200	200
Sentences	6,723	6,723
Words	116,246	125,762
Tokens	116,246	164,447
Sure Alignments	70,238	70,238
Possible Alignments	88,964	88,964
Words in Alignments	90,581	121,271
Tokens in Alignments	90,581	151,905

Table 3: Details of Aligned EVBCorpus at word level

In this evaluation, the data is tagged and aligned automatically at the word level between English and Vietnamese and we just focus on the set of alignments and amount of annotation rather than evaluate the quality of the linguistic tagging modules. The number of alignments in 200 news articles is 89,222 alignments, which are aligned automatically by the word alignment module (as mentioned in Section 2.3.2) and checked and linked manually by annotators.

Alignments are annotated with both sure alignments *S* and possible alignments *P*, with $S \subseteq P$. These two types of alignment are annotated to evaluate the alignment models by the Alignment Error Rates (AER) according to the specifications described by Och and Ney (2003). In 200 annotated news articles, there are 70,238 sure alignments, accounting for 78% of possible alignments (as shown in Table 3). These alignments mainly come from nouns, verbs, adverbs, and adjectives which are meaningful words in sentences. On the other hand, the 22% remaining possible alignments are mainly from prepositions in both English words and Vietnamese words.

5. Conclusion

In this paper we introduced a design of a visualizing method for word alignment annotation and a complete workflow to build an English-Vietnamese bilingual corpus: from collecting data, tagging chunks, aligning words in bilingual text, and developing an annotation tool for bilingual corpora. We showed that the size of our corpus with 200 English-Vietnamese aligned news article pairs at the word level is a valuable contribution to build a high quality corpus in the future. We pointed out that linguistic information tagging based on our procedure, including tagging and annotation, so far, stops at the chunk level.

6. References

Dan Klein and Christopher D. Manning (2003). Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Dien Dinh, (2002). Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation. In Proceedings of Workshop on Machine Translation in Asia, pp. 26-32.

Dien Dinh, Kiem H., Ngan N. L. T., Quang X., Toan N. V., Quoc Hung N., and Hoi P. P. (2002). Word alignment in English – Vietnamese bilingual corpus. Proceedings of EALPIIT'02, HaNoi, Vietnam, pp. 3-11.

Freeman, Andrew T., Sherri L. Condon and Christopher M. Ackerman (2006). Cross Linguistic Name Matching in English and Arabic: A "One to Many Mapping" Extension of the Levenshtein Edit Distance Algorithm, HLT-NAACL, New York, NY.

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. Proceedings of the First International Language Resources and Evaluation Conference, Granada, Spain, pp. 463 – 470.

Ker, Sue J. and Jason S. Chang (1997). A class-based approach to word alignment. Computational Linguistics, 23(2):313–343.

Germann, Ulrich (2008). Yawat: Yet Another Word Alignment Tool. Proceedings of the ACL-HLT 2008.

Le M. Nguyen and Cao, T. H. (2008), Constructing a Vietnamese Chunking System. Proceedings of the 4th National Symposium on Research, Development and Application of Information and Communication Technology, Science and Technics Publishing House, pp. 249-257.

Le M. Nguyen, Huong T. Nguyen, Phuong Thai Nguyen, Tu Bao Ho and Akira Shimaz (2009). An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models. The 7th Workshop on Asian Language Resources (in conjunction with ACL-IJCNLP).

Le H. Phuong, N. T. M. Huyen, R. Azim, H. T. Vinh (2008). A hybrid approach to word segmentation of Vietnamese texts. Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, LATA 2008, Springer LNCS 5196, Tarragona, Spain.

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning (2006). Generating Typed Dependency Parses from Phrase Structure Parses. LREC 2006.

Nguyen Huong Thao, Nguyen Phuong Thai, Nguyen Le Minh, and Ha Quang Thuy (2009). Vietnamese Noun Phrase Chunking based on Conditional Random Fields. Proceedings of the First International Conference on Knowledge and Systems Engineering (KSE 2009).

Rimell, L., Clark S., and Steedman M. (2009). Un-bounded dependency recovery for parser evaluation. Proceedings EMNLP, pp. 813–821.

Smith, Noah A. and Michael E. Jahr (2000). Cairo: An alignment visualization tool. Second International Conference on Linguistic Resources and Evaluation (LREC-2000).

Van B. Dang, Bao Quoc Ho (2007). Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining. Research, Innovation and Vision for the Future (RIVF), IEEE International Conference. pp. 261-266.

Yujie Zhang, Zhulong Wang, Kiyotaka Uchimoto, Qing Ma, Hitoshi Isahara (2008). Word Alignment Annotation in a Japanese-Chinese Parallel Corpus. LREC 2008.

Yvonne Samuelsson and Martin Volk (2007). Alignment tools for parallel treebanks. Proceedings of GLDV Frühjahrstagung 2007.

SMT systems for less-resourced languages based on domain-specific data

Lene Offersgaard[1], Dorte Haltrup Hansen[2]
[1, 2] University of Copenhagen, Center for Language Technology
Njalsgade 140, DK-2300 Copenhagen
E-mail:[1]leneo@hum.ku.dk, [2]dorteh@hum.ku.dk

Abstract

In this paper we show that good SMT systems for less-resourced languages can be obtained by using even small amounts of high quality domain-specific data. We suggest a method to filter newly collected data for parallel corpora, using the internal alignment scores from the aligning process. The filtering process is easy to use and is based on open-source tools. The domain-specific data are used in combination with other public available resources for training SMT systems. Automatic evaluation shows that relatively small amounts of newly collected domain-specific data result in systems with promising BLEU scores in the range of 52.9 to 60.9.
The LetsMT! platform is used to create the presented machine translation systems, where the flexible platform allows uploading the user's own data for training. The paper shows that the platform is a promising way of making SMT systems available for less-resourced languages.

Keywords: SMT, less-resourced languages, domain-specific SMT

1 Introduction

LetsMT! is an EU-project[1] with the aim of building a platform for user tailored machine translation and online sharing of training data. Here we report on resent results of training and evaluating SMT systems for three less-resourced European languages within the LetsMT! platform, all systems based on newly collected domain-specific data. When data are collected from new sources it is a challenge to ensure good parallel data quality. In this paper we suggest a method for filtering the collected data, using alignment scores. The filtered data are then used in combination with other public available resources for training SMT systems. The systems are trained in the LetsMT! platform, which enables the Moses SMT software to do the training with in-domain and out-of-domain language models. Automatic evaluation shows that relatively small amounts of good quality domain-specific data result in systems with promising evaluation scores. In this paper we therefore focus on the process from data collection, data filtering to the SMT system training and evaluation within the LetsMT! platform, a platform which gives the opportunity for new users to create their own domain specific SMT system of fairly good quality by means of limited quantity of in-domain data.

2 LetsMT! platform

The LetsMT! platform[2] allows users to upload their own data into a repository, which converts, store and handle data in a safe and functional way to prepare data for training standard SMT engines (Tiedemann et al. 2012). From an easy-to-use web-interface registered users can configure an SMT engine based on a combination of large public resources and other resources uploaded to the platform - either by the user itself or other users. An efficient cloud-based training can then be carried out based on the Moses SMT software[3] with the in-domain and out-of domain data handling described in Koehn and Schroeder, 2007 The LetsMT! platform also allows for integration in SDL Trados - integration with other CAT tools is under development, enabling easy use of the LetsMT! system for localization. For testing purpose and minor translation tasks a web-interface is available.

3 Domain issues in SMT

In LetsMT! data has been collected for a number of subject domains. Our assumption is the quality of automatic translation increases if the systems are trained on domain-specific data. In (Pecina et al., 2011) an approach of tuning existing general-domain systems with domain-specific data did not seem promising. In (Offersgaard et al., 2008) systems were trained on domain-specific data, but here a method weighing a domain-specific phrase table higher than a more general phase table showed an increase in BLEU and TER scores. In this paper we focus on the options given in the LetsMT! platform (Koehn and Schroeder, 2007), where in-domain and out-of-domain language models are weighted.

An important issue is to classify the data in named subject domains. In an ideal world it would be preferable if collected data could be classified in the same large-scale general subject classification system. Not only would it ease the identification of consistent and representative bilingual training data, it would also, via the fine-grained subject classification, increase the probability that the lexical coverage of a given SMT-system would be tuned for the texts to be translated. But unfortunately a large

[1]LetsMT! is supported by the European Commission's ICT Policy Support Programme and is running from Mar. 1st 2010 until Aug. 31 2012
[2] See http://letsmt.eu for theLetsMT! platform
[3] http://www.statmt.org/moses/

universal classification system involves too much administrative work (Rirdance&Vasiljevs, 2006) being a difficult task to classify collected data. In addition, subject classification systems do not take into account possible divergences in the data within the same subject domain, e.g. different companies may have chosen to have different specific company terminologies.

Besides, texts from the same subject domain will make use of very different writing styles in terms of sentence types and varieties in language usage according to the genre of the text. Marketing texts, for instance, may praise the features of the product while manuals focus on strict instructions on how to use the product.
Consequently, in principle it would be preferable to train SMT systems on texts with almost identical writing styles and within the same subject domain.

In LetsMT! we decided to have the limited number of 15 subject domains available. These subject domains include the 10 domains used in TAUS[4].When only a few broad domains are available while uploading data the user can easily select the most appropriate subject domain.

As a supplement to the subject domain specification, the user can also specify text type, a description of the corpus and other metadata for the corpus. This allows users to give detailed information, and to use this information when selecting data for training a specific SMT system.

4 Data collection

The LetsMT! platform gives the opportunity to train domain-specific systems based on data uploaded to the LetsMT! resource repository. The available data in the repository consist of the large and well-known publicly available corpora e.g. Europarl, DGT-TM Acquis Communitare and the Opus corpora, all resources often used for SMT systems. In the LetsMT! platform these resources serve as backbone for training the phrase table and building the language model. In addition to the public available resources domain-specific data for under-resourced languages is collected by the project.

One of these domains is *Business and financial news*. This domain is chosen as a use case for an on-line translation service of financial news into less-resourced languages. The data collected for the domain is annual reports, which have been harvested automatically from a selected list of web sites. Annual reports are mostly freely available on companies' web sites in pdf format.

Another subject domain in focus is *Education* for which administrative documents from Danish Universities were collected, mainly curricula. This use case takes advantage of the LetsMT! plug-in to SDL Trados. Danish universities have an increasing demand for translation of curricula since a large number of courses are now taught in English allowing foreign students an easy access to education in Denmark.

The data collection was not done by web crawling systems but by systematic conduction of relevant web sites to secure high quality of in domain parallel resources.

5 Filtering data

When data is collected automatically noise arises from different sources: the files might be broken or have different content than expected, the translations might not be totally parallel, the layout might have destroyed the text in the pdf-to-xml conversion etc. These factors consequently lead to bad sentence-alignment. Normally large amounts of data ensure to blur bad alignment, but in our set up where only little domain specific data is available, high quality data is required.

As filter we used the alignment types and alignment scores from the HunAligner[5] (Varga et al. 2005). The HunAligner first does a Gale&Church sentence-length based alignment and then builds an automatic dictionary based on this alignment and realigns the text. The aligner produces 0-alignments, when segments have no corresponding segments in the other language, and n:m-alignments, specifying that n segments in the source language correspond to m segments in the target language.

After collection the data were first converted into text, tokenized, converted into xml and aligned by the Uplug tools (Tiedemann, 2002). Then 0-alignments were removed and the average scores calculated for each document. In the filtering process our aims were twofold: we wanted to provide good quality data to the LetsMT! platform and we wanted to find methods for filtering the data automatically.

From the average alignment scores we have done manual inspection of documents with a low average score (< 2). It seemed, however, that this wasn't a sufficient clue for alignment quality. In Pecina et al., 2011 an absolute score of 0.4 was used to filter out bad alignment. Our observations are, however, that especially positive low scores are not reliable while negative scores, high positive scores and average scores for the entire document are more useable. We therefore investigated documents with a high per cent of negative alignments (> 10%). In this case all parallel documents were of a bad quality. We also inspected documents without negative alignments. Absence of negative alignments can either indicate a perfectly parallel translation, an English-English "translation" (the same file) or empty files. Finally we searched for the English word *the* in the non-English documents to spot false translations or

[4] http://www.translationautomation.com/

[5] http://mokk.bme.hu/resources/hunalign/

language pairs being swopped.

Annual reports	Swedish	Danish	Dutch
Av. score, all reports	2.92	3.1	3.57
Av. score < 2	17 %	8 %	14 %
Neg. scores > 10%	13.5 %	7.7 %	7.4 %
Neg. scores = 0%	4 %	3.5 %	14.8 %
% of documents with mixed languages	1.6 %	4.2 %	2.8 %
% of documents filtered out	16.1 %	14.7 %	19 %

Table 1: Parameters for data filtering.

Table 1 shows the distribution of average alignment scores for Swedish, Danish and Dutch annual reports and the percentage of documents filtered out on the background of the findings.

We suggest that high quality data in terms of being parallel and in-domain, can be filtered by using the negative alignment scores from the HunAligner. Our findings are that positive alignment scores are less reliable than negative scores and that the average percentage of negative scores is a very good indicator for the alignment quality of the document and therefore of the data quality. It is difficult to set a fixed cut-off limit but our manual investigations showed that a threshold of around 10 % negative alignments per document was the upper limit. The table below shows the sizes of the domain-specific corpora after filtering. These corpora are used for training the SMT systems described in the next sections.

Language pair and domain	**Words (English)**
English-Danish Annual reports	3 022 233
English-Dutch Annual reports	5 753 369
English-Swedish Annual reports	11 503 078
Danish-English Education	635 685

Table 2: Size of domain-specific corpora after filtering

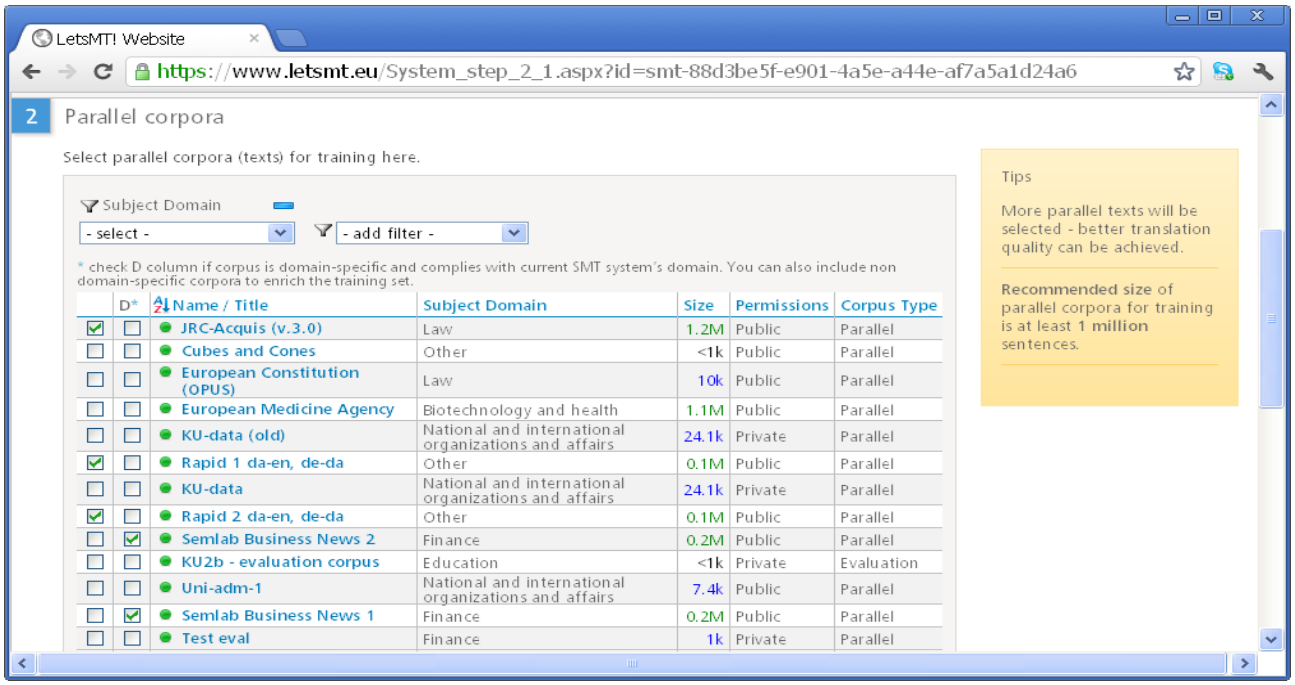


Figure 1: Selecting parallel corpora at the LetsMT! platform

Figure 1 shows how the LetsMT! platform allows the user to select parallel domain-specific and parallel general corpora when training an English-Danish finance SMT system.

6 LetsMT! system training

As reported in section 5 the collected in-domain data are of a relatively small size compared to often suggested amounts of training data for SMT systems. A minimum of 1 M parallel segments and 5 M mono-lingual segments for the language model are normally recommended by LetsMT!.

The LetsMT! Platform enables two ways of applying evaluation and tuning sets to the training process. Either the user can define the sets when configuring the training

process or the system can automatically extract sets of 1000 segments from the in-domain training corpora. In both cases - user-defined or automatic - the training data is afterwards cleaned-up for potential overlap between the training data and the selected tuning and evaluation sets. The evaluation sets used for the systems in section 7 and 9 are extracted automatically from the in-domain training data. In section 8 the same evaluation set is used for all systems.

7 Financial SMT systems

As a starting point we have trained three comparable financial SMT systems covering three different language pairs for the financial sub-domain 'Annual reports'. These three systems are trained using both in-domain and out-of-domain data. In table 3 and 4 the amounts of training data are shown. The in-domain data used are the corpora described in section 5. The out-of-domain parallel data for the English-Danish system is a corpus of EU press releases from Rapid, which can be seen as text from a general domain. For the English-Dutch and the English-Swedish systems the EU DGT Acquis corpus was used as out-of-domain data since we did not have a general corpus of original written text for these languages. The monolingual training data are a combination of the target language of the parallel data and the EU DGT Acquis corpus.

System	In-domain parallel data	Out-of-domain parallel data	Total
English-Danish Annual reports I	113 509	194 239	307 748
English-Dutch Annual reports I	307 807	360 449	668 256
English-Swedish Annual reports I	504 572	398 063	902 632

Table 3: Parallel training data (segments)

System	In-domain mono-lingual data	Out-of-domain mono-lingual data	Total
English-Danish Annual reports I	113 509	1 170 532	1 284 041
English-Dutch Annual reports I	307 807	379 225	687 032
English-Swedish Annual reports I	504 572	403 570	908 142

Table 4: Monolingual training data (segments)

The three systems are evaluated using the automatic measures: BLEU (Papineni et al. 2002), Meteor (Denkowski & Lavie 2011) and TER (Snover et al. 2006) (see table 5). The evaluation sets are extracted automatically.

System	BLEU	Meteor	TER
English-Danish Annual reports I	59.75	0.493	48.6
English-Dutch Annual reports I	52.89	0.368	52.3
English-Swedish Annual reports I	55.25	0.384	47.6

Table 5: Evaluation scores for the financial systems

Both the BLEU and the Meteor scores are calculated case-insensitive, to leave out casing issues from the evaluation. Meteor is used with the language independent option, not bringing all Meteor modules into play. Please mark that the TER score indicates the number of edits needed to adjust the translation output according to the reference translation. Therefore a lower TER score is better.

The scores show that all three systems have relative high BLEU and Meteor scores. The TER score reveals that even with these high BLEU scores the number of edits needed to adjust the translation outputs according to the reference translations are substantial – 48% to 52% changes. The evaluation scores cannot be compared among the three systems - and therefore we cannot state that one of the systems is better than the other two systems - as the evaluation corpora are different for the three systems. But we can see that the evaluation scores are very promising for these systems covering the financial sub-domain of 'Annual reports', even with different amount of in-domain and out-of-domain data.

8 More data or in-domain data?

The generally good results for the trained financial systems led us to train additional systems to see which factors made the biggest impact on the translation quality: the amount of data, the domain-specific data or the filtering of data.

For this experiment we focused on the English-Danish annual reports and trained 4 different systems: a baseline system containing only out-of-domain data (the EU DGT Acquis corpus and the EU press releases from Rapid), Annual reports I (as described in section 7), Annual reports II (only in-domain data) and Annual reports III (only in-domain data filtered for bad aligned files).

Systems (En-Da)	In-domain parallel data	Out-of-domain parallel data	Total
Baseline	-	897 548	897 548
Annual reports I	113 509	194 239	307 748
Annual reports II	113 509	-	113 509
Annual reports III	109 644	-	109 644

Table 6: Parallel training data (segments) for the En-Da annuals report systems

Systems (En-Da)	In-domain mono-lingual data	Out-of-domain mono-lingual data	Total
Baseline	-	1 170 532	1 170 532
Annual reports I	113 509	1 170 532	1 284 041
Annual reports II	113 509	-	113 509
Annual reports III	109 644	-	109 644

Table 7: Monolingual training data (segments) for the En-Da annuals report systems

The systems were tested on the same 1000 in-domain segments.

Systems (En-Da)	BLEU	Meteor	TER
Baseline	17.12	0.210	86.2
Annual reports I	59.75	0.493	48.6
Annual reports II	60.04	0.409	49.3
Annual reports III	60.91	0.413	46.8

Table 8: Evaluation scores for the En-Da annuals report systems

The evaluation scores show very clearly that domain-specific data increases the translations quality significantly. It is more surprising that the quality remains at the same level even when only a little amount of in-domain data is used. We believe that this might have to do with the special text type we are dealing with, namely annual reports. The vocabulary and the syntactic structures for this text type are relatively narrow. Finally, the evaluation scores show that filtering out the bad aligned documents gives a small additional improvement in both BLEU and TER even though only 3865 segments were removed.

9 Educational domain

To test if the same kind of quality can be achieved for other subject domains, we trained a system on the relatively small amount of curricula from Danish universities. The results show that with an in-domain parallel corpus containing only 0.6 M words (19,415 segments) and a general parallel corpus containing 526,302 segments, a BLEU score of 56.3 can be achieved.

System	BLEU	Meteor	TER
Danish-English Education with Acquis DGT	56.31	0.408	53.9

Table 9: Evaluation scores for Educational domain

Translators from the translation department on University of Copenhagen have inspected the translated evaluation set and find that the translations are very usable. They are currently evaluating the Danish-English Education system integrated in SDL Trados by the LetsMT! plug-in and report it being a very efficient way to include SMT in their translation workflow.

10 Conclusions

In this paper we describe the process from collection of new domain-specific data for less-resourced languages, filtering the data based on alignment scores, to training systems using the LetsMT! platform. Three systems for the same text type (annual reports) but for three different language pairs (Danish, Swedish, Dutch) were trained. The combination of in-domain and out-of-domain data shows promising automatic evaluation scores with BLEU scores from 52.9 to 55.3. The TER scores are 48% to 52%, revealing that even with the high BLEU scores the output still need quite some editing to match the translation references.

When collecting data from the web, some documents turn out to be lesser parallel as they might look at the first glace. We therefore present a usable method for filtering the collected data based on the negative alignment scores from the HunAligner. Our findings are that the average percentage of negative scores is a very good indicator for the alignment quality of the document. We suggest a cut-off limit of 10% negative alignments per document.

We also investigated the effect of both in-domain data and the amount of data on the translation quality. Results from a baseline system without in-domain data and a system with a combination of all available in-domain data and the same out-of-domain data as the baseline are presented. The system including in-domain data was significantly better than the baseline system with BLEU scores going from 17.1 up to 59.8. Furthermore systems based only on in-domain data – filtered and unfiltered - were trained. Surprisingly the BLEU score remained at the same level for the system with only in-domain data, namely 60.0 compared to 59.8 for the system with the much bigger amount of out-of-domain data. The filtered system showed a small additional improvement with a BLEU of 60.9.

We will conclude by saying that using the LetsMT! platform is a promising way of making SMT systems available for less-resourced languages. Users can now easily create tailored machine translation system taking advantage of the flexible way of including their own data for training SMT systems.

11 Acknowledgements

We want to thank our project partners in the LetsMT! project for collaboration. Especially we would like to thank Thomas Dohmen, SemLab, The Netherlands for web crawling the annual reports and Jörg Tiedeman, University of Uppsala for providing the Uplug tools.

12 References

Anil Kumar Singh and Samar Husain. Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs. In *Proceedings of ACL 2005 Workshop on Parallel Text*. Ann Arbor, Michigan. June 2005

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590-596.

Denkowski, M. and Lavie, A. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems, In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, 2011

Gale, W.A. and K.W. Church. 1994. .A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):pp75-102.

Koehn, P. and Schroeder J. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, 2007*, pages 224–227, Prague, Czech Republic

Lavie, A and Denkowski, M. The METEOR Metric for Automatic Evaluation of Machine Translation, *Machine Translation*, 2010

NIST 2005. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Retrieved 2010-04-17. *Machine Translation Evaluation Official Results*.

Offersgaard, L., Povlsen, C., Almsteen, L., Maegaard, B., Domain specific MT in use. In *Proceedings of the 12th EAMT conference*, 22-23 September 2008, Hamburg, Germany

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311–318.

Pecina,P et.a. Towards UsingWeb-Crawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the15th EAMT conference*, 2011.

Public project report LetsMT! D1.1 *Report on requirements analysis*, 2010, http://project.letsmt.eu

Rirdance, S. Vasiljevs, A,: Towards Consolidation of European Terminilogy Resources. *Experiments and Recommoondations from EuroTermBank Project*. Riga 2006

Snover, M., Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, A Study of Translation Edit Rate with Targeted Human Annotation, *Proceedings of Association for Machine Translation in the Americas*, 2006.

Tiedemann, J. Uplug - a modular corpus tool for parallel corpora. In *Parallel Corpora, Parallel Worlds*, pages 181-197, Rodopi, 2002.

Tiedemann, J, Hansen, D.H., Offersgaard, L., Olsen, S., Zumpe, M. A Distributed Resource Repository for Cloud-Based Machine Translation. . In *Proceedings of LREC 2012*

Towards a Wikipedia-extracted Alpine Corpus

Magdalena Plamada, Martin Volk

Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zürich
{plamada, volk}@cl.uzh.ch

Abstract

This paper describes a method for extracting parallel sentences from comparable texts. We present the main challenges in creating a German-French corpus for the Alpine domain. We demonstrate that it is difficult to use the Wikipedia categorization for the extraction of domain-specific articles from Wikipedia, therefore we introduce an alternative information retrieval approach. Sentence alignment algorithms were used to identify semantically equivalent sentences across the Wikipedia articles. Using this approach, we create a corpus of sentence-aligned Alpine texts, which is evaluated both manually and automatically. Results show that even a small collection of extracted texts (approximately 10 000 sentence pairs) can partially improve the performance of a state-of-the-art statistical machine translation system. Thus, the approach is worth pursuing on a larger scale, as well as for other language pairs and domains.

Keywords: Comparable corpus, Alpine texts, Wikipedia, Information Retrieval, Sentence alignment, Statistical machine translation, French-German

1. Introduction

The performance of Statistical Machine Translation (SMT) systems depends strongly both on the quality and the quantity of the training data. A well-known problem of SMT systems for most language pairs is the limited amount of bilingual parallel training data. The existing parallel corpora cover a relatively small percentage of possible language pairs and very few domains, thus building new ones involves considerable efforts, both in terms of time and costs.

In the last decade, less expensive but very productive methods of creating such sentence-aligned bilingual corpora have been proposed, based on the extraction of parallel texts from comparable texts. Zhao and Vogel (2002) introduced an adaptive approach for mining parallel sentences from a bilingual news collection, which combines a sentence length model with an IBM Model 1 translation model. Fung and Cheung (2004) combine bootstrapping methods and an IBM Model 4 model in order to exploit "very-non-parallel corpora" consisting of news stories from different sources.

The availability of comparable corpora and their potential for creating parallel corpora have sparked the interest of the SMT community. Munteanu and Marcu (2005) propose a maximum entropy-based classifier for identifying parallel sentences in newspaper articles by referring to a bilingual dictionary. They evaluate the extracted corpus by using it as training data for an SMT system. A similar approach is presented in (Abdul Rauf and Schwenk, 2011), with the difference that the authors of the latter paper use automatic translations instead of bilingual dictionaries and the selection relies on other metrics, such as word or translation error rate (WER, TER).

The approaches mentioned up to this point have been tested only on news corpora, but the expansion of the Web has drawn the attention towards another fruitful resource: web corpora. Adafre and de Rijke (2006) describe an MT based approach to find corresponding sentences in Wikipedia based on sentence similarity, without investigating the improvements of their method for a specific task (e. g. SMT, information extraction). Alternatively, Fung et al. (2010) also crawl comparable web sites (in particular, Wikipedia) in order to extract potential parallel sentences. The authors mention the improvement of SMT systems as one of the main objectives, but do not report any results.

As previously discussed, work in this field has focused mainly on two types of corpora (news and web corpora), with the purpose of extracting good training material for SMT. Nevertheless, not all papers present their results in terms of SMT improvements. There is also no claim about the performance of these approaches for a different domain. This represents our motivation to develop an approach inspired by earlier work, with the aim to extract a parallel corpus of mountaineering texts from Wikipedia. Moreover, we are interested in investigating to what extent the extracted corpus improves the performance of a domain-specific SMT system.

Wikipedia is an important multilingual resource available for a variety of domains, in almost 300 languages. It is not a parallel corpus because its articles in different languages are edited independently by users and are not literal translations of each other. However, often an article in one language contains a number of sentences translated from its corresponding article in another language. We identify and extract the parallel sentences in the Wikipedia articles and, moreover, we reduce the search space to one specific domain: Alpine texts (i. e. mountaineering reports, hiking recommendations, popular science articles about the biology and the geology of mountainous regions).

In the project Domain-specific Statistical Machine Translation[1] we have developed an SMT system trained for the Alpine domain. The training data comes from the Text+Berg corpus[2], which contains the digitized publications of the Swiss Alpine Club (SAC) from 1864 until

[1] http://www.cl.uzh.ch/research_en.html
[2] See www.textberg.ch

2011. The most relevant part for SMT training is the parallel German-French one representing a sizable corpus of approx. 5 million words. We therefore have the expertise to use in-house developed tools for the purpose of this experiment.
This article describes our approach for exploiting Wikipedia in order to produce more parallel texts for the Alpine domain. In section 2. we describe the extraction workflow, and in the subsequent section we evaluate the resulting corpus. The last section discusses future experiments and further improvements of the extraction method.

2. Extraction Methods

The general architecture of our parallel sentence generation process is shown in Figure 1. The approach was applied only to the language pair German-French, as these are the main languages of the Text+Berg corpus. However, the procedure can be applied with little effort to any of the available Wikipedias and any other domain. In our case, the input consists of German and French Wikipedia dumps[3], which are available in a special XML format, called MediaWiki[4].

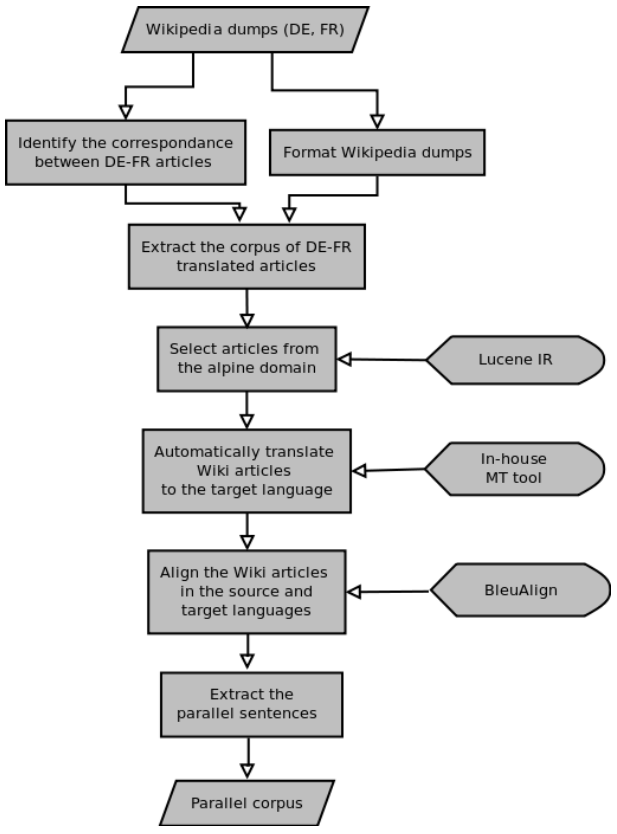


Figure 1: The workflow of the extraction algorithm

In the first step, we identify Wikipedia articles available in both languages by using the procedure described in (Lopez and Otero, 2010). We had to adapt the configuration files for French and German, as the original tool was developed for English, Spanish and Portuguese. The relevant output for our task represents the mapping between the titles of the articles available in both languages. The simplified XML structure proposed by CorpusPedia (Lopez and Otero, 2010) cannot be validated by usual XML parsers (e.g. DOM, SAX, ElementTree), so we need an additional tool for converting MediaWiki to valid XML.
For this purpose we used WikiPrep[5], a preprocessing tool that transforms the Wikipedia dumps to a simple XML format (Gabrilovich and Markovitch, 2006). The content of Wikipedia pages is converted to plain text with XML markups for section headers and internal links. MediaWiki is localized for all the languages supported in Wikipedia. We therefore had to customize the configuration files for German and French, so that MediaWiki elements (namespaces, templates, date and number formats etc.) can be correctly identified. After updating the files, we run the tool over the two Wikipedia dumps and then filter the articles available in both German and French.
Upon completion of this step, we have extracted a bilingual corpus of approximately 400 000 articles per language. The corpus is subsequently used for information retrieval (IR) queries aiming to identify the articles belonging to the Alpine domain. This procedure is detailed in the following subsection. Once we extract the Alpine comparable corpus, we proceed to the sentence alignment of the articles. The aim is to obtain a reasonably-sized set of sentence pairs that are likely to contain good data for our parallel corpus. This step is described in subsection 2.3.

2.1. Article classification in Wikipedia

In Wikipedia, articles are organized into topics and therefore they are assigned to one or more categories. This classification could allow us to extract articles on similar topics, in our case the topics of interest could be *Alpen*, *Berge* or *Ort*. However, many articles lack a category tag. This is the case with disambiguation articles, which distinguish between several contexts associated with an article title. For example, the article *Morgenstern* can refer to a planet, a medieval weapon, a magazine, a music band or a common last name for several personalities. Neither redirect articles, which automatically send the reader to another article, fall in any Wikipedia category. For instance, both pages *Wintersonnenwende* and *Sommersonnenwende* are linked to the more general article *Sonnenwende* (English: solstice). Apart from these cases, there are some arbitrary articles in our Wikipedia dump[3] that have no category tags. In fact, only 51.5% of the articles in the German Wikipedia have an assigned category. The remaining part consists of 33% redirect articles, 10% miscellaneous articles and 5.5% disambiguation articles. The percentages are similar in the French Wikipedia: 52.5% of the articles are categorised, 40% represent redirect articles, 4% mixed articles and 3.5% disambiguation articles.
Another interesting aspect is that articles are usually not placed in the most general category they logically belong to, if they are tagged as a subcategory thereof. For example, the article *Rosengartengruppe* is tagged with the following categories: *Bergmassiv (Dolomiten), Gebirge in*

[3] Accessed in September 2011
[4] http://www.mediawiki.org/wiki/MediaWiki
[5] http://sourceforge.net/projects/wikiprep/

Südtirol, Gebirge im Trentino, Dolomiten (English: massif in the Dolomites, mountains in South Tyrol, mountains in Trentino, Dolomites), but there is no reference to the Alps, although it is obvious that this mountain range belongs to the Alps. If we would like to use the Wikipedia classification as criterion for the extraction of domain-specific articles, we should come up with an extensive list of relevant categories. The categories in Wikipedia are sometimes very specific (e. g. *Berg im Kanton Appenzell Innerrhoden*), so compiling the list is not a trivial task. Besides, we would need an automatic classifier able to distinguish between relevant (e. g. *Bergführer*) and irrelevant categories (e. g. *Berg bei Neumarkt in der Oberpfalz*) for our corpus.

Another challenge for this task is that the categories assigned to the same article in different languages do not overlap. For example, the article *Trois Vallées* is tagged in German as *Wintersportgebiet in Frankreich, Alpen* (English: winter sports resort in France, Alps), whereas in French it belongs to the following categories: Tourisme en Savoie, Domaine skiable (English: tourism in Savoy, ski area). Identifying the semantic relationships between the German and the French categories is also not an easy task for a reasoner. One would therefore need to compile separate category lists for both German and French, as a simple translation of the categories from the other language would not help. This is not an isolated case in Wikipedia, but a general trend, as tables 1 and 2 show. They illustrate the distribution of the Wikipedia categories for the first 10 000 articles extracted with our approach (see section 2.2.). The German part contains 17 000 categories and the French one 16 000 categories, but more than 50% of them appear only once.

Category	Number of articles
Mann	1 278
Berg in Europa	325
Deutscher	279
Berg in den Alpen	210
Autor	190
Schweizer Gemeinde	157

Table 1: The most frequent categories in the top 10^4 German articles retrieved by Lucene

In the German Wikipedia, however, the leading category *Mann* (English: man, person) covers approximately 13% of the articles. As this category is rather general, we inspected the other categories assigned to these articles. We found that more than 90% of the articles were tagged with categories such as *Bergsteiger, Geograph, Entdecker, Extremsportler, Bergführer* (English: alpinist, geographer, explorer, extreme athlete, mountain guide). This proves that the retrieved articles are consistent with our domain of interest. Approximately 20% of them were also tagged with *Deutscher* (English: German), an expected percentage considering their corresponding values in table 1.

The next ranked categories cover significantly less articles (approx. $2 - 3\%$), but they obviously represent what we would expect in such a corpus (e. g. mountains in Europe and in the Alps, respectively). These results prove the accuracy of our extraction approach.

Category	Number of articles
Film américain	140
Ville de Bade-Wurtemberg	134
Ville de Rhénanie-du-Nord-Westphalie	121
Sommet des Alpes autrichiennes	98
Ville de Bavière	75
Sommet des Alpes suisses	64

Table 2: The most frequent categories in the top 10^4 French articles retrieved by Lucene

The French categories are much more diverse, therefore none of them covers a significant percentage of the articles. The leading category in the French Wikipedia, *Film américain* (English: American movie), is rather unexpected for this domain and belongs to the false positives in our results. However, the value is comparable to the following positions in the hierarchy, which are all relevant for our domain.

Taking all these aspects into consideration, we considered the extraction of domain-specific articles by means of the Wikipedia categorization time-consuming. We therefore decided to use an information retrieval-based approach, which will be detailed in section 2.2.

2.2. Extracting domain-specific articles

In order to extract the articles belonging to the Alpine domain, we have performed IR queries over the French and German Wikipedia. The input queries contained the 100 most frequent mountaineering keywords in the Text+Berg corpus (e. g. *Alp, Gipfel, Meter, Berg* in German and *montagne, sommet, mètre, cas* in French). The keyword lists are not translations of each other, as the term frequencies have been computed separately for German and French, respectively. However, they share common terms in the Alpine domain, such as *mountain, peak, meter*.

The extraction tool is based on the Lucene API[6], an open-source IR library. As Lucene does not have a module for morphological analysis, the reported results are based only on word-matching. We have decided to restrict the keywords to common nouns due to their limited inflectional variation. Lucene returns a list of the articles relevant to our query, ranked by their similarity score[7]. The score takes into consideration several factors such as term frequency, inverse document frequency, number of matched query terms etc.

Upon completion of this step our corpus was reduced to approx. 150 000 parallel articles. This value should be regarded with caution, as it stands for all articles that contain at least one occurrence of the top 100 Text+Berg keywords. Therefore in our experiments we use only articles that report a Lucene score above a certain threshold. The choice of the threshold depends highly on the targeted accuracy

[6] http://lucene.apache.org

[7] http://lucene.apache.org/core/old_versioned_docs/versions/3_0_0/scoring.html

and the task itself, as the similarity scores are sometimes misleading. It is possible that a short article about less important mountains (e. g. *Gurktaler Alpen*, similarity score: 0,010 97) receives a lower score than a long article about a collection of novels (e. g. *Die Arbeiten des Herkules*, similarity score: 0,034 29). Table 3 shows a selection of the articles with the highest scores in the German Wikipedia.

Title	Score
Reinhold Messner	0,089 43
Britische Mount-Everest-Expedition 1924	0,080 52
Hans Kammerlander	0,070 07
Ortler	0,069 66
Mount Everest	0,062 15
Mont Blanc	0,053 64

Table 3: The best ranked Alpine articles in the German Wikipedia according to Lucene

In contrast, table 4 presents the best articles in the French Wikipedia, sorted by their relevance according to Lucene. The French ranking differs from the German one firstly because the keyword lists partially contain different nouns. On the other hand, the content of the articles (including their structure and length) highly varies among the language variants of Wikipedia.

Title	Score
Lure	0,059 58
Parc national de Glacier	0,059 40
Mont Kenya	0,057 72
Nez-Percés	0,057 53
Mont Ventoux	0,057 15
Mont Blanc	0,057 09

Table 4: The best ranked Alpine articles in the French Wikipedia according to Lucene

However, the hit lists may also contain overlapping content, such as the article about *Mont Blanc* in the previous examples. An interesting finding is that the first hit for the French Wikipedia is an article about the city of Lure, which apparently does not have much in common with our topic, mountains. Taking a closer look at the whole article explains the score, as is contains thorough sections about the geology, the topography, the hydrology, and the climatology of the place, which are all areas closely related to mountaineering.

2.3. Extracting aligned sentence pairs

We use the Bleualign algorithm (Sennrich and Volk, 2010) for extracting parallel sentences from two Wikipedia articles. The aligned sentences (beads) are identified by means of an intermediary machine translation of the source. In our case, the translation is performed by our in-house SMT system trained on Alpine texts. Bleualign generates all possible sentence pairs between the automatic translation and the targeted article and computes for each of them the BLEU score (Papineni et al., 2002). Subsequently it reduces the search space by keeping only the 3 best-scoring alignment candidates for each sentence. Finally the algorithm returns the alignment pair which maximizes the BLEU score and respects the monotonic sentence order.

The algorithm can be applied in both directions. Translation direction does not matter in general, but we have decided to translate from French to German. We chose the French texts as the source texts because they are generally shorter. As the algorithm tries to align as much sentences as possible, this choice of the source texts allows us to maximize precision. In order to obtain a high-precision sentence alignment, Sennrich and Volk (2010) proposed computing the alignments in both directions, intersecting the results and then discarding all beads that differ between the two runs. For our purposes, however, we compute the alignments in a single direction (French-German).

In the end we filter the results once more by choosing only the 70% best-ranked alignments. The resulting set of alignment pairs represents a corpus containing semantically equivalent sentences.

As an example, the following sentence pair is a candidate for our parallel corpus which obtains the highest BLEU score.

FR: ainsi , la partie nord de l' himmelschrofenzug se compose de dolomite tandis que la partie sud se compose de roches du lias de la couche de l' allgäu

Automatic translation: damit ist der nördliche teil des himmelschrofenzug besteht aus dolomit , während der südliche teil besteht aus felsen des lias der schneedecke , das allgäu

DE Reference: so besteht der nördliche teil des himmelschrofenzugs aus hauptdolomit. der südliche teil besteht aus liasgesteinen der allgäudecke , die auf den hauptdolomit aufgeschoben worden sind

It is worth noting that the BLEU score is not computed between the source and target sentences, but between the automatic translation and the target sentence. This is how the BLEU values in Table 5 should be interpreted. Although the automatic translation is not perfectly correct, one notices that the word overlap between the translation and the target sentence is rather high. This explains why the extra tail in the German reference *die auf den hauptdolomit aufgeschoben worden sind* is not penalized by the BLEU score. Moreover, this example clearly shows that the algorithm deals not only with 1-to-1 alignments, but also with 1-to-n alignments.

3. Experiments and Results

3.1. Experimental setting

In this experiment we selected the top 4 000 ranked Wikipedia articles retrieved by Lucene. For this purpose we have merged the German and French lists and sorted the resulting list by the similarity score that Lucene provides. The articles have been sent to Bleualign for sentence alignment, using a customized configuration. We put great value on translations' fluency, so we measured the BLEU score on 3-grams, instead of 2-grams, as proposed by Sennrich and Volk (2010). In addition, we decided not to use any gap filling heuristics, because of the great variation of article structure between the Wikipedias.

French sentence	German sentence	BLEU Score
sur ce point , Andrée se démarque non seulement des explorateurs qui lui succéderont , mais aussi de bien de ceux qui l'ont précédé	darin unterschied sich Andrée nicht nur von den späteren sondern auch von vielen früheren Entdeckungsreisenden	0.5555
lors d' une conférence donnée en 1895 à l' académie royale des sciences de Suède, il fit grosse impression devant un public composé de géographes et météorologues	er hielt Vorlesungen bei der Königlichen Akademie der Wissenschaften und bei der schwedischen Gesellschaft für Anthropologie und Geologie und erhielt breite Zustimmung	0.6010
cinquante-sept personnes trouvèrent la mort et 200 habitations, 47 ponts, 24 km de chemin de fer et 300 km de routes furent détruits	in dem dünn besiedelten und zuvor evakuierten Gebiet verloren 57 Menschen ihr Leben und 200 Häuser, 47 Brücken, 24 km Eisenbahngleise sowie 300 km Highways wurden zerstört	0.4143
il est ainsi le premier homme à gravir trois sommets de plus de 8000 m en une même saison	mit dieser Besteigung war Messner der erste Mensch überhaupt , der mehr als zwei Achttausender bestiegen hatte	1.0
cette montagne est avec le plateau de Gottesack voisin l'attraction majeure du sous-groupe	dieser Berg ist zusammen mit dem benachbarten Gottesackerplateau auch die markanteste Erscheinung der Untergruppe	1.0

Table 5: Alignment pairs identified by Bleualign

Specifically, the dataset consists of 555 000 German and 290 000 French sentences. Bleualign identified 24 000 alignments out of them. For the evaluation, we manually check a set of 200 automatically aligned sentences and we report the precision of the algorithm for this dataset.

3.2. Results

Out of the 200 sentence pairs under consideration, 30% represent perfect translations, 45% contain only aligned segments (partial alignments) and 25% represent missalignments. We can therefore count on 75% precision of the alignment procedure. A large-scale automatic evaluation of the alignment quality could be indirectly performed by measuring the improvements of a SMT system trained with the aligned data.

Table 5 presents a selection of the alignment pairs identified by our approach, together with the BLEU score computed over the translation. An interesting finding is that a high BLEU score does not always correlate with a perfect translation. BLEU has been previously criticized as a measure of translation quality, and it is not considered reliable on sentence level (Callison-Burch et al., 2006). Take, for example, the fourth sentence in the table, whose automatic translation received the maximum alignment score. This is a perfect example of a comparable text, but not a translation. The topic is clearly the same: the first man ascending more than two (e. g. three) peaks, but the rest of the sentence modifies its meaning in different directions. This finding brings again in discussion the relativity of the BLEU scores and the central question whether this sort of alignments can be considered good training material for SMT.

On the other hand, the last sentence pair correctly receives the maximum BLEU score, as all the words in the French sentence have a correspondent in the German one. In fact, the French article that contains the sentence in question is a faithful translation of its German correspondent, performed by a human translator. This is not an unique case in Wikipedia, but part of the initiative *Projet:Traduction* aiming to enrich the French Wikipedia with translations from other Wikipedias. This information is marked up in the page source with *{{Traduction/Référence|de|Allgäuer Alpen|28915176|9 mars 2007}}*. For quality reasons, the translated articles are subject to double reviewing. These sentence pairs are therefore the ideal parallel data that we aim to find in Wikipedia.

Moreover, the first two sentence pairs also represent valid translations, but receive lower alignment scores due to the different construction types. For example, the French relative clause *des explorateurs qui lui succéderont* is replaced by a nominal phrase in German: *den späteren Entdeckungsreisenden*. And the passive voice *une conférence donnée* is expressed as active voice in the German sentence: *er hielt Vorlesungen*. However, as long as these results are in the upper part of the ranking, the differences between BLEU scores should not be a problem for our task.

Between the extremes we find example number three, situated in the second half of the BLEU ranking. In this case, the German sentence has one significant extra segment compared to the French one: the nominal phase *in dem dünn besiedelten und zuvor evakuierten Gebiet*. The rest of the sentence is perfectly translated into French, but the rather poor BLEU score can be ascribed to be a penalty for the length difference. This finding highlights the need of more fine-grained alignments, at sub-sentential level. Munteanu and Marcu (2006) proposed a method to extract these segments and demonstrated the relevance of the task by reporting improvements in SMT performance.

3.2.1. SMT Experiments

In addition to the manual evaluation discussed in the previous subsection, we have run preliminary investigations with regard to the usefulness of the extracted corpus for SMT. The results discussed in this section refer only to the trans-

lation direction German-French. Our Baseline MT system is trained on the Text+Berg corpus (approx. 200 000 sentence pairs) and is the same used for the automatic translations required in the extraction step (see section 2.3.). We then train another MT system on the initial corpus plus 10 000 sentence pairs from the extracted corpus. In the following we will refer the latter one as ExtractedPlus. Both systems were tested on a test corpus of 1 000 sentences from the Text+Berg corpus. The translation performance was measured using the automatic BLEU evaluation metric on a single reference translation.

The system trained with the addition of the comparable texts has not achieved the expected improvements in performance, most probably because of the small amount of new training material (compared to the existing training data). Therefore we have manually inspected the performance of the two systems in terms of word coverage. The Baseline system failed to translate 700 words from the test corpus, whereas the ExtractedPlus system reports only 600 out-of-vocabulary words, most of them proper nouns and compounds.

An example is presented below. Both systems produce an imperfect output, following the same grammatical structure. The differences consist mainly in the choice of words. The baseline system leaves untranslated 3 words: *spitzenrouten, anziehungspunkte, bschüttigütti* and omits some words (e. g. *kletterer*). The ExtractedPlus system, however, can handle the domain specific terms like the ones mentioned before and translates them correctly: *voies extrêmes, points d' impact, grimpeurs*. Although it still cannot translate the proper noun *bschüttigütti*, the output sentence is still easier to understand and therefore the ExtractedPlus system can be considered better in this case.

DE: dasselbe gilt für die von den rein klettertechnischen schwierigkeiten her gesehenen spitzenrouten und anziehungspunkte für leistungsstarke kletterer bschüttigütti (10) und fusion (10 -).

Reference: cela vaut également pour bschüttigütti (10) et fusion (10 -) , voies extrêmes par leurs difficultés techniques , et objectifs de rêve pour de forts grimpeurs .

Baseline: il en est de même pour les difficultés purement techniques venant spitzenrouten anziehungspunkte par et pour doués bschüttigütti (10) et à s'être illustrée dans fusion (10 -) .

ExtractedPlus: il en est de même pour les difficultés purement techniques venant de la voies extrêmes et de points d' impact pour grimpeurs doués bschüttigütti (10) et de fusion (10 -) .

4. Conclusions and Outlook

We have presented our efforts in extracting a parallel corpus of Alpine texts from Wikipedia. Wikipedia, and, in general, comparable corpora are inherently heterogeneous collections of texts, where the same topic can be expanded in different ways. The differences can be found not only on the content level, but also on the formal level (i. e. MediaWiki syntax). One major problem of freely available resources like Wikipedia is that they can be edited independently by non-experts and there are no unification efforts. This makes it difficult, in the first place, to process the different Wikipedias in an uniform manner.

We demonstrated that it is difficult to use the Wikipedia categorization for the extraction of domain-specific articles from Wikipedia. Our method proposes a IR approach in order to achieve a solution to this task. However, an interesting research direction for the future is to combine these two approaches, in order to increase the reliability of the extraction method.

We have identified semantically equivalent sentences from the German and French Wikipedia articles by computing alignments between them. The reported results support our claim that this approach is worth pursuing. The procedure can be refined by training a classifier based on the Bleualign algorithm to automatically distinguish between useful and less useful alignment pairs (without the need to manually set thresholds). Moreover, as shown in section 3.2., an important improvement step is to allow the alignment of subsentential segments.

After collecting a sizable collection of Alpine texts, we will investigate the contribution of the extracted corpus for SMT performance on a larger scale. Finally, the use of the improved SMT system in our extraction algorithm could allow us to compute new and better alignments in the next development cycle.

5. References

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25:341–375. 10.1007/s10590-011-9114-9.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *In Proceedings of EACL*, pages 249–256.

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proceedings of EMNLP*.

Pascale Fung, Emmanuel Prochasson, and Simon Shi. 2010. Trillions of comparable documents. In *Proceedings of the the 3rd workshop on Building and Using Comparable Corpora (BUCC'10)*, Malta.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1301–1306. AAAI Press.

Isaac Gonzalez Lopez and Pablo Gamallo Otero. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the LREC 2010*, Malta.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploit-

ing non-parallel corpora. *Computational Linguistics*, 31:477–504, December.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–748, Washington, DC, USA. IEEE Computer Society.

Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness

Sanja Štajner and Ruslan Mitkov

Research Group in Computational Linguistics, RIILP
University of Wolverhampton, UK
S.Stajner@wlv.ac.uk, R.Mitkov@wlv.ac.uk

Abstract

This study from the area of language variation and change is based on exploitation of the comparable diachronic and synchronic corpora of 20th century British and American English language (the 'Brown family' of corpora). We investigate recent changes of lexical density and lexical richness in two consecutive thirty-year time gaps in British English (1931–1961 and 1961–1991) and in 1961–1992 in American English. Furthermore, we compare the diachronic changes between these two language varieties and discuss the results of the synchronic comparison of these two features between British and American parts of the corpora (in 1961 and in 1991/2). Additionally, we explore the possibilities of these comparable corpora by using two different approaches to their exploitation: using the fifteen fine-grained text genres, and using only the four main text categories. Finally, we discuss the impact of the chosen approaches in making hypotheses about the way language changes.

Keywords: corpus analysis, language change, lexical richness

1. Introduction

Kroch (2008) defines language change as "a failure in the transmission across time of linguistic features" and states that "over historical time languages change at every level of the language structure: vocabulary, phonology, morphology and syntax". He states that in principle, language change can occur within groups of adult native speakers of language as the result of the substitution of one feature with another as in the case of the substitution of old words with new ones, though he raises a doubt in the validity of this hypothesis in the case of syntactic and grammatical changes.

1.1. Lexical density and lexical richness

In this study, our focus was only at the vocabulary level of the language change. We wanted to investigate how the lexical density and lexical richness were changing during the 20th century. Lexical density is one of the most commonly used features for describing diversity of a vocabulary (Stamatatos et al., 2000). Smith and Kelly (2002), for instance, used this feature for dating works. Lexical density is calculated as the ratio between the number of unique word types and the total number of tokens in the given text. Therefore, a higher lexical density would indicate a wider range of used vocabulary. However, as lexical density counts morphological variants of the same word as different word types (tokens), Corpas Pastor et al. (2008) suggested the use of another measure – lexical richness, instead. The lexical richness is computed as the ratio between the number of unique lemmas and the total number of tokens in the given text. This second measure does not take into account different morphological counts of the same word as different word types. Therefore, Corpas Pastor et al. (2008) believed that it would be a more appropriate indicator of the vocabulary variety of an author.

1.2. Diachronic corpora of 20th century English language

There are several corpora of English language consisting of the texts published in the 20th century, compiled principally for purposes of grammatical researches, but they are usually not publicly available or they cover only a specific genre. The ARCHER corpus (Biber et al., 1994), for instance, belongs to the first of the mentioned groups. It covers a wide range of genres - drama, medical, historical and news reportage texts, from 1650 to 1990 divided into fifty-year blocks, but is not available to the research community (Leech and Smith, 2005). The Corpus of Late Modern English Prose (Denison, 1994), a collection of informal private letters written in British English between 1861 and 1919 is, on the other hand, available to the research community, but it covers only one genre and belongs more to the 19th than to the 20th century. The Corpus of English Newspaper Editorials – CENE (Westin, 2002; Westin and Geisler, 2002), which consists of institutional editorials of three 'broadsheet' British newspapers - The Times, The Guardian and The Daily Telegraph, sampled at ten-year intervals across the 20th century (Leech and Smith, 2005) and the Bauer's corpus of The Times (Bauer, 1994), also consisting of editorials sampled at decade intervals (Leech and Smith, 2005), both belong to the intersection of the above two types as they cover only a specific genre and they are not publicly available.

1.3. The 'Brown family' of corpora

The 'Brown family' of corpora is comprised of five mutually comparable corpora. The American part consists of two corpora:

- The Brown University corpus of written American English – Brown (Francis, 1965)
- The Freiburg - Brown Corpus of American English –

Main category	Code	Genre	Number of texts		
			(F/B)LOB	Brown	Frown
PRESS	A	Press: Reportage	44	44	44
	B	Press: Editorial	27	27	27
	C	Press: Review	17	17	17
PROSE	D	Religion	17	17	17
	E	Skills, Trades and Hobbies	38	36	36
	F	Popular Lore	44	48	48
	G	Belles Lettres, Biographies, Essays	77	75	75
	H	Miscellaneous	30	35	30
LEARNED	J	Science	80	80	80
FICTION	K	General Fiction	29	29	29
	L	Mystery and Detective Fiction	24	24	24
	M	Science Fiction	6	6	6
	N	Adventure and Western	29	30	29
	P	Romance and Love Story	29	29	29
	R	Humour	9	9	9

Table 1: Structure of the corpora

Frown (Hundt et al., 1998).

The British part consists of three corpora:

- The Lancaster1931 – BLOB (Leech and Smith, 2005)
- The Lancaster-Oslo/Bergen Corpus – LOB (Johansson et al., 1978)
- The Freiburg-LOB Corpus of British English – FLOB (Sand and Siemund, 1992).

The corpora contain texts published in years 1931±3 (Lancaster1931), 1961 (LOB and Brown), 1991 (FLOB) and 1992 (Frown) divided into 15 different genres (Table 1). These five corpora comply with the formal criteria of comparability as the texts have been compiled on the basis of the same sampling frame and with similar balance and representativeness. In particular, the texts have been selected to match the same domain and topics, and are of comparable size. Therefore, they fulfill all the necessary conditions for being widely used throughout the linguistic community – they are a diachronic corpora of 20th century written English texts, which cover a wide range of genres and are publicly available as part of the ICAME Corpus Collection[1].
The Brown corpus was published first, back in 1964. One of the ideas of compiling the Brown corpus was to help "to have a common body of material on which studies of various sorts can be based" (Leech and Smith, 2005) and in that way to provide some kind of 'standard' for the following parallel corpora of British English or for English of other periods to be matched (Francis, 1965 in Leech and Smith, 2005). It was a one-million-word corpus, consisting of 500 texts of about 2000 running words each, selected at random points from the original source and the texts covered fifteen different text genres. Following that idea, the LOB corpus (Johansson at al. 1978) of written British English was compiled as the first corpus to match the Brown corpus, respecting the year of sampling (1961) and its sampling frame and representation of different text types (Leech and Smith, 2005). The release of the LOB corpus enabled synchronic comparison between two major English language varieties across a wide range of text genres. In the 1990s, the FLOB and Frown corpora were compiled at Freiburg University representing, respectively, written British English in 1991 and American English in 1992. As their design matched closely to the design of the LOB and Brown corpora, this provided the opportunity to investigate and compare diachronic changes between two major varieties of the written English language. The exact procedure for diachronic matching applied during the compilation of the FLOB and Frown corpora could be found in (Leech and Smith, 2005, p.8). Later on, the research to extend the Brown model backwards in time, undertaken at the Lancaster University, led to the compilation of the Lancaster1931 corpus to match the design of the LOB and FLOB corpora. The target sampling year in this case was 1931 (± three years), in order to maintain the thirty-year gap already established between LOB and FLOB corpora, as well as between Brown and Frown corpora.
Being all mutually comparable, these five corpora (BLOB, LOB, FLOB, Brown and Frown) create the possibility for several different types of investigation:

- Synchronic comparison between British and American English in 1961 and in 1991/2
- Diachronic comparison among the texts published in 1931, 1961 and 1991 in British English
- Diachronic comparison among the texts published in 1961 and 1992 in American English
- Comparison of diachronic changes in 1961–1991/2 between British and American English

1.4. Structure of the corpora

Each of the corpora (BLOB, LOB, FLOB, Brown and Frown) consist of approximately 1,000,000 words – 500 texts of about 2000 running words each. The texts cover fifteen different text genres (Table 1), which could be further grouped into four more generalised categories: Press

[1] http://icame.uib.no/newcd.htm

(A–C), Prose (D–H), Learned (J) and Fiction (K–R). This structure of the corpora allows three different approaches to the exploitation of the corpora in diachronic studies:

1. Differentiating between texts only across two different language varieties or two different years of publication (without differentiating between texts across the text genres/categories).

2. Differentiating between texts across the four main text categories (Press, Prose, Learned and Fiction), thus exploring diachronic changes separately in each of the four main text categories.

3. Differentiating between texts across all fifteen fine-grained text genres (A–R), thus exploring diachronic changes separately in each of the fifteen fine-grained text genres.

2. Related work

The 'Brown family' of corpora has already been used in many diachronic studies of various lexical, grammatical, stylistic and syntactic features, e.g. (Mair and Hundt, 1995; Mair, 1997; Mair et al., 2002; Smith, 2002; Smith, 2003b; Smith, 2003a; Leech, 2003; Leech, 2004; Leech and Smith, 2006; Mair and Leech, 2006; Leech and Smith, 2009; Leech et al., 2009; Štajner and Mitkov, 2011). A large set of these studies shared the same methodology. The corpora were part-of-speech tagged, the change was presented in terms of the absolute and relative differences and the statistical significance was measured by the log likelihood function. The first attempt for a completely automated feature extraction from the raw text version of the 'Brown family' of corpora in diachronic studies was reported by Štajner and Mitkov (2011). The corpora were parsed with the state-of-the-art Connexor's Machinese Syntax parser[2] and the features were automatically extracted from the parser's output. Statistical significance of the results was measured by the t-test.

However, all of these previous studies used the aforementioned second approach, differentiating only between texts across the four main categories (Press, Prose, Learned and Fiction). Following the discussion in (Stajner, 2011) about the impact of the chosen genre granularity (aforementioned approaches 1–3), we decided to use the third approach and differentiate between texts across all fifteen fine-grained text genres (A–R), in order to obtain a better understanding of how lexical density and richness change. To the best of our knowledge, this is the first diachronic study conducted on the 'Brown family' of corpora using this approach.

Of the most relevance for this work was the study conducted by Štajner and Mitkov (2011), where the authors investigated diachronic changes of lexical density (LD) and lexical richness (LR) in the period 1961–1991/2 and used the same methodology for feature extraction. However, they only differentiated between texts across the four main text categories (Press, Prose, Learned and Fiction). In this study, we went one step further, by differentiating between texts across all fifteen fine-grained text genres (A–R). This approach allowed us to obtain a better insight into the way language changes. It also gave us the opportunity to compare the results obtained by these two different approaches and draw attention to the possible pitfalls in making hypotheses by differentiating between texts only across the four main text categories. In that sense, the results presented in this study could also be taken as an additional support for the claims made in (Štajner, 2011).

In this study, we also extended the time span in British English by using the Lancaster1931 corpus. Therefore, we were able to compare the trends of change in two consecutive thirty-year time gaps (1931–1961 and 1961–1991) in British English and examine whether the trend of change was stable during the whole sixty-year period.

3. Methodology

In this study, we followed the methodology for feature extraction proposed by Štajner and Mitkov (2011). All five corpora were used in their initial raw text format and then parsed with the state-of-the-art Connexor's Machinese Syntax parser for the purposes of tokenisation and lemmatisation. The main reason for using the same parser and the same methodology, although the tokenisation and lemmatisation could have been done by some lighter tools, was to be able to compare our results obtained for all fifteen text genres (the aforementioned third approach) with those results reported by Štajner and Mitkov (2011) when the authors were differentiating only between the texts across the four main text categories (the aforementioned second approach). As the performance of the parser in this task and its specificities regarding the tokenisation and lemmatisation processes were already discussed in details in (Štajner and Mitkov, 2011), here we will just highlight the most important ones in order to facilitate a better understanding of the presented results.

The lexicon of the Connexor's Machinese Syntax parser was built using various large corpora of different text genres – news, bureaucratic documents, literature etc. (Connexor, 2006) and contains hundreds of thousands of base forms. The words which are not found in the lexicon are assigned their word class and base form by using the heuristic methods (Connexor, 2006). The software which was used as a base for the current version of the parser reported an excellent accuracy (Samuelsson and Voutilainen, 1998) and the parser itself reported the POS accuracy of 99.3% on Standard Written English (benchmark from the Maastricht Treaty) with no ambiguity (Connexor, 2006).

3.1. Tokenisation

The Connexor's Machinese parser treats the contracted negative form (*n't*) and its antecedent verb as two separate tokens. E.g. *aren't* would be separated into two tokens *are* and *not* and assigned two separate base forms – *be* and *not*. The *'s* is treated in two different ways, depending on the role it has in the sentence. When it represents a genitive form, e.g. *"... Isaac's illness..."* (FLOB: K02), it is treated as one token and is assigned the corresponding lemma *isaac*. In other cases where *'s* represents the contraction of the verb *to be (is)* or *to have (has)*, e.g. *" He's at a table over there."* (FLOB: K01), the personal pronoun

[2] http://www.connexor.eu

and verb contraction are treated as two separate tokens *he* and *is* and assigned two separate base forms *he* and *be*, accordingly.

3.2. Lemmatisation

The output of the lemmatisation process done by the Connexor's Machinese parser expresses certain differences between the earlier versions and the current version of the parser. The main difference is in the way that possessive pronouns, derived adverbs, and EN and ING forms are treated.

While the earliest versions of the parser would assign the corresponding personal pronoun as the lemma of the given possessive pronoun (e.g. the word *theirs* would be assigned *their* as its lemma), the current version of the parser assigns their own base forms to possessive pronouns (e.g. the word *theirs* is assigned *theirs* as its lemma).

A similar rule applies to derived adverbs. In the previous versions of the parser, derived adverbs, such as *absolutely* or *directly* would be assigned *absolute* and *direct* as their lemmas, while in the current version of the parser, these same derived adverbs are assigned their own base forms – *absolutely* and *directly*.

The EN and ING forms, which can represent either present and past participles or corresponding nouns and adjectives, are assigned a POS tag (EN, ING, N or A) and different base forms in the current version of the parser, according to their usage in that particular case. For example, if the word *meeting* is recognised as a noun by the parser, it will be assigned *meeting* as the corresponding lemma. In case that the same word is recognised as a present participle of the verb *to meet*, it will be assigned *meet* as its corresponding lemma. The results would be similar in the case of an EN form. For example, if the word *selected* represents an adjective in the given context, it will be assigned *selected* as its lemma. In another case, if it represents a past participle, it will be assigned *select* as the corresponding lemma.

These differences between previous and current versions of the parser in lemmatising certain word forms is reflected in the differences between the lexical richness and lexical density. It is reasonable to expect that the calculated LD and LR will be much closer if we use the current version than if we use an earlier version of the parser.

3.3. Feature extraction

The lexical density (LD) and lexical richness (LR) were calculated for each text separately in order to enable later applied statistical tests. Lexical density was calculated as the total number of unique word forms (tokens) divided by the total number of tokens in the given text (eq.1).

$$LD = \frac{number\_of\_unique\_tokens}{total\_number\_of\_tokens} \quad (1)$$

Lexical richness was calculated similarly, this time using the total number of unique lemmas divided by the total number of tokens (eq.2).

$$LR = \frac{number\_of\_unique\_lemmas}{total\_number\_of\_tokens} \quad (2)$$

4. Experimental settings

The purpose of this study was two-fold: (1) to investigate diachronic changes of lexical density and lexical richness in 20th century English language in each of the fifteen fine-grained text genres, and (2) to compare the results of two different approaches to the exploitation of these comparable corpora. Therefore, we had two different sets of experiments. The first set of experiments consisted of investigating the following five changes using the third approach (differentiating between the texts across the all fifteen fine-grained text genres):

- Diachronic changes in British English in the period 1931–1961
- Diachronic changes in British English in the period 1961–1991
- Diachronic changes in American English in the period 1961–1992
- Synchronic differences between British and American English in 1961
- Synchronic differences between British and American English in 1991/2.

The second set of experiments consisted of the same five experiments, but this time using the second approach (differentiating between the texts only across the four main text categories).

4.1. Statistical significance testing

For each of the aforementioned five experiments we calculated the statistical significance of the mean differences between the two corresponding groups of texts. Statistical significance tests are divided into two main groups: parametric (which assume that the samples are normally distributed) and non-parametric (which do not make any assumptions about the sample distribution). In the cases where the samples follow the normal distribution, it is recommended to use parametric tests as they have greater power than non-parametric tests (Garson, 2012a). Therefore, we first applied the the Shapiro-Wilk's W test (Garson, 2012b) offered by SPSS EXAMINE module in order to examine in which cases/genres/categories the features were normally distributed. This test is a standard test for normality, recommended for small samples. It shows the correlation between the given data and their expected normal scores. If the result of the W test is 1, it means that the distribution of the data is perfectly normal. Significantly lower values of W (≤ 0.05) indicate that the assumption of normality is not met. Those cases are shown in bold (Table 2).

Following the discussion in (Garson, 2012c), for both approaches we used the following strategy: if the two data sets we wanted to compare were both normally distributed we used the t-test for the comparison of their means; if at least one of the two data sets was not normally distributed ($W \leq 0.05$ in Table 2), we used the Kolmogorov-Smirnov Z test (a non-parametric test) for two independent samples to calculate the statistical significance of the differences between their means.

Approach	Genre	LD					LR				
		British			American		British			American	
		1931	1961	1991	1961	1992	1931	1961	1991	1961	1992
III	A	.320	**.003**	.807	.448	.737	.221	**.015**	.963	.345	.575
	B	.935	.905	.326	.263	.958	.776	.644	.322	.256	.371
	C	.399	.716	.428	**.002**	.369	.370	.786	.692	**.002**	.574
	D	.777	.679	.643	.711	.089	.706	.409	.178	.816	**.047**
	E	.115	**.026**	**.011**	.238	.725	.289	**.047**	.093	.664	.353
	F	.818	.639	.319	.338	**.000**	.883	.652	.401	.383	**.000**
	G	**.013**	.065	.170	.054	.072	**.017**	**.018**	.285	.236	.240
	H	.202	.892	.952	.119	.303	.261	.992	.970	.109	.266
	J	.051	.883	.252	**.002**	.470	.127	.803	.158	**.003**	.826
	K	.403	.835	.511	.283	.523	.304	.722	.916	.353	.630
	L	.333	.599	.291	.230	.529	.365	.457	.359	.141	.277
	M	.528	.290	.940	.179	.812	.601	.55	.792	.107	.835
	N	.966	.127	.287	.990	.314	.886	.087	.183	.789	.572
	P	.587	.084	.322	.279	.362	.300	.068	.316	.379	.386
	R	.291	.913	.580	.555	.962	.182	.873	.683	.421	.805
II	Press	.834	.068	**.012**	.112	.490	.856	.230	**.014**	**.044**	.660
	Prose	**.000**	**.001**	**.000**	**.000**	**.000**	**.002**	**.002**	**.011**	**.001**	**.000**
	Learned	.051	.883	.252	**.002**	.470	.127	.803	.158	**.003**	.826
	Fiction	.756	.116	.850	.087	.169	.591	.101	.645	**.011**	.181

Table 2: Normal distribution testing (Shapiro-Wilk's W test results)

It is interesting to note that in some cases, even if the data in fine-grained text genres follow the normal distribution (e.g. genres A–C in columns LD and LR of British English in 1991), the data in that whole text category (Press in columns LD and LR of British English in 1991) do not follow the same distribution. Also, we can find examples of the opposite situation when some of the data in the fine-grained text genres (e.g. genre A in columns LD and LR of British English in 1961) do not follow the normal distribution, but the data in the corresponding broader text category (Press in columns LD and LR of British English in 1961) are normally distributed. This second case is intuitively more expected as we know that the bigger the data set, the more chance there is that the data would be normally distributed. However, both the cases force us to use different statistical significance tests for the second and for the third approach.

5. Results and discussion

Our study basically has two main parts: diachronic comparison (1931–1961 and 1961–1991 in British English; 1961–1992 in American English) and synchronic comparison of British and American English (in 1961 and in 1991/2). Therefore, we will present the results separately for diachronic (separately for LD and LR) and synchronic comparisons (together for LD and LR) in the next three subsections. In each of these subsections, together with our main results obtained by using the third approach (differentiating across fifteen fine-grained text genres) we will also present the results of the alternative second approach (differentiating across only four main text categories), in order to be able to compare the differences in the conclusions drawn from these two approaches. Statistically significant changes at a 0.05 level significance (sign. ≤ 0.05) are shown in bold.

5.1. Diachronic changes of lexical density (LD)

The results of the investigation of diachronic changes of lexical density (LD) in British and American English are presented in Table 3 (using the third approach) and Table 4 (using the second approach). In both cases we followed the same pattern of representing the results. Columns '1931', '1961' and '1991' under 'British English', and columns '1961' and '1992' under 'American English' represent the calculated average LD in those years for the corresponding language variety. Columns '1931–1961', '1961–1991' and '1961–1992' contain the information about the changes of LD in those periods for the corresponding language varieties. Their subcolumn 'sign.' represent the calculated two-tailed statistical significance of the differences between the corresponding means, by using t-test or Kolmogorov-Smirnov Z test, according to Table 2 and the discussion in Subsection 4.1. The subcolumn 'change' contains the relative change in the observed period, calculated as a percentage of the starting value. The sign '+' stands for an increase and the sign '−' for a decrease over the time.

5.1.1. British English

The results presented in Table 3 indicate several interesting phenomena. First, we can notice that diachronic changes in British English were generally not stable in the two subsequent periods 1931–1961 and 1961–1991. Most of the genres demonstrated significant changes only in one of the two observed periods. In genres G (Belles Lettres, Biographies, Essays) and R (Humour), LD had changed (increased) only in the first period 1931–1961, while in genres A (Press: Reportage), B (Press: Editorial), C (Press: Review), D (Religion) and P (Romance and Love Story) it had changed (increased) only in the second period 1961–1991. Genre E (Skills, Trades and Hobbies) was the only genre that showed a stable increase of LD throughout both periods

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	0.352	0.316	+0.64%	**0.355**	**0.000**	**+6.90%**	**0.379**	0.369	0.940	+0.13%	0.368
B	0.354	0.202	+1.95%	**0.361**	**0.000**	**+7.94%**	**0.389**	**0.378**	**0.031**	**+3.64%**	**0.392**
C	0.382	0.158	+2.83%	**0.392**	**0.001**	**+7.95%**	**0.424**	**0.395**	**0.006**	**+4.18%**	**0.411**
D	0.312	0.427	−2.79%	**0.304**	**0.027**	**+8.47%**	**0.329**	0.323	0.381	+3.26%	0.334
E	**0.327**	**0.045**	**+4.66%**	**0.342**	**0.002**	**+6.99%**	**0.366**	**0.331**	**0.014**	**+7.72%**	**0.357**
F	0.342	0.916	+0.23%	0.342	0.421	+1.91%	0.349	**0.342**	**0.027**	**+5.84%**	**0.362**
G	**0.341**	**0.047**	**+2.79%**	0.350	0.065	+2.66%	0.359	0.347	0.279	+1.42%	0.351
H	0.286	0.593	+1.79%	0.292	0.792	+1.01%	0.295	0.294	0.688	+1.74%	0.299
J	0.295	0.600	+1.34%	0.299	0.236	+2.84%	0.307	0.298	0.329	+4.69%	0.312
K	0.315	0.295	−2.81%	0.307	0.118	+4.51%	0.320	0.327	0.370	−2.99%	0.317
L	0.299	0.458	+1.93%	0.304	0.434	−2.31%	0.297	0.299	0.493	+2.17%	0.306
M	0.328	0.810	+1.46%	0.333	0.574	+4.48%	0.348	0.323	0.779	−1.55%	0.318
N	**0.314**	**0.048**	**−4.90%**	**0.299**	**0.020**	**+7.06%**	**0.320**	0.315	0.768	−0.92%	0.313
P	0.298	0.089	−4.46%	**0.285**	**0.010**	**+7.66%**	**0.307**	0.302	0.528	−1.86%	0.297
R	**0.311**	**0.000**	**+14.16%**	**0.355**	0.545	−1.96%	0.348	**0.359**	**0.011**	**−18.39%**	**0.293**

Table 3: Diachronic changes of lexical density (LD) – third approach

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
Press	0.358	0.168	+1.48%	**0.364**	**0.000**	**+7.43%**	**0.391**	0.376	0.084	+2.03%	0.384
Prose	**0.328**	**0.007**	**+2.02%**	**0.335**	**0.000**	**+3.54%**	**0.347**	**0.333**	**0.007**	**+3.76%**	**0.346**
Learned	0.295	0.600	+1.34%	0.299	0.236	+2.84%	0.307	0.298	0.329	+4.69%	0.312
Fiction	0.308	0.297	−1.35%	**0.304**	**0.009**	**+3.92%**	**0.316**	0.315	0.105	−2.51%	0.307

Table 4: Diachronic changes of lexical density (LD) – second approach

1931–1961 and 1961–1991. The most interesting might be the case of genre N (Adventure and Western) which demonstrated a significant change of LD in both periods although these changes had opposite directions. While in the first period (1931–1961) LD had decreased, in the second period (1961–1991) it had increased. At the same time, the decrease of LD in this genre is the only observed significant decrease of LD in British English in this study.

5.1.2. American English

In American English, the results (Table 3) indicated a significant increase of LD in four genres: B (Press: Editorial), C (Press: Review), E (Skills, Trades and Hobbies) and F (Popular Lore), and a significant decrease of LD in genre R (Humour). At the same time, this change of LD in genre R was of a significantly higher intensity than the changes reported in other genres.

5.1.3. British vs. American English

The comparison of diachronic changes of LD between British and American English in the period 1961–1991/2 indicates that the most of the genres did not undergo the same changes at the same time. For instance, genres A (Press: Reportage), D (Religion), N (Adventure and Western) and P (Romance and Love Story) demonstrated a change only in British English, while genres F (Popular Lore) and R (Humour) demonstrated a change of LD only in American English during the same period 1961–1991/2. The only genres which reported a significant increase of LD in both language varieties during that period were genres B (Press: Editorial), C (Press: Review) and E (Skills, Trades and Hobbies).

5.1.4. Second vs. third approach

The first obvious difference in conclusions drawn from the results of the second approach (Table 4) and those of the third approach (Table 3) is that by using solely the results of the second approach we would conclude that whenever there was a change, LD has increased. By closer examination of the corpora (Table 3), we notice that in fact a significant decrease of LD is also likely to happen, as in the case of genre N (Adventure and Western) in British English (1931–1961) and genre R (Humour) in American English (1961–1992).

The other differences between the conclusions drawn from these two approaches are more subtle but maybe even more important to mention. The most drastic difference can be noticed in Fiction category of British English (1931–1961), and Fiction and Press categories in American English (1961–1992). While the results of the second approach (Table 4) reported no changes of LD in these particular cases, the results of the third approach (Table 3) revealed some interesting phenomena in the corresponding genres. In American English, a very intensive decrease of LD in genre R (present in the results of the third approach), was probably masked in the second approach by the constancy of LD in other genres of this category (genres K–P) which have a greater number of texts than genre R (Table1

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	0.317	0.808	+0.09%	**0.317**	**0.001**	**+5.95%**	**0.336**	0.331	0.334	−1.84%	0.325
B	0.316	0.263	+1.81%	**0.321**	**0.000**	**+8.27%**	**0.348**	0.337	0.117	+2.93%	0.347
C	0.345	0.141	+3.21%	**0.356**	**0.002**	**+8.42%**	**0.386**	**0.358**	**0.017**	**+3.49%**	**0.371**
D	0.278	0.362	−3.59%	**0.268**	**0.030**	**+9.40%**	**0.293**	0.286	0.240	+3.92%	0.297
E	**0.290**	**0.012**	**+4.74%**	**0.303**	**0.005**	**+6.99%**	**0.324**	**0.292**	**0.024**	**+8.00%**	**0.316**
F	0.303	0.993	+0.02%	0.303	0.389	+2.36%	0.310	0.304	0.249	+5.55%	0.320
G	**0.304**	**0.018**	**+3.21%**	**0.313**	**0.004**	**+3.35%**	**0.324**	0.310	0.550	+0.89%	0.312
H	0.254	0.649	+1.75%	0.258	0.840	+0.78%	0.260	0.261	0.772	+1.28%	0.265
J	0.262	0.550	+1.61%	0.267	0.413	+2.11%	0.272	0.265	0.436	+4.87%	0.278
K	0.277	0.246	−3.54%	0.268	0.168	+4.56%	0.280	0.287	0.411	−3.17%	0.278
L	0.261	0.521	+1.88%	0.265	0.427	−2.63%	0.259	0.260	0.490	+2.49%	0.267
M	0.290	0.879	+1.06%	0.293	0.562	+5.40%	0.309	0.285	0.699	−2.47%	0.277
N	**0.276**	**0.030**	**−6.18%**	**0.259**	**0.020**	**+8.26%**	**0.281**	0.275	0.826	−0.79%	0.273
P	0.259	0.066	−5.44%	**0.245**	**0.014**	**+8.24%**	**0.266**	0.264	0.359	−3.05%	0.256
R	**0.271**	**0.000**	**+16.78%**	**0.317**	0.555	−2.06%	0.310	**0.320**	**0.012**	**−21.14%**	**0.253**

Table 5: Diachronic changes of lexical density (LR) – third approach

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
Press	0.322	0.279	+1.25%	**0.326**	**0.000**	**+7.17%**	**0.349**	0.338	0.387	+0.71%	0.341
Prose	**0.291**	**0.001**	**+2.06%**	**0.297**	**0.000**	**+3.95%**	**0.309**	0.296	0.096	+3.52%	0.307
Learned	0.262	0.550	+1.61%	0.267	0.413	+2.11%	0.272	0.265	0.436	+4.87%	0.278
Fiction	0.270	0.201	−1.89%	**0.265**	**0.012**	**+4.28%**	**0.276**	0.276	0.202	−3.04%	0.268

Table 6: Diachronic changes of lexical density (LR) – second approach

in Section 1.4). The differences in the Prose category of American English could be explained in the same way. In British English, however, the situation was even more complex. The results of the second approach did not only mask the changes of LD in certain genres (N and R), but they also hid the fact that the changes in these two genres went in opposite directions (an increase of LD in genre R and a decrease of LD in genre N).

Less pronounced, but still worth mentioning, were the differences between the results of the second and third approaches in Prose (1931–1961, 1961–1991) and Fiction (1961–1991) categories of British English, and Prose category of American English. In these cases, the results of the second approach reported significant changes of LD in these categories (Table 4), while the more detailed analysis used in the third approach (Table 3) actually demonstrated that these changes were present only in certain genres of the mentioned categories and not in all of them.

5.2. Diachronic changes of lexical richness (LR)

Diachronic changes of lexical richness (LR) in British and American English are presented in the same manner as in the case of lexical density. Table 5 contains the results of the third approach and Table 6 the results of the second approach.

5.2.1. British English

Similar to the case of LD, LR did not show the same trends of changes in both observed periods 1931–1961 and 1961–1991 in most of the genres. In genre R (Humour) a change was present only in the first period (1931–1961), while in genres A (Press: Reportage), B (Press: Editorial), C (Press: Review), D (Religion) and P (Romance and Love Story) a change was present only in the second period (1961–1991). In genres E (Skills, Trades and Hobbies) and G (Belles Lettres, Biographies, Essays), LR had increased in both periods, while in genre N (Adventure and Western) it first had decreased (in period 1931–1961) and then increased (in the period 1961–1991).

If we compare these results for LR with those obtained for LD (Table 3), we can notice that in most genres, LD and LR demonstrated similar diachronic changes. The only exception to this was reported in genre G (Belles Lettres, Biographies, Essays) in the period 1961–1991 , where LD did not show any statistically significant change, while LR reported an increase of +3.35%.

5.2.2. American English

The results of the investigation of diachronic changes of LR in American English (Table 5) reported a higher lexical richness in 1992 than in 1961 in genres C (Press: Review) and E (Skills, Trades and Hobbies). In genre R (Humour) the situation was the opposite. In this genre, LR was reported to be higher in 1961 than in 1992.

The comparison of diachronic changes between LD and LR (Table 3 and Table 5) indicate similar behaviour of these two features in all three genres in which a significant change of LR was reported. Additionally, LD demon-

Year	Genre	LD				LR			
		Br.	sign.	change	Am.	Br.	sign.	change	Am.
1961	A	0.355	0.043	+3.92%	0.369	0.317	0.012	+4.39%	0.331
	B	0.361	0.012	+4.79%	0.378	0.321	0.020	+4.91%	0.337
	K	0.307	0.035	+6.66%	0.327	0.268	0.041	+7.41%	0.287
	P	0.285	0.037	+6.04%	0.302	0.245	0.023	+7.45%	0.264
1991/2	G					0.324	0.031	−3.53%	0.312
	R	0.348	0.004	−15.91%	0.293	0.310	0.002	−18.63%	0.253

Table 7: Synchronic comparison of LD and LR in 1961 and 1991/2 (British vs. American English)

strated a change in genres B (Press: Editorial) and F (Popular Lore), in which LR did not report any changes.

5.2.3. British vs. American English

The results of the comparison of diachronic changes of LR between British and American English indicates that this feature underwent similar changes in both language varieties (in the period 1961–1991/2) in only two genres – C (Press: Review) and E (Skills, Trades and Hobbies). The number of genres in which LR reported a change in only one of the two language varieties was significantly higher, thus indicating different trends of change between these two varieties in general. On one side we have genres A (Press: Reportage), B (Press: Editorial), D (Religion), G (Belles Lettres, Biographies, Essays), N (Adventure and Western) and P (Romance and Love Story) for which the results (Table 5) indicate a significant increase of LR in the period 1961–1991 only in the British part of the corpora. On the other side we have genre R (Humour) in which a significant change (in this case a decrease) of LR was reported only in American English.

5.2.4. Second vs. third approach

The investigation of diachronic changes of LR revealed the same possible pitfalls in making conclusions solely based on the results of the second approach (Table 6) as in the case of LD (Section 5.1.4). For example, these results (Table 6) did not show any significant differences of LR between 1931 and 1961 in Fiction category, while the results of the third approach (Table 5) indicated a significant decrease of LR in genre N (Adventure and Western) and a significant increase in genre R (Humour). In this case not only did the results of the second approach fail to report significant changes in some genres of the Fiction category, but even more importantly, they failed to report that different genres which belong to the same broad category, exhibit different trends of change – increase and decrease, in the same period of time.

In American English, the results of the second approach (Table 6) did not indicate any changes of LR in the observed period 1961–1992, while the results of the third approach (Table 5) reported significant changes in one of the genres in each of the Press, Prose and Fiction categories – genres C (Press: Review), E (Skills, Trades and Hobbies) and R (Humour). In the case of Prose category (in both periods, 1931–1961 and 1961–1991) and Fiction category (in the period 1961–1991) in British English, the results of the second approach (Table 6) which reported a significant increase of LR were less misleading than in the previous case, though still hiding the fact that these changes were present only in certain genres of this category and not in all of them (Table 5).

5.3. Synchronic comparison

The results of synchronic comparison of LD and LR between British and American English are presented in Table 7. As LD and LR were already presented for both of these language varieties in the previous two sections (5.1 and 5.2), here we presented only the genres in which a statistically significant difference between British and American English was reported for at least one feature and one year.

It is interesting to note that the results (Table 7) did not report any genre in which a significant difference of LD or LR between these two language varieties was present in both years – 1961 and 1991/2. Actually, in 1961, a significant difference in LD and LR between British and American was reported in only four genres – A (Press: Reportage), B (Press: Editorial), K (General Fiction) and P (Romance and Love Story). In all these genres, the texts written in American English used a wider vocabulary than those written in British English. In 1991/2, a significant difference of LD between British and American English was reported in only one genre – genre R (Humour). In this genre, texts written in British English had a greater vocabulary variety than those written in American English. In the same year (1991/2), LR was reported to be significantly higher in British than in American English for two genres – genre G (Belles Lettres, Biographies, Essays) and R (Humour).

It is also interesting to notice that all reported differences in 1961 went in favour of a larger vocabulary used in American English, while all those differences reported in 1991/2 went in favour of a larger vocabulary used in British English.

6. Conclusions

The results of the experiments presented in this paper enabled us to make two different types of relevant conclusions. The first type of conclusions would be those regarding the investigated diachronic changes of lexical density and lexical richness and their behaviour in British and American English. The second type would be those regarding the influence of the chosen approach (chosen way of exploitation of the comparable corpora) – using only four main broad text categories (second approach) or using all fifteen fine-grained text genres (third approach), on making hypotheses about the way English language changes.

On the basis of the results of the third approach to the investigation of diachronic changes of LD and LR (Tables 3 and 5), we can conclude that the changes of these two stylistic features were very heterogeneous in various ways – across the genres (A–R), language varieties (British and American) and periods observed (1931–1961, 1961–1991/2). Most importantly, these results indicated different trends of change even among the genres which belong to the same broad text category, e.g. genres N and P in Fiction category reported a decrease and an increase of LD and LR in the same period 1931–1961. Furthermore, the investigated genres did not report many constant ongoing changes during the two consecutive periods 1931–1961 and 1961–1991. Genre N (Adventure and Western) reported a significant decrease in the first period 1931–1961 and then a significant increase of both features (LD and LR) in the second period (1961–1991) in British English. In other genres, a significant change was usually reported in only one of the two observed periods. The only exceptions were noticed in genre E (Skills, Trades and Hobbies), where LD and LR had increased in both periods, and in genre G (Belles Lettres, Biographies, Essays), where LR reported a significant increase during both periods.

Genre R (Humour) reported different behaviour between the two language varieties (no change in British English and a significant decrease in American English) for the same period 1961–1991/2, and different behaviour in two consecutive time periods in British English (an increase in 1931–1961 and no significant change in 1961–1991). Even more interestingly, the reported changes in British and American English (although not for the same period, but for 1931–1961 in British and for 1961–1992 in American English) did not follow the same direction, i.e. in British English, LD and LR had increased (in the period 1931–1961), while in American English, both of these features had decreased (in the period 1961–1992). Therefore, we cannot even say that the changes reported in British and American English were shifted in time (for thirty years). The results presented in this study actually indicate that the changes of LD and LR in British and American English were not mutually influenced.

All these findings lead to the conclusion that the time gap in diachronic studies of lexical density and lexical richness should ideally be smaller if we wish to gain a better insight into the way they change. They also indicate that different language varieties should be investigated separately as they generally do not follow the same patterns of change. Similarly, the presented results emphasise the necessity for separate investigation of the genres which belong to the same broad text category as they demonstrate different trends of changes among themselves.

The comparison between the results obtained by using the second approach (differentiating only across the four main broad categories) and those obtained by using the third approach (differentiating across all fifteen fine-grained text genres) clearly stated some of the potential pitfalls in making hypotheses about the way language changes solely on the basis of the results of the second approach. It pointed out two possible problems in using the second approach. The first problem would be the case in which the results of the second approach do not report any changes in the relevant text category, while a closer examination of the same category (using the third approach) clearly indicates significant changes in some of the genres belonging to that category. The second problem would be the case in which the results of the second approach again do not report any changes, while the results of the third approach not only indicate significant changes in some of the genres of that category, but also indicate different trends of changes among them (increase, decrease and no change). In the second approach these changes are probably masked by unbalanced distribution of texts or by a high heterogeneity of changes across different genres of that category.

Finally, this study presented various possibilities of the comparable 'Brown family' of corpora and different approaches to their exploitation in diachronic and synchronic language studies. Most of these ideas and the methodology used could also be applied to other existing comparable corpora in order to enable their better exploitation in various tasks.

7. References

Laurie Bauer. 1994. *Watching English change: An introduction to the study of linguistic change in standard English in the twentieth century.* London: Longman.

Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994. ARCHER and its challenges. compiling and exploring a representative corpus of historical English registers. In G. Tottie U. Fries and P. Schneider, editors, *Creating and using English language corpora*, pages 1–14. Amsterdam: Rodopi.

Connexor. 2006. Machinese language analysers. *Connexor Manual*.

Gloria Corpas Pastor, Ruslan Mitkov, Afzal Naveed, and Pekar Victor. 2008. Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the AMTA*.

David Denison. 1994. A Corpus of Late Modern English Prose. In *Corpora Across the Centuries. Proceedings of the First International Colloquium on English Diachronic Corpora*, pages 7–16. Amsterdam: Rodopi.

Nelson W. Francis. 1965. A Standard Corpus of Edited Present-Day American English. *College English*, 26(4):267–273.

David G. Garson. 2012a. Significance. Statnotes: Topics in Multivariate Analysis. [http://faculty.chass.ncsu.edu/garson/PA765/signif.htm].

David G. Garson. 2012b. Testing of assumptions: Normality. Statnotes: Topics in Multivariate Analysis.

David G. Garson. 2012c. Tests for two independent samples: Mann-Whitney U, Wald-Wolfowitz runs, Kolmogorov-Smirnov Z, & Moses extreme reactions tests. Statnotes: Topics in Multivariate Analysis. [http://faculty.chass.ncsu.edu/garson/PA765/mann.htm].

Marianne Hundt, Andrea Sand, and Rainer Siemund, 1998. *Manual of Information to Accompany the Freiburg-LOB Corpus of British English*. Freiburg.

Stig Johansson, Geoffrey Leech, and Helen Goodluck, 1978. *Manual of Information to Accompany the*

Lancester-Oslo/Bergen corpus of British English. Department of English, University of Oslo.

Anthony S. Kroch, 2008. *Syntactic Change*, pages 698–729. Blackwell Publishers Ltd.

Geoffrey Leech and Nicholas Smith. 2005. Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal*, 29:83–98.

Geoffrey Leech and Nicholas Smith. 2006. Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English. *Language and Computers*, 55(1):185–204.

Geoffrey Leech and Nicholas Smith. 2009. Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991. *Language and Computers*, 69(1):173–200.

Geoffrey Leech, Marianne Hundt, Christian Mair, and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Geoffrey Leech. 2003. Modality on the move: the English modal auxiliaries 1961-1992. In R. Facchinetti, M. Krug, and F. Palmer, editors, *Modality in contemporary English*, pages 223–240. Berlin/New York: Mouton de Gruyter.

Geoffrey Leech. 2004. Recent grammatical change in English: data, description, theory. *Language and Computers*, 49(1):61–81.

Christian Mair and Marianne Hundt. 1995. Why is the progressive becoming more frequent in English? A corpus-based investigation of language change in progress. *Zeitschrift f´ur Anglistik und Amerikanistik*, 43:111–122.

Christian Mair and Geoffrey Leech. 2006. Current change in English syntax. In B. Aarts and A. McMahon, editors, *The Handbook of English Linguistics*, page Ch. 14. Oxford: Blackwell.

Christian Mair, Marianne Hundt, Geoffrey Leech, and Nicholas Smith. 2002. Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7:245–264.

Christian Mair. 1997. The spread of the going-to future in written English: a corpus-based investigation into language change in progress. In R. Hickey and St. Puppel, editors, *Language history and linguistic modelling: a festschrift for Jacek Fisiak on his 60th birthday*, pages 1537–1543. Berlin: Mouton de Gruyter.

Christer Samuelsson and Atro Voutilainen. 1998. Comparing a linguistic and a stochastic tagger. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eight Conference of the European Chapter of the Association for Computational Linguistics (ACL '98)*, pages 246–253. Association for Computational Linguistics.

Andrea Sand and Rainer Siemund. 1992. LOB-30 years on ... *ICAME Journal*, 16:119–122.

Joseph A. Smith and Colleen Kelly. 2002. Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, 36:411–430.

Nicholas Smith. 2002. Ever moving on? The progressive in recent British English. In P. Peters, P. Collins, and A. Smith, editors, *New frontiers of corpus research: papers from the twenty first International Conference on English Language Research on Computerized Corpora, Sydney 2000*, pages 317–330. Amsterdam: Rodopi.

Nicholas Smith. 2003a. Changes in the modals and semi-modals of strong obligation and apistemic necessity in recent British English. In R. Facchinetti, M. Krug, and F. Palmer, editors, *Modality in contemporary English*, pages 241–266. Berlin/New York: Mouton de Gruyter.

Nicholas Smith. 2003b. A quirky progressive? a corpus-based exploration of the will + be + -ing construction in recent and present day British English. In D. Archer, P. Rayson, A. Wilson, and T. McEnery, editors, *Proceedings of the Corpus Linguistics 2003 Conference*, volume 16, pages 714–723. Lancaster University: UCREL Technical Papers.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis T. 2000. Automatic text categorization in terms of genre and author.

Sanja Štajner and Ruslan Mitkov. 2011. Diachronic stylistic changes in British and American varieties of 20th century written English language. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage at RANLP 2011*, pages 78–85.

Sanja Štajner. 2011. Towards a better exploitation of the Brown 'family' corpora in diachronic studies of British and American English language varieties. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 17–24.

Ingrid Westin and Christer Geisler. 2002. A multi-dimensional study of diachronic variation in British newspaper editorials. *ICAME Journal*, 26:133–152.

Ingrid Westin. 2002. *Language Change in English Newspaper Editorials*. Amsterdam: Rodopi.

Mining for Term Translations in Comparable Corpora

Dan Ştefănescu
Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie, No. 13, Bucharest 050711, ROMANIA
E-mail: danstef@racai.ro

Abstract

This paper presents the techniques currently developed at RACAI for extracting parallel terminology from the comparable collection of Romanian and English documents collected in the ACCURAT project. Apart from being used for enriching translation models, parallel terminology can be (and very often is) a goal in itself, since such resources can be used for building dictionaries or indexing technical or domain-restricted documents.

Keywords: Terminology Extraction, Terminology Mapping, Comparable Corpora

1. Introduction

The construction of any Statistical Machine Translation System requires two types of statistical models: language models and translation models, whose parameters are usually derived from the analysis of parallel corpora. However, large parallel corpora are only available for a quite small number of languages with rich resources (English, French, German, Spanish, etc.) and so, there is an increasing need in gathering parallel data for under resourced languages. One of the recent approaches in solving this task is to extract parallel data from comparable corpora. Such corpora consist in documents covering the same topic or subject, using more or less parallel expressions, entities or terminology. For instance, one can easily find Wikipidia[1] or news articles which are examples of strongly and respectively weakly comparable documents. The goal is to extract, if possible, the existing parallel data and use it to enrich poor translation models.

This paper presents the techniques currently developed at RACAI for extracting parallel terminology from the comparable collection of Romanian and English documents in the ACCURAT project (Skadiņa et al., 2012). Apart from being used for enriching translation models, parallel terminology can be (and very often it is) a goal in itself, since such resources can be used for building dictionaries or indexing technical or domain-restricted documents.

First, the terminology is monolingually extracted, taking into consideration both single and multi-word terms, while in the second step the extracted terms are mapped based on string similarity and existing dictionaries. The methods described are language independent as long as language specific resources are provided. The paper is structured as follows: the next section presents the monolingual terminology extraction, while section 3 describes the terminology mapping. Experiments and results are presented in section 4. The paper ends with conclusions and references sections.

[1] http://en.wikipedia.org/wiki/Romania vs. http://ro.wikipedia.org/wiki/Rom%C3%A2nia (27.03.2012)

2. Terminology Extraction

Terminology extraction is the subtask of Information Extraction which refers to extracting terms from a given corpus, relevant to the genre / domain of the corpus. This task dates back to the 70s and it was most studied in the 90s. This latter period saw an explosion of various approaches (Schütze, 1998) based on raw frequency and part of speech filters (Dolby et al., 1973; Justeson and Kats, 1995), low variance in relative position for multi-word terms (Smadja, 1993), hypothesis testing and mutual information (Church and Hanks, 1989), likelihood ratios on assumed distributions (Dunning, 1993), inverse document frequency on assumed distributions (Church, 1995), finite-state automaton parsing (Grefenstette, 1994), full parsing (Bourigault, 1993; Strzalkowski, 1995), semantic analysis (Pustejovsky et al., 1993), etc. Recent work includes that of Park et al. (2002), who focus on all possible parts-of-speech terminology taking into account out-of-vocabulary words, Wong et al. (2007), who use a probabilistically-derived measure – *Odds of Termhood*, for scoring and ranking term candidates for term extraction, or Velardi et al. (2008), who see the Web as a huge corpus of texts that can be processed to create and update specialized glossaries.

While the existence of various commercially available terminology extraction tools[2] might suggest that this is a sufficiently studied problem, in practice, users complain about the amount of manual work required to filter out much of the terms returned by such systems[3].

Our solution makes a clear distinction between single-word and multi-word terms, since their identification and extraction is usually performed by using different approaches.

[2] http://www.translationzone.com/en/translator-products/sdlmultitermextract/ (27.03.2012)
http://www.e-kern.com/en/kern/translations/terminology/terminology-extraction.html (27.03.2012)
http://www.wordfast.net/ (27.03.2012)

[3] http://www.proz.com/forum/software_applications/96347-terminology_extraction_software.html (27.03.2012)

2.1 Single-word terminology extraction

We approached the task of single-word terminology extraction by improving Damerau's method (Damerau, 1993) as it has been reported to yield very good results (Schütze, 1998; Paukkeri et al., 2008). Damerau's approach compares the relative frequency in the documents of interest (user corpus – C_U) to the relative frequency in a reference collection (reference corpus – C_R). The original formula for computing the score of a word *w* is:

$$score(w) = \frac{f(w, C_U)}{|C_U|} \div \frac{f(w, C_R)}{|C_R|} \quad (1)$$

where $f(w, C)$ is the frequency of *w* in corpus *C*, and $|C|$ is the total number of words in *C*. One can immediately notice that the score for a word is calculated according to the likelihood ratios of occurring in both corpora (that of the user and the reference). The main idea is to compare the maximum likelihood estimates (MLE) computed on the user corpus to the ones on the reference corpus. Consequently, the reference corpus should be a large, balanced and representative corpus for the language of interest. Essentially, the MLE on such a corpus is equivalent with a unigram language model:

$$P_{MLE}(w) = \frac{f(w, C_R)}{|C_R|} \quad (2)$$

In practice, such models are usually used in information retrieval to determine the topic of documents. Thus, Damerau's formula works by comparing two unigram language models.

It has been proven however, that due to data sparseness, the unigrams language models constructed only by the means of MLE behave poorly and that a proper smoothing should be performed (Chen and Goodman, 1998). To do this, we employ a variant of Good-Turing estimator smoothing (Kochanski, 2006) :

$$P_{GT}(w) = \frac{f(w, C_R) + 1}{|C_R| + |V_R|} \cdot \frac{E(f(w, C_R) + 1)}{E(f(w, C_R))} \quad (3)$$

where V_R is the vocabulary (the unique words in C_R) and $E(n)$ is the probability estimate of the word to occur exactly *n* times.

Let us consider a slightly modified example from (Kochanski, 2006): let us say we have a (reference) corpus with 40,000 English words which contains only one instance of the word "*unusual*": $f(w, C_R) = 1$. Let us also say that the corpus contains 10,000 different words that appear once and so, $E(1)$ = 10,000 / 40,000, and that we have 5,500 words that appear twice, giving $E(2)$ = 5,500 / 40,000. Again, let us consider that the total number of the unique words in the corpus is 15,000 ($|V_R|$ = 15,000). The Good-Turing estimate of the probability of "*unusual*" is:

$$P_{GT}(unusual) = \frac{1+1}{40{,}000 + 15{,}000} \cdot \frac{5{,}500/40{,}000}{10{,}000/40{,}000}$$
$$= \frac{2}{55{,}000} \cdot \frac{5{,}500}{10{,}000} = \frac{1}{50{,}000}$$

But using MLE, we would have had a larger value:

$$P_{MLE}(unusual) = \frac{1}{40{,}000}$$

Because the sum of the probabilities must be 1, we have a remaining probability mass (P_R) to be reassigned to the unseen words (*U*). Consequently, for computing the estimated probability of a single unseen word *u_w*, we should divide this mass to the estimated number of unseen words $|U|$:

$$P_{GT}(u\_w) = \frac{P_R}{|U|} = \frac{E(1)}{(|C_R| + |V_R|) \cdot |U|} \quad (4)$$

Going back to Damerau's formula, we have now that:

$$score(w) = \frac{f(w, C_U)}{|C_U|} \div P_{GT}(w \text{ in } C_R) \quad (5)$$

The words having the highest scores are terminological terms. In case C_U is a large corpus, we can also compute Good Turing estimators for the numerator. For small corpora, this is however unreliable since one cannot compute the estimates $E(n)$ with high enough confidence.

This approach can be improved by additional preprocessing of the corpora involved. First, for better capturing the real word distribution, it is better to use word lemmas (or stems) instead of the occurrence forms. Second, the vast majority of the single terminological terms are nouns and therefore one can apply a Part of Speech (POS) filtering in order to disregard the other grammatical categories. Both can be resolved by employing stand-alone applications that can POS-tag and lemmatize the considered texts. As our research and development is mainly focused on English and Romanian, we usually make use of the TTL preprocessing Web Service (Ion, 2007; Tufiş et al., 2008) when dealing with these languages.

The method presented above can be reinforced with the well-known *TF-IDF* (*term frequency – inverse document frequency*) approach (Spärck Jones, 1972), provided that the corpus of interest is partitioned into many documents or that this partitioning can be automatically performed.

As reference corpora we used the *Agenda* corpus (Tufiş and Irimia, 2006) and a collection of *Wikipedia* documents for Romanian, while for English, we also used *Wikipedia* documents.

2.2 Multiple-word terminology extraction

Terminology extraction does not limit to single-word terms and so, one must be able to extract multi-word terminology, too. Smadja (1993) was among the first to advocate that low variance in relative position is a strong indicator for multi-word terminological expressions, which can be found among the collocations of a corpus. These are expressions which sometimes cannot be translated word-by-word using only a simple dictionary and a language model, because they might be characterized by limited compositionality – the meaning of the expression is more than the sum of the meaning of the words composing the collocation.

Different methods have been proposed for finding

collocations. Some counted the occurrences of bigrams and then used a part-of-speech filter in order to rule out those bigrams which cannot be phrases (Justeson and Krats, 1995). Smadja (1993) employed a method based on the mean and the variance of the distances between pairs of words, while others (Church et al., 1991) used *t Test*, *chi square Test*, *Log-Likelihood* or *Mutual Information* for finding pairs of words which appear together in the text more often than expected by chance.

Our approach for the identification and extraction of collocations has been described in several papers (Ştefănescu et al., 2006; Todiraşcu et al., 2009; Ştefănescu, 2010). For the purposes of the current task, we define a collocation as a pair of words for which:
– the distance between them is relatively constant;
– they appear together more often than expected by chance: *Log-Likelihood*.
Looking at this definition, one can notice, that from a strict linguistic point of view, such a construction can be seen as a strong co-occurrence, rather than a collocation.

The first component of our solution is based on a method developed by Smadja (1993). This uses the average and the standard deviation computed on distances between words to identify pairs of words that regularly appear together at the same distance, a fact which is considered to be the manifestation of a certain relation between those words. Collocations can be found by looking for such pairs for which standard deviation is small.

In order to find terminological expressions, we employ a POS filtering, computing the standard deviation for **only** the *noun-noun* and *noun-adjective* pairs within a window of 11 non-functional words length, and we keep all the pairs for which standard deviation is smaller than 1.5 – a reasonable value according to (Manning and Schütze, 1999). This method allows us to find good candidates for multi-word expressions but not good enough. We want to further filter out some of the pairs so that we keep only those composed by words which appear together more often than expected by chance. We do this by computing the Log-Likelihood (LL) scores for all the above obtained pairs, by taking into account only the occurrences of the words having the selected POS-es. We take into consideration the pairs for which the LL values are higher than 9, as for this threshold the probability of error is less than 0.004 according to the *chi square* tables.

We further keep as terminological expressions only those for which at least one of the words composing them can be found among the *single-word* terminological terms, disregarding their context. In this way we aim at filtering out commonly used expressions which have no terminological value.

3. Terminology mapping

Lately, automatic terminology mapping has been well-studied using methods like compositional analysis (Grefenstette, 1999; Daille and Morin, 2008) or contextual analysis (Fung and McKeown, 1997). Still, terminology mapping for languages with scarce resources is less researched (Weller et al., 2011).

Our terminology mapping tool was developed under the name TEA (TErminology Aligner). Given two lists containing monolingually extracted terminology, it is designed to find (in those lists) pairs of expressions which are reciprocal translations. In order to do this, TEA analyzes candidate pairs, assigning them translation scores (*tScore*) based on (i) translation equivalence estimation and (ii) cognates that can be found in those pairs (eq. 6).

$$tScore(pair) = max(te(pair), cg(pair)) \quad (6)$$

The translation equivalence score (*te*) for two expressions is computed based on the word-level translation equivalents existing in the expressions (eq. 7). Each word w_s in the source terminological expression e_s is paired with its corresponding word w_t in e_t such that the translation probability is maximal, according to a GIZA++ (Och and Ney, 2000) like translation dictionary.

$$te(e_s, e_t) = \frac{\sum_{w_s \in e_s} \max_{w_t \in e_t} dicScore(w_s, w_t)}{length(e_s) + \delta} \quad (7)$$

where *dicScore* is the translation equivalence score from the dictionary. The score should be normalized with the length of expression e_s. Still, we modify the denominator in order to penalize (δ) candidate pairs according to the length difference between source and target expressions:

$$\delta = \frac{|length(e_s) - length(e_t)|}{2} \quad (8)$$

The cognate score for two expressions is computed as the Arithmetic mean between two different string similarity measures (eq. 9). The first one (*sm_ld*) is calculated as the *Levenshtein Distance* (LD) in which the expressions are normalized (*norm*) by removing double letters and replacing some character sequences: "*ph*" by "*f*", "*y*" by "*i*", "*hn*" by "*n*" and "*ha*" by "*a*". This type of normalization is often employed by spelling and alteration systems (Ştefănescu et al., 2011). In practice, we modify this function in order to obtain values in the [0,1] interval, which we want to be high in case strings are similar and approach 0 for high differences (eq. 10). The second string similarity measure is simply the *longest common substring* of the two expressions, normalized by the maximum value of their lengths (eq. 11).

$$cg(e_s, e_t) = \frac{sm\_ld + sm\_lcs}{2} \quad (9)$$

$$sm\_ld = 1 - \frac{LD\big(norm(e_s), norm(e_t)\big)}{\min\big(length(e_s), length(e_t)\big)} \quad (10)$$

$$sm\_lcs = \frac{length(LCS(e_s, e_t))}{\max(length(e_s), length(e_t))} \quad (11)$$

The values of *te*(*pair*) and *cg*(*pair*) are taken into account only if they are higher than a threshold, the value of which regulates the tradeoff between precision and recall.

4. Experiments and Results

Evaluation of parallel terminology extraction requires the existence of a *Gold Standard* (GS) containing bilingual mapped terminology relevant to a collection of bilingual comparable texts. The only freely available such GS we know of is Eurovoc (Steinberger et al., 2002). This is "*the thesaurus covering the activities of the EU and the European Parliament in particular*" and it has been described in (Steinberger et al., 2002). We conducted two experiments: the first one was designed to assess the performance of the monolingual terminology extraction, while the second one, the performance of the mapping.

In the first experiment we considered 950 English-Romanian parallel documents from the *JRC-Acquis* corpus (Steinberger et al., 2006). They are all from 2006 and contain about 3.5 million tokens per language (approx. 55 Mb of preprocessed text). To assess the performance of the tool, we generated lists containing only those Eurovoc terms that appeared in these documents for both languages and counted how many of the recognized terms were found in these corresponding restricted lists (Table 1).

	English	**Romanian**
#documents	950	
Size of preprocessed	3.55 mil. tokens 55.1 Mb	3.34 mil. tokens 61.8 MB
Eurovoc terms identified out of those found in the collection having at least 1 occurrence	793 / 2699 29.38%	744 / 1961 37.93%
… 10 occurrences	289 / 1185 24.38%	252 / 815 30.92%
… 50 occurrences	65 / 507 12.82%	63 / 326 19.32
… 100 occurrences	24 / 318 7.54%	33 / 213 15.49%

Table 1: Eurovoc terms identified as terminological

If a word becomes more and more frequent, approaching its occurrence probability in the reference corpus, the tool cannot consider it terminological. This means, that some of the terminology that is valid for the entire JRC-Acquis cannot be discovered by considering only the documents from a single year, even though that terminology appears in those documents.

Regarding this evaluation methodology, one has to keep in mind that the list of Eurovoc terms is neither exhaustive nor definitive and as such, there might be valid non-Eurovoc terms that our application discovers. Examples for English include "*Basel convention*", "*standards on aviation*", "*Strasbourg*", "*national safety standards*", "*avian influenza*" etc. This is the reason for which we are not evaluating this module in terms of standard precision and recall.

For the second experiment, we considered the ideal case in which the monolingual terminology contains only and all the Eurovoc terms. We conducted this experiment for two language pairs: English-Romanian and English Latvian, computing precision (P), recall (R) and F-measure (F1) values. The next tables summarize the results.

Threshold	**P**	**R**	**F1**
0.1	0.563	0.069	0.122
0.2	0.426	0.101	0.163
0.3	0.562	0.194	0.288
0.4	0.759	0.295	0.425
0.5	**0.904**	**0.357**	**0.511**
0.6	0.964	0.298	0.456
0.7	0.986	0.216	0.359
0.8	0.996	0.151	0.263
0.9	0.995	0.084	0.154

Table 2: Terminology Mapping Performance for English-Romanian

Threshold	**P**	**R**	**F1**
0.1	0.347	0.068	0.114
0.2	0.357	0.108	0.166
0.3	0.636	0.210	0.316
0.4	0.833	0.285	0.425
0.5	**0.947**	**0.306**	**0.463**
0.6	0.981	0.235	0.379
0.7	0.996	0.160	0.275
0.8	0.996	0.099	0.181
0.9	0.997	0.057	0.107

Table 3: Terminology Mapping Performance for English-Latvian

We should mention that these ideal experiment settings, in which we deal with parallel data, allow us to assess the performance of our approach in situations which **can be compared** for the languages of interest. The described methodology for terminology identification is **monolingual** and therefore, it does not matter if the initial data is parallel, or merely comparable. The idea here is to allow for comparable scenarios. As the mapping process does not depend on the document collection, but only on the lists of monolingually extracted terms, again, it does not depend directly upon the comparability level of the initial data. In the mapping experiment described above, we were interested in the limit case where the extracted terminology can be entirely mapped. In the case of comparable corpora, the comparability level and the collection genres have both an important impact on the comparability of the monolingually extracted term lists. Accordingly, many terms may not be present in both lists and so, they cannot and should not be mapped. We might even end up with completely unmappable lists. This issue is the subject of further research.

5. Conclusions

This paper presents the techniques currently used for extracting parallel terminology from the comparable collection of Romanian and English documents in the ACCURAT project. The purpose of the this task is to improve the automatic alignment process of comparable corpora, which finally aims at developing better translation models for Statistical Machine Translation systems.

Future work will be focused on improving this approach by introducing a filtering step for eliminating some of the terms which are incorrectly found as terminological, as a consequence of the error propagation caused by the chaining of the statistical modules involved. We are also working on improving the evaluation process and on estimating the performance of our method for several other language pairs.

The mapping module is the basic terminology mapping tool in the ACCURAT project and it is currently involved in mapping terminology extracted for all the languages involved: English, Estonian, German, Greek, Croatian, Latvian, Lithuanian, Romanian and Slovenian.

6. Acknowledgements

This work has been supported by the ACCURAT project (http://www.accurat-project.eu/) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under the Grant Agreement no. 248347.

7. References

Bourigault D. (1993). *An endogenous corpus-based method for structural noun-phrase disambiguation*, in Proceedings of EACL-93, pp. 81--86.

Chen S.F., Goodman J. (1998). *An empirical study of smoothing techniques for language modeling*, Technical Report TR-10-98, Harvard University.

Church K. (1995). *One term or two?*, in Proceedings of SIGIR-95, pp. 310--318.

Church K., Gale W., Hanks P., Hindle D. (1991).*Parsing, word associations and typical predicate-argument relations*, Current Issues in Parsing Technology. Kluwer Academic, Dordrecht.

Church K., Hanks P. (1989). *Word Association Norms, Mutual Information, and Lexicography*, in Proceedings of the 27th Annual Meeting of the ACL.

Daille, B. and Morin, E. (2008). *Effective Compositional Model for Lexical Alignment*. In Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India.

Damerau F. (1993). *Generating and evaluating domain-oriented multi-word terms from text*. Information Processing and Management, 29(4), pp. 433--447.

Dolby J. L., Ross I.C., Tukey J. W. (1973, 1973, 1975, 1973). Index to Statistics and Probability, Vol. 1 The Statistics Cumindex, 2 Citation Index, 3-4 Permuted Title, 5 Locations and Authors. R and D Press.

Dunning T. (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics 19(1), pp. 61–74.

Fung, P. and McKeown, K. (1997). Finding terminology translations from non-parallel corpora.In Proceedings of the 5th Annual Workshop on Very Large Corpora, pp. 192--202.

Grefenstette G. (1994). *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Press, Boston.

Grefenstette, G. (1999). *The World Wide Web as a resource for example-based machine translation tasks*. Translating and the Computer 21, London, UK.

Ion R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD thesis (Romanian), Romanian Academy, Bucharest.

Justeson J.S., Katz S.M. (1995). *Technical Terminology: some linguistic properties and an algorithm for identification in text*. Natural Language Engineering (1), pp. 9--27. Cambridge University Press.

Kochanski G. (2006). *Lecture 4 - Good-Turing probability estimation*, Oxford.

Manning C., Schütze H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge.

Morin, E. and Prochasson, E. (2011). *Bilingual Lexicon Extraction from Comparable Cor-pora Enhanced with Parallel Corpora*. ACL HLT 2011, page 27.

Och, F. J. and Ney, H. (2000). *Improved Statistical Alignment Models*. In Proceedings of the ACL 2000, Hong Kong, China, pp. 440--447.

Park, Y., Byrd, R. J., Boguraev B. (2002). *Automatic glossary extraction: beyond terminology identification*. Proceedings of the 19th International Conference on Computational Linguistics - Taipei, Taiwan.

Paukkeri M., Nieminen I.T., Pöllä M., Honkela T. (2008). A *Language-Independent Approach to Keyphrase Extraction and Evaluation*, in Proceedings of COLING- 08.

Pustejovsky J., Bergler S., Anick P. (1993). *Lexical semantic techniques for corpus analysis*, Computational Linguistics, 19(2), pp. 331--358.

Schütze, H. (1998). *The Hypertext Concordance: A Better Back-of-the-Book Index*. In Proceedings of Computerm '98 (Montreal, Canada, 1998), D. Bourigault, C. Jacquemin, and M.-C. L'Homme, Eds., pp. 101--104.

Skadiņa, I., Aker, A., Glaros, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A., Babych, B. (2012). *Collecting and Using Comparable Corpora for Statistical Machine Translation*, in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.

Smadja F. (1993). *Retrieving Collocations from Text: Xtract*.Computational Linguistics 19, pp. 143--175.

Spärck Jones, K. (1972). *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation 28 (1), pp. 11--21.

Ştefănescu D. (2010). *Intelligent Information Mining from Multilingual Corpora*, PhD thesis (Romanian), Romanian Academy, Bucharest.

Ştefănescu, D., Ion R., Boroş, T. (2011). *TiradeAI: An*

Ensemble of Spellcheckers, in Proceedings of the Spelling Alteration for Web Search Workshop, pp. 20--23, Bellevue, USA.

Ştefănescu, D., Tufiş, D., Irimia, E. (2006). *Automatic Identification and Extraction of Collocations from Texts*, in Proceedings of the 2nd Romanian Workshop for Linguistic Tools and Resources Volume, 3 Nov. 2006, Bucharest, Romania.

Steinberger, R., Pouliquen, B., Hagman, J. (2002). *Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc*, Springer-Verlag.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 2142--2147. Genoa, Italy, 24-26 May 2006.

Strzalkowski T. (1995). *Natural Language information retrieval*, IP&M, 31(3), pp. 397--417.

Todiraşcu, A., Gledhill, C., Ştefănescu, D. (2009). *Extracting Collocations in Contexts*, in Human Language Technology. Challenges of the Information Society, LNCS Series, Springer, Vol. 5603/2009, pp. 336--349. ISBN 978-3-642-04234-8.

Tufiş D., Ion R., Ceauşu A., Ştefănescu D. (2008). *RACAI's Linguistic Web Services*, in Proceedings of the 6th Language Resources and Evaluation Conference (LREC), Marrakech, Morocco, pp. 28--30.

Tufiş D., Irimia E. (2006). *RoCo_News - A Hand Validated Journalistic Corpus of Romanian*, in Proceedings of the 5th LREC Conference, Genoa, Italy, pp. 869--872.

Velardi, P., Navigli, R., D'Amadio, P. (2008). *Mining the Web to Create Specialized Glossaries*, IEEE Intelligent Systems, 23(5), IEEE Press, pp. 18--25.

Weller, M., Gojun, A., Heid, U., Daille, B., Harastaniv, R. (2011). *Simple methods for dealing with term variation and term alignment*. In Proceedings of TIA 2011: the 9th International Conference on Terminology and Artificial Intelligence, November 8-10, Paris, France.

Wong, W., Liu, W. and Bennamoun, M. (2007). *Determining Termhood for Learning Domain Ontologies in a Probabilistic Framework*. In 6th Australasian Conference on Data Mining (AusDM).

Accurate phrase alignment in a bilingual corpus for EBMT systems

George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos and Marina Vassiliou
Institute for Language and Speech Processing, Athena Research Center
6 Artemidos & Epidavrou Str., Paradissos Amaroussiou, 151 25, Athens, Greece
giorg_t@ilsp.gr; mtroullinos@ilsp.gr; s_sofian@ilsp.gr; mvas@ilsp.gr

Abstract

An ongoing trend in the creation of Machine Translation (MT) systems concerns the automatic extraction of information from large bilingual parallel corpora. As these corpora are expensive to create, the largest possible amount of information needs to be extracted in a consistent manner. The present article introduces a phrase alignment methodology for transferring structural information between languages using only a limited-size parallel corpus. This is used as a first processing stage to support a phrase-based MT system that can be readily ported to new language pairs. The essential language resources used in this MT system include a large monolingual corpus and a small parallel one. An analysis of different alignment cases is provided and the solutions chosen are described. In addition, the application of the system to different language pairs is reported and the results obtained are compared across language pairs to investigate the language-independent aspect of the proposed approach.

Keywords: phrase alignment, bilingual corpus, machine translation, EBMT systems,

1. Introduction

The current trend in MT systems is that of automatically extracting as much linguistic information as possible from corpora, either monolingual or bilingual ones. This applies to both Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT). The monolingual corpora are substantially easier to assemble, but cannot be used to create the translation models required by SMT systems, while bilingual corpora provide potentially more information and can be used to produce SMT translation models, but are more expensive to either collect from the web or create manually. The use, as far as possible, of monolingual rather than bilingual corpora can alleviate the need for expensive language resources. Hence, the motivation of the present article is to support the design of an MT system using as far as possible information extracted from monolingual corpora, while also maximising the utilisation of a small parallel corpus of a limited size (typically of a few hundred sentences).

The MT concept adopted here comprises a two-stage process. In the first stage, the structure of the sentence to be translated is transformed from the source language (SL) to the corresponding structure of the target language (TL), while in the second stage this structure is modified and enriched at a sub-sentential level to create the final translation. The entire process is data-driven and draws on linguistic information residing in two types of resources, namely (i) a limited-size bilingual corpus, the processing of which offers the essential information for transforming the SL sentence structure to the TL one, and (ii) a large monolingual corpus, compiled via web crawling, which is exploited in order to refine the translation at a sub-sentential level. This is in summary the concept of the PRESEMT project (www.presemt.eu), which aims to create an MT system that can be readily ported to new language pairs, using an EBMT-type approach.

The processing of this bilingual corpus to establish structural correspondences from the source to the target language is the main theme of the present article. In addition, it is useful to assess automatically the fidelity of translation from SL to TL for each sentence pair, so as to identify pairs where the match is not sufficiently accurate to provide information on the structure transformation from SL to TL.

In the remainder of this article, initially a survey of related research work is performed. This is followed by a description of the principles of the proposed approach. A detailed algorithmic description of the step-wise phrase alignment process is then provided. The required resources that have been assembled for experiments are subsequently presented, followed by the experimental results. This section includes an investigation of the approach accuracy when applied to different language pairs. In addition, an analysis of the source of errors (itemised in terms of the processing steps) is performed. Finally, potential extensions are investigated, such as the ability to assess the suitability of individual sentence pairs to serve as reference material for defining the TL structure, via the phrase aligner approach. This allows the creation of a more appropriate set of bilingual sentences.

2. Processing of bilingual corpora in MT

The majority of current MT systems, encompassing both statistical MT (SMT) and non-statistical MT systems, implement the translation of sentences by operating at sub-sentential level, for instance syntactic phrases, into which these sentence are split. In early SMT, the phrases were derived automatically based on sequences of tokens (Koehn, 2010). However, more recently, improvements are attained by introducing syntactically-valid phrasing. It has been found that the introduction of a parser in an SMT system enables the reordering of the SL side to better match the TL side of the corpus, thus conferring an improvement in translation quality (Collins et al., 2005).

A similar improvement in MT quality by introducing parsers has been identified in other MT paradigms such as EBMT systems, where sentences in SL are provided together with their reference translations in TL. However,

in EBMT systems the definition of appropriate phrases necessitates either (i) the development of matched segmentations that give similar outputs for SL and TL or (ii) the definition of a mapping between SL and TL segmentation schemes. Both these approaches constrain the applicability of an MT system to language pairs for which the segmentation schemes are either directly compatible or are rendered compatible via additional processing (for example by generating transformation rules, mainly by trial-and-error, until a desired level of matching is achieved). A typical example of introducing phrasing in an EBMT approach is the METIS-II data-driven MT system (Markantonatou et al., 2006), where pre-existing parsing tools are employed for both the source and the target languages, but the tools' outputs are further processed to render them compatible. By definition, this heavily constrains the portability of an MT system to new language pairs, due to the need to ensure compatibility between the outputs of tools for different languages in advance.

An alternative solution, which is presented in this article, adopts a novel paradigm that circumvents this bottleneck of parsing scheme agreement, and thus can support the straightforward development of MT systems for new language pairs. This solution employs pattern recognition principles to create matching segmentations for the two languages, which then provide the basis for the transfer from the SL structure to the TL one. Relying on the use of a small bilingual corpus, which typically comprises a few hundred sentences aligned at sentence level, this approach is based on identifying sub-sentential segments in both SL and TL. Rather than trying to harmonise two already existing parsers, it uses a parser only in one language and maps this parsing information to the other language of a given language pair. In other words, given a parser (or more generally a phrasing model) in one of the two languages (either SL or TL), the aim is to generate an appropriate phrasing model for the other language. This is the main principle behind the PRESEMT approach (Tambouratzis et al., 2011). In the proposed implementation of the phrase alignment process, it is assumed that only a TL parser is available. The current work is based on the PAM approach proposed in Tambouratzis et al. (2011), though here the methodology has been extensively reworked to achieve a higher alignment accuracy coupled with enhanced language independence.

The process of defining SL-TL correspondences is achieved by grouping together SL elements (words) to sub-sentential segments (phrases) in accordance to the TL ones rendered by the parser. This approach exploits pattern recognition-based clustering techniques to extend these correspondences so that they cover the entire source language structure, dividing it into TL-based phrases.

3. Literature survey

A number of studies related to the phrase alignment approach proposed in this article have been carried out in the general field of linguistics, to determine the optimal alignment for bilingual corpora, by defining word phrases. A conceptually similar process to the one presented here has been proposed for parse trees by Yamada and Knight (2001), who assume a channel model. According to this model, during the machine translation process the segments (which are tree-based) are modified via three operations, namely reordering, insertion and translation. The information in this case is extracted via statistical methods.

Yarowski and Ngai (2001) propose projecting linguistic annotations from a resource-rich language to a resource-sparse one, in the case of parallel corpora of sentences. These projections are used to support the implementation of linguistic tasks in languages where the annotated material is sparse, via raw bilingual corpora which are automatically aligned. Yarowsky and Ngai (2001) have aimed at transferring shallow-processing tools such as noun phrase chunkers on the basis of word-level alignment between the languages.

The motivation of Tillmann (2003) is to determine blocks of corresponding words in the source and target languages that can then be used to perform statistical machine translation. This is achieved by a two-stage Viterbi-type approach which initially establishes high-precision alignments in terms of words that are in a second phase supplemented by incorporating lower-precision alignments to provide higher word coverage, thus generating blocks of words.

Och and Ney (2004) propose a data-driven approach that operates on corpora that are not linguistically-annotated to determine corresponding sequences of words. The definition of the sequences is performed via a two-stage process, where initially an alignment of words is performed and then aligned phrase pairs are extracted, employing a dynamic programming-type algorithm.

In contrast, Simard et al. (2005) propose a translation method using non-contiguous phrases, which is claimed to allow the coverage of additional linguistic phenomena in comparison to only allowing contiguous phrases. Ganchev et al. (2009) propose a methodology for inducing grammar knowledge for resource-poor languages. This methodology is based on bitexts between the resource-poor target language and a resource-rich language (such as English), where the resource-rich information is transferred to the resource-poor language. Ganchev et al. investigate the effect of introducing language-specific constraints for disambiguating annotation choices as compared to using only the bitext-based knowledge.

Melamed (1997) has studied the problem of correspondence of words in different languages with the aim of estimating a partial translation model that accounts for translational equivalence, only at a word level, based on word co-occurrences. Taskar et al. (2005) have proposed a discriminative method for defining word alignment models based on a selection of features of word pairs and compared this method to statistics-based models such as Giza++. Finally, DeNero et al. (2007) propose an alignment approach aimed to support syntactic machine

translation, using HMM modelling.
An alternative approach for identifying corresponding words has been proposed for EBMT as the Marker Hypothesis. In this hypothesis, specific words are used for signalling phrase boundaries in both the SL and TL (see for instance Gough and Way, 2004). This approach however presupposes the compilation of marker word lists per language; besides, in the approach proposed in the present article, the SL text segmentation is guided by the TL text parsing scheme.

4. Extracting alignments from a bilingual corpus

The methodology proposed here, henceforth called Phrase aligner (PA), aims at extracting phrasal information via mutual alignment of the SL sentences and the TL ones of a parallel corpus. The Phrase aligner requires only one side of the parallel corpus to contain phrasing information that will be provided by an appropriate parser, while the other side only contains lemma and Part-of-Speech (PoS) tag information. By performing word alignments between the sentences of the parallel corpus and clustering all words into phrases based on the phrases found on the parsed side of the corpus, the Phrase aligner effectively extracts a phrasing scheme for the corpus side that has no phrasing information, on the condition that the given phrases in the two languages do not overlap. The extracted alignment information is then exploited to (a) create a phrasing model that can be applied for processing any input sentences for the parser-less language side and (b) create an SL-TL model for structural reordering during the machine translation process.

4.1 Design of the PA algorithm

The Phrase aligner needs three resources, namely an SL-TL bilingual lexicon, a tagger and lemmatiser for both the SL and TL sides of the corpus and a TL parser for yielding the appropriate phrasing scheme. Based on these resources, the following information is available:

- Likely SL-TL word correspondences, as furnished by the bilingual lexicon. These correspondences may be
  - one-to-one (a single SL word translates into exactly a single TL word)
  - one-to-many (a single SL word corresponds to a multi-word TL unit)
  - many-to-one (an SL multi-word unit corresponds to a TL single one)
- SL-to-TL tag correspondence; for languages with rich morphology, possibly additional morphological information.
- In-sentence distances between two words, measured in terms of the number of intervening tokens.
- Decomposition of the TL sentence in sub-sentential segments depending on the parser employed.

Based on this set of inputs, PA needs to decide on the optimal segmentation of the source sentence into phrases. A multi-criterion-type comparison must be performed, where the different inputs are accordingly prioritised and combined. Naturally, not all aforementioned inputs need to be present for the PA to generate results, though use of all inputs yields a more accurate alignment.

4.2 Implementation of the PA algorithm

Similarly to several of the aforementioned systems (cf. Och and Ney, 2004; Ganchev et al., 2009), PA employs a multi-stage process, according to which the establishment of word correspondences is performed in the first stage, and these correspondences are then extended in subsequent stages to eventually cover the entire sentence. More specifically, a three-stage process is implemented, where (i) SL-TL word correspondences are established based on the lexicon, (ii) alignments exploit the similarity of grammatical features and (iii) SL words aligned within the first two stages are used as the nuclei of phrases to which still unaligned SL words are assigned. Each of the three stages is described in detail below.

Stage 1: Alignment of single words

The word aligner algorithm performs alignment of SL words to TL ones based on the information of the bilingual lexicon. It is often the case that SL words have more than one candidate translations. So, let us assume that a given SL word '*A*' has two candidate translations, '*B*' and '*C*', in the bilingual lexicon. If in the TL side of the sentence pair both '*B*' and '*C*' exist, then this multiple word alignment cannot be resolved without additional information, such as, for instance, the information residing in the neighbourhood of words '*A*', '*B*' and '*C*' in the SL and TL sentences.
Figure 1 illustrates an example of such a case, where the SL side comprises four words, denoted '*SL1'* to '*SL4'*, and the TL side comprises four words denoted '*TL1'* to '*TL4'*. According to the lexicon, words '*SL1'* and '*SL4'* have each a single candidate translation (words '*TL3'* and '*TL4'* respectively); but the word '*SL2'* has two candidate translations in the TL sentence, namely '*TL1'* and '*TL2'*. Exploiting information on the environment of '*TL1'* and '*TL2'* to choose between the two candidate translations, a distance-based principle is used to determine the TL word (either '*TL3'* or '*TL4'*) to which an SL word within the vicinity of '*SL2'* is single-aligned and which has a minimum distance from one of the candidate words. In this example, the two distances corresponding to single-aligned words are *dis(SL2,TL1)* and *dis(SL2,TL2)*. Hence, the distance between the SL side and the TL side is expressed as the distance of the candidate translations ('*TL1'* and '*TL2')* from those TL words, to which other SL words, within a given neighbourhood to the SL word in question ('*SL2'*), have already been single-aligned.
In the example of Figure 1, if a neighbourhood size of 1 is used, then only one neighbouring word, namely '*SL1'*, is single-aligned, to '*TL3'*. Since '*TL3'* is situated closer to '*TL2'* than to '*TL1'*, then '*TL2'* will be chosen as the most likely translation of '*SL2'*.
If a neighbourhood size of 2 is used, two neighbouring words are single-aligned, namely '*SL1'* and '*SL4'*, which translate into '*TL3'* and '*TL4'* respectively). In that case, the choice will be based on the smallest mean distance of the two candidate translations, '*TL1'* and '*TL2'* from the

two reference points '*TL3'* and '*TL4'*. The computed distances are as follows:

dis(SL2,TL1)=[dis(TL3,TL1)+dis(TL4,TL1)]/2= [2 + 3]/2 = 2.5

dis(SL2,TL2)=[dis(TL3,TL2)+dis(TL4,TL2)]/2= [1 + 2]/2 = 1.5

Thus, based on the principle of smallest distance, word 'SL2' will again be chosen as the most likely translation of '*TL2'*.

In the general case, for an assignment to be made, this cumulative distance must be below a given threshold, which is a system parameter, so as to avoid aligning words at a large distance to each other.

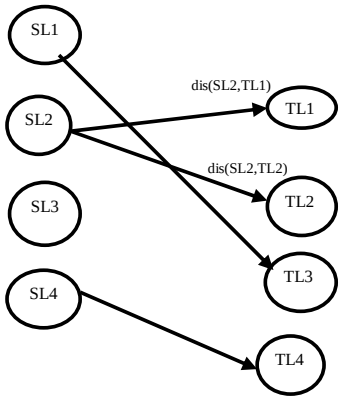


Figure 1: Example of multiple alignments

A similar process is followed in the case of multiple words from the SL side being translated to a TL single word. In this case, a mirror-application of the algorithm is performed, with words in the environment of the SL side being used to establish the minimum distance solution.

Naturally, within a sentence several multiple alignments may exist; their resolution is carried out in the first stage of PA so as to minimise the mean value of distances for all words being examined. In addition, a necessary property is that of independence to the order with which the sentence words are processed. To that end, all decisions aimed at resolving (some of) the multiple alignments are performed while ensuring that the collective distances for the entire sentence are minimised.

Given that (i) a single application of the algorithm will very likely not resolve all ambiguities within a sentence and (ii) the resolution of certain multiple alignments can facilitate the resolution of other pending ones, this algorithm is applied iteratively on a sentence basis, until there exist no further multiple alignments.

An example of a more complex situation is depicted in Figure 2 (distances between TL and SL elements are indicated on the relevant edges, while already aligned words are not shown in order to simplify the figure). There are two SL words, for each of which multiple possible alignments exist, and these alignments overlap. If it is attempted to resolve first the multiple alignment of '*SL1'*, the achievement of a global minimum cannot be guaranteed. On the contrary, by examining word '*SL2'*, it can be seen that '*TL3'* is at a smaller distance than '*TL2'*, and that this is the lowest global distance. By removing the possible association between '*SL2'* and '*TL2'* (as '*SL2'* has already been aligned to '*TL3'*), there remain two candidates for '*SL1'*, namely '*TL1'* and '*TL2'*. Thus, by examining in the second iteration their relative distances, it can be seen that '*TL2'* is a preferable alignment to '*TL1'*. Consequently, in a total of two iterations the entire sentence is disambiguated.

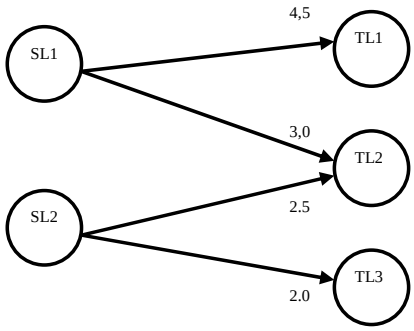


Figure 2: Example of resolvable multiple alignments needing more than one iterations to be resolved

A different situation is depicted in Figure 3. More specifically, though the number of words and of multiple alignments is exactly the same, the relevant distances differ. So, though the first iteration will again assign '*SL2'* to '*TL3'*, the second iteration cannot decide on a TL word to which word '*SL1'* should be assigned. This illustrates the effect of the relevant magnitude of distances on the disambiguation process. To avoid reaching sub-optimal solutions it has been decided not to force the resolution of such cases in stage 1, but re-examine candidate solutions at later stages.

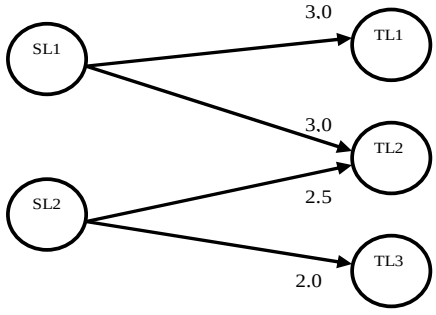


Figure 3: Example of non-fully resolvable multiple alignments needing more than one iterations

These examples illustrate the approaches that the PA employs in order to resolve as many as possible multiple alignments provided by the lexicon. The limited coverage of the lexicon is overcome through two language-independent mechanisms:

(i) Matching of numeric words, when their actual strings

match. As this mechanism is almost certain to lead to the correct assignment, its application precedes accessing the lexicon.

(ii) Transliteration process to a common character set, when SL and TL differ in terms of alphabet (for instance English and Greek). A comparison between the transliterated words of the SL and TL sides is performed to map so far unmatched words, provided their transliterations have a similarity exceeding a given threshold. This operation is applied at the very end of stage 1, after all lexicon-extracted information has been used. This allows the similarity threshold to be set to a lower value without affecting the output of the lexicon-matching process.

At the end of Step 1, alignments using single-word information are resolved to the greatest extent possible. Any words that cannot be unambiguously aligned are forwarded to the next two stages for resolution.

Stage 2: Alignment based on feature similarity

Stage 2 processes the output of Stage 1, with the aim of increasing the percentage of words aligned between the SL and TL sentences. In this stage, the resolution of so far unassigned SL words is based on similarity of grammatical features (e.g. case, number etc.), to be found in the extended PoS tags. Hence, the extended tags of still unassigned SL words are matched to those of other SL words that have been unambiguously aligned in the previous stage. Amongst these matches, the one with the highest similarity is selected, since that indicates a high likelihood of association between the matched words. The tag similarity is normalised by multiplying with a Gaussian function that takes as its input the distance in terms of tokens of the two words on the sentence. Consequently, the tag similarity is reduced as the physical distance in the sentence increases. This normalisation allows the assignment of SL words to the same phrase, provided that they match to an acceptable extent in terms of grammatical features but are also relatively closely situated within the sentence. The variance of the Gaussian function is tuneable to the application requirements.

The aforementioned algorithm is effective only for inflected words such as verbs, nouns, adjectives, pronouns and yields good results in the case of morphologically rich languages. However, it can still be applied in morphologically poor languages without loss of generality, though naturally the number of words aligned by it will be limited.

The present phrase alignment stage is aimed to maximise the coverage and accuracy of word alignments. Hence, an additional effort involves aligning yet-to-be-assigned SL words to TL ones based on the inter-language tag correspondence. This type of information is of a statistical nature and is extracted in an unsupervised manner from the bilingual lexicon by studying macroscopically the average frequency with which any SL word of PoS type 'X' is translated to an also unaligned TL word of PoS type 'Y'. Assuming that the majority (exceeding a chosen threshold) of words of PoS type 'X' do translate into words of PoS type 'Y', then an unaligned SL word of PoS type 'X' could be assigned to a TL word of PoS type 'Y' to improve the phrase aligner coverage. If there exist more than one TL words of PoS type 'Y', the most likely one can be determined by applying the neighbourhood-based principle, as described in Stage 1.

Stage 3: Alignment based on neighbourhood

Stage 3 operates on the output of Stage 2, with the aim of grouping the residual unaligned SL words to TL phrases. This is achieved via two methods. In the first method, grammatical feature similarity is taken into account, as introduced in stage 2, the difference being that at this third stage the principle of normalising over the distance applies to TL phrases instead of TL words. The second method forces an unaligned SL word to be assigned to the TL phrase to which the majority of its SL side immediate neighbours belong.

5. Experimental setup

Since the PA methodology is language-independent, the Phrase aligner module has been tested so far on three language pairs, Greek – English and German – English and English – German, all of which involve languages with a different word order (English has a fixed word order, Greek has a free word order, while German is a V2 language). In the present article, the experiments on the first two pairs are reported. For each pair a bilingual parallel corpus has been built from the web. For both the SL and TL sides the corpus has been processed using readily available language tools as detailed below.

The SL side of the corpus is then manually edited so that it would be "close" to the TL one, removing metaphors or elliptical constructions and smoothing out divergences between the two languages. Moreover, for the reported experiments, the corpus NLP annotations have been manually corrected, so as to focus on testing the PA performance on data devoid of errors. Future experiments will study the effect of the actual annotations (which will unavoidably contain errors) on the performance of the phrase aligner.

Greek - English corpus: Extracted from a multilingual website[1], this corpus comprises 200 sentences. The SL side of the corpus has been tagged and lemmatised by the FBT Tagger-Lemmatiser (Papageorgiou et al., 2000), while the TL side has been processed with the TreeTagger for English (Schmid 1994), yielding tag, lemma and phrase annotations.

German - English corpus: Also extracted from a multilingual website[2], it comprises 164 sentences. The SL side of the corpus has been tagged and lemmatised by the TreeTagger and the RFTagger (Schmid and Laws, 2008), while the TL side has been processed with the TreeTagger for English, generating tag, lemma and phrase annotations.

5.1 Experimental results

For assessing the segmentation accuracy obtained by the

[1] http://europa.eu/abc/history/index_en.htm
[2] http://europa.eu/abc/12lessons/index_en.htm

phrase aligner, its output was compared with a gold-standard reference set. This set included all SL sentences of the aforementioned corpora manually segmented into phrases in accordance to the TL side phrasal segmentation. In other words, the SL side was segmented in those phrases, which PA was expected to generate.

For the purposes of the experiment, two gold-standard sets have been created, of 50 sentences each, for the Greek – English corpus (EL-EN), and two sets, of 50 sentences each, for the German – English (DE-EN) corpus. The degree of match of the PA result to the gold-standard for both language pairs is reported in Table 1, where the best accuracies are denoted in boldface.

Different configurations have been examined, using different values for system parameters. Among the system parameters used, the configurations reported here vary in terms of only certain parameters to which the system is more sensitive, namely (i) the maximum distance for a single alignment to be made, (ii) the minimum required transliteration similarity, (iii) the minimum extended tag similarity threshold, and (iv) the minimum required number of lexicon entries of a given SL tag for which the most likely TL tag is defined in the latter part of Stage 2. The values of these parameters are listed in Table 2 for a number of experimental configurations.

Configuration	Accuracy			
	EL-EN Set1	EL-EN Set2	DE-EN Set1	DE-EN Set2
A	93.74	91.64	**88.50**	88.96
B	**94.51**	92.16	88.23	88.11
C	**94.51**	**93.09**	88.23	88.11
D	94.38	92.28	88.49	89.46
E	94.32	**93.09**	87.92	**90.09**

Table 1: PA experimental results for the EL-EN and DE-EN corpora with variant configurations

System Parameters	Configuration				
	A	B	C	D	E
Distance threshold	3.0	2.0	2.0	3.0	2.0
Translit. similarity threshold	0.5	1.0	1.0	0.5	0.5
Extended tag threshold	0.1	0.5	0.5	0.5	0.5
Threshold of lexicon entries per SL tag	100	0	100	10000	10000

Table 2: Configurations tested for the system parameters

A first observation is that the results across the two sets for each language pair are very similar, indicating that the PA behaviour can be expected to be consistent over a variety of texts. Furthermore, all tested configurations of parameter values (the configurations reported in Table 2 are the more effective ones out of the set examined) give rise to similar results, with a deviation of less than 2% in terms of accuracy.

Another observation concerns the actual accuracy of the phrase aligner. This averages over 94.5% in the case of the Greek – English language pair over a given set of sentences. Since certain sentences give very low alignment accuracies, the actual accuracy over the 'better' sentences is even higher. So, if the sentences to be aligned and then used in the translation process are filtered in advance to remove those with low alignment accuracy, the collective alignment can be substantially higher.

In the case of German – English, the peak accuracy is just over 90%. This is lower than the accuracy reported for the Greek – English pair but still represents a high accuracy. The reduced accuracy for German – English can be mainly attributed to the more complex alignments involved due to the very productive compounding mechanism of the German language, which increases the difficulty of identifying word-to-word alignments.

5.2 Studying the system performance

By analysing the system operation, it is possible to determine which stages are the more effective ones, and which may provide the basis for further improvement. The results summarised in Table 3 are yielded by the optimal configuration (configuration 'C') for the Greek – English corpus; those in Table 4 derive from the same configuration, when applied to the German – English corpus.

In both cases, the accuracy reported is calculated over the entire set of 100 sentences for which gold-standard phrases have been defined.

Greek – English			
	Erroneous alignments	Correct alignments	Accuracy
Stage 1	29	1198	97.6%
Stage 2	15	134	89.9%
Stage 3	61	324	84.2%
Total	**105**	**1656**	**94.0%**

Table 3: Accuracy of each stage of the alignment process for the EL-EN corpus

German – English			
	Erroneous alignments	Correct alignments	Accuracy
Stage 1	82	1601	95.1%
Stage 2	5	13	72.2%
Stage 3	191	325	63.0%
Total	**278**	**1939**	**87.5%**

Table 4: Accuracy of each stage of the alignment process for the DE-EN corpus

According to Tables 3 and 4, as the PA operation proceeds from stage 1 to stage 3, the alignment accuracy decreases in each subsequent stage. This is expected, as in each

stage, less reliable information is employed to perform the alignments, in order to improve the coverage in terms of aligned words. However, for the given phrase alignment result to be useful in the MT process, it is essential to achieve a full coverage of the SL sentences and to this end all stages must be applied.

6. Comparison to Existing Methods

Comparative experiments have been performed in order to obtain a better perspective of the accuracy achieved by the Phrase aligner. GIZA++ was used as a baseline to perform alignments between the SL and TL phrases of the corresponding sentences in the bilingual corpora that PA has been developed on (even though it should be mentioned that GIZA++ is not primarily designed for such a task). The comparison results (cf. Table 5) are promising, as, for both Greek – English and German – English corpora, the accuracy attained by PA is substantially higher than that of GIZA++.

Comparison to Baseline	Corpus	GIZA++
Precision	EL-EN	**72.21%**
Recall		60.98%
Precision	DE-EN	**74.64%**
Recall		71.01%

Table 5: Giza-based experimental results

7. Evaluating sentence pairs' suitability

In the PRESEMT architecture, the limited-size parallel corpus determines the structure of the translation. As the creation of a parallel corpus is a labour-intensive process, it is essential to be able to determine the level of direct correspondence between the SL and TL sides. As described before, alignments are performed in three distinct stages, with each subsequent stage having a lower dependability than previous ones. Consequently, by measuring the percentage of words aligned after each stage for each sentence pair, an estimate of the sentence pair dependability is provided. This can then be used to filter out corpus sentence pairs with a low correspondence between SL and TL, this being reflected by the resolution of alignments for many sentence words in later stages (for instance stage 3). Of course, this estimate also depends on the coverage of the bilingual lexicon used, which can affect the accuracy of the given sentence pair alignments.

8. Further Extensions

In this article, a phrase alignment approach has been presented which generalises the phrasing scheme drawn from the parsed TL side of a bilingual corpus to the non-segmented SL side. This approach is used as a first processing stage to support a phrase-based MT system that is readily portable to new language pairs. A detailed analysis of alignment phenomena, coupled with the application of the system to different language pairs indicate the language independence of the proposed approach.

Within the next period, it is aimed to integrate this mechanism to the PRESEMT system in order to investigate the effectiveness of the approach.

Algorithm-specific improvements possibly entail the refinement of the distance definition, in order to take into account the phrase boundaries when identifying the limits of a word environment. Besides, it is planned to apply the algorithm to more language pairs, including Greek-to-German and English-to-German, with the aim of gaining further insight with respect to the characteristics of the proposed approach.

Up to date, the developed MT language pairs in PRESEMT have been based on the use of parallel corpora. In the following period, it is intended to employ SL-TL comparable corpora, with the aim of evaluating the PA performance on non-strictly parallel corpora and the consequent effect on the performance of the PRESEMT system. Provided the translation accuracy is of a sufficient level, this may allow the simpler development of new language pairs, potentially reducing the effort required for generating high-quality parallel corpora.

Upon completion, the phrase aligner will also be released as public software, available to be incorporated in other applications, with the expectation that it will be of interest and of benefit to the wider research community.

9. References

DeNero, J. and Klein, D. (2007). Tailoring Word Alignments to Syntactic Machine Translation. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007, pp. 17--24,.

Clause Restructuring for Statistical Machine Translation (2005) Collins, M., Koehn, P., Kucerova, I. (2005). Proceedings of the 43rd Annual Meeting of the Association for Computational Linguists (ACL-05), Ann Arbor, USA, June 2005, pp. 531--540.

Ganchev, K., Gillenwater, J., and Taskar, B. (2009). Dependency Grammar Induction via Bitext Projection Constraints. Proceedings of the 47th Annual Meeting of the ACL, Singapore, 2-7 August 2009, pp. 369--377.

Gough, N. and Way, A. (2004). Robust Large-Scale EBMT with Marker-Based Segmentation. Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04), pp. 95--104.

Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press, Cambridge.

Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, G., Vassiliou, M. and Yannoutsou, O. (2006). Using patterns for machine translation (MT). Proceedings of the 11th annual Conference of the European Association for Machine Translation. Oslo, Norway, pp. 239-246.

Melamed, D. (1997). A Word-to-Word Model of Translational Equivalence, Proceedings of the 35th Conference of the Association for Computational Linguistics, Madrid, Spain, pp. 490--497.

Och, F.J., and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, 30(4), pp. 417--449.

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. (2000). A Unified POS Tagging Architecture and its Application for Greek. LREC-2000 Conference Proceedings, Athens, Greece, pp. 1455--1462.

Schmid, H., and Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained PoS Tagging. Proceedings of COLING 2008, Manchester, Great Britain, pp. 777--784.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Simard, M., Cancedda, N., Cavestro, B., Dymetman, M., Gaussier, E., Goutte, C., Yamada, K., Langlais, P. and Mauser, A. (2005). Translating with Non-Contiguous Phrases. Proceedings of the Conferences on Human Language Technology and on Empirical Methods in Language Processing, Vancouver, Canada, pp. 755--762.

Tambouratzis, G., Simistira, F., Sofianopoulos, S., Tsimboukakis, N. and Vassiliou, M. (2011). A resource-light phrase scheme for language-portable MT, Proceedings of the 15th International Conference of the European Association for Machine Translation, (eds. M. L. Forcada, H. Depraetere and V. Vandeghinste) 30-31 May 2011, Leuven, Belgium, pp. 185--192.

Taskar, B., Lacoste-Julien, S. and Klein, D. (2005). A Discriminative Matching Approach to Word Alignment. Proceedings of the HLT/EMNLP Conference, Vancouver, October 2005, pp. 73--80.

Tillmann, C. (2003). A Projection Extension Algorithm for Statistical Machine Translation. Proceedings of the EMNLP Conference, pp. 1--8.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. Proceedings of the 39th Annual ACL Meeting, July 9-11, Toulouse, France, pp. 523--530.

Yarowsky, D. and Ngai, G. (2001). Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. Proceedings of NAACL-2001 Conference, pp. 200--207.

10. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248307.

Using Czech-English Parallel Corpora in Automatic Identification of *It*

Kateřina Veselovská, Ngụy Giang Linh, Michal Novák

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800
{veselovska,linh,mnovak}@ufal.mff.cuni.cz

Abstract

In this paper we have two goals. First, we want to present a part of the annotation scheme of the recently released Prague Czech-English Dependency Treebank 2.0 related to the annotation of personal pronoun *it* on the tectogrammatical layer of sentence representation. Second, we introduce experiments with the automatic identification of English personal pronoun *it* and its Czech counterpart. We design sets of tree-oriented rules and on the English side we combine them with the state-of-the-art statistical system that altogether results in an improvement of the identification. Furthermore, we design and successfully apply rules, which exploit information from the other language.

Keywords: personal pronoun *it*, pleonastic *it*, automatic identification, parallel corpus, coreference resolution

1. Introduction

In the majority of cases in English, the pronoun *it* illustrates nominal anaphora, tending to refer back to another noun phrase in the text. These cases have been surveyed as a part of anaphora resolution research and described e.g. in (Mitkov, 2002) or (Kučová et al., 2003). However, in a minor but still large enough class of cases, the pronoun *it* is used in exceptional ways that fail to demonstrate strict nominal anaphora and can be used without referring to any specific entity. In the present study we investigate mainly these occurrences.

Needless to say that the identification of pronouns to nominal expressions constitutes an important component of the process of coreference resolution, which has been found to be crucial in the fields of information extraction (Hirschman, 1997), machine translation (Peral et al., 1999), and automatic summarization (Harabagiu and Maiorano, 1999).

The English personal pronoun *it* can be translated into Czech as a demonstrative pronoun *to (this / that)* or a personal pronoun in singular *on / ona / ono (he / she / it)*, since English third person singular pronouns are distinguished according to animacy and gender, whereas Czech third person singular pronouns are used to identify grammatical gender only.

(1) Vezmu si **to**.
I will take RFLX **it**.
'I will take **it**.'

(2) (**Ono**) Je těžké v době krize sehnat práci.
(**It**) is difficult in times of crisis to get job.
'**It** is difficult in times of crisis to get a job.'

(3) Společnost Faulding uvedla, že (**ona**) vlastní
Company Faulding said, that (**she**) owns
33 % akcií společnosti Moleculon.
33% of voting stock of company Moleculon.
'Faulding said **it** owns 33% of Moleculon's voting stock.'

The Czech demonstrative pronoun *to* is usually used to refer back to a substantial section of a text, hence in this work we have decided to focus on the third person singular pronouns as the equivalents of the English *it* only. As mentioned before, the automatic identification of personal pronouns (coreferential or not) in English as well as in Czech plays an important role in coreference resolution.

In the present paper, the occurrences of personal pronoun *it* are identified using a parallel Czech-English dependency data collected in the Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) (Hajič et al., 2011). The English part of PCEDT 2.0 contains the entire Penn Treebank-Wall Street Journal Section (Marcus et al., 1999). The Czech part consists of Czech translations of all of the Penn Treebank-WSJ texts. The corpus is 1:1 sentence-aligned. PCEDT 2.0 is a collection of linguistically annotated tree structures which is based on the theoretical framework of Functional Generative Description (FGD) (Sgall et al., 1967; Sgall, 1969). The annotation scheme of the PCEDT 2.0 consists of three layers: morphological, analytical and tectogrammatical. In the present study, we will mostly pursue the tectogrammatical layer (i.e. underlying structure).

The goal of this work is to use the benefits of the manually annotated parallel data in PCEDT 2.0 to construct a tool to determine anaphoricity of *it* or its Czech counterpart, even on the automatically analyzed data. Furthermore, our long-term objective is to improve the coreference resolution using bilingual parallel data not only from PCEDT 2.0, but also from much larger parallel corpus CzEng 1.0 (Bojar et al., 2011).

This paper is organized as follows. The English *it* and its Czech equivalent classification is described in Section 2. Section 3. provides a brief survey of related work. Section 4. presents the data we use for our system development. Description of the experiments for English and Czech is given in Section 5. and Section 6. Section 7. follows with the use of the parallel data. In Section 8., conclusions and ideas for future work are presented.

2. Theoretical Background

There have been several uses of *it* in English identified in the literature (Quirk et al., 1985; Sinclair, 1995; Swan, 1995). In FGD, we distinguish five basic types of personal pronoun *it* according to their function. They are described by the examples below:

1. The **anaphoric *it*** refers to a preceding noun denoting an inanimate entity or a not personalized animal.

   (4) I bought a new hat but my husband did not like *it*.

2. The **anticipatory *it*** anticipates on a part of the sentence which appears later in subject as well as in object position:

   (5) *It* is no good bothering about *it*.

   (6) *It* is feared that the ship was wrecked.

3. The **deictic *it*** belongs to deictic personal pronouns in general. It is used for deixis out of the language. The deictic pronoun as well as the copula verb must be in morphological agreement with the entity *it* refers to. The need of number agreement is typical of the deictic *it*.

   (7) Is *it* your suitcase (over there)?

4. The **exclamative *it*** is also used in deictic contexts but it refers to a situation implicitly known in the discourse rather than immediately to the given entity:

   (8) (Knock knock knock...) "*It*' s me, open the door!"

5. The **prop *it*** has little or no semantic content. It occurs in clauses which do not require any subject. It is typically clauses signifying time, atmospheric conditions and distance where the copula verb to be is regarded:

   (9) *It* is not far to New York.

   (10) *It* is 5 o'clock.

   (11) *It* is our wedding anniversary next month.

   (12) *It* is Sunday.

In Czech, it is natural to drop out personal pronouns in subject position of the clause. An overt subject pronoun indicates an emphasis of the speaker. Nevertheless the unexpressed subject pronoun can be understood from the verb morphological information thanks to its morpheme that identifies person, number and in some cases also gender.[1] In Nguy and Ševčíková (2011) four types of unexpressed subjects are distinguished:

1. The **implicit subject** most often stands for an entity already mentioned in the text or can be deictic.

   (13) Jana$_i$ ráda peče. Dnes Ø$_i$
   Jane gladly bakes. Today (she)
   upekla jablečný koláč.
   baked$_{3.SG.FEM}$ apple pie.
   'Jane likes to bake. Today she has baked an apple-pie.'

2. The **general subject** does not refer to any concrete entity; it has a general meaning, so it can be omitted in the surface structure.

   (14) S rizikem se Ø počítá.
   With risk RFLX (one) counts$_{3.SG}$.
   'Risk is counted in. (One counts risk in.)'

3. The **unspecified subject** denotes an entity more or less known from the context which is however not explicitly referred to.

   (15) Ø Hlásili to v rádiu.
   (They) Announced$_{3.PL.ANIM}$ it on radio.
   'It was announced on radio. (They announced it on radio.)'

4. The **null subject** does not refer to any entity in the real world. It is neither phonetically realized, nor can be lexically retrieved. In this case the predicate is an impersonal (weather) verb.

   (16) Zítra Ø bude oblačno.
   Tomorrow (it) will$_{3.SG}$ cloudy.
   'Tomorrow it will be cloudy.'

For the coreference resolution purpose, the personal pronoun distinction is simplified to **referential** and **non-referential**. As shown in (Evans, 2001; Nguy and Ševčíková, 2011), the automatic identification of other types has a poor accuracy because of its low occurrence.
The non-referential *it* is also referred to as **non-anaphoric** (Mitkov, 2002), **pleonastic** (Lappin and Leass, 1994) or **prop it** (Quirk et al., 1985).
We adopted the categorization from the PCEDT 2.0 annotation, which is as follows:

anaphoric – English anaphoric and anticipatory *it* and its equivalent Czech anaphoric unexpressed implicit third person singular subject.

non-anaphoric – English deictic and exclamative *it* and Czech deictic unexpressed implicit third person singular subject.

pleonastic – English prop *it* and Czech unexpressed general and null subject.

3. Related Work

Pleonastic pronouns have been resolved in a number of research on anaphora resolution. Lappin and Leass (1994)'s and Denber (1998)'s algorithm is based on pattern recognition, e.g. 'It is {a modal adjective} that'. Paice and Husk (1987)'s approach improves the pattern-matching process

[1] Gender is recognizable in past participle form of verbs only.

by constraints. As an illustration, a pronoun *it* is identified as non-referential if it occurs in the sequence 'it ... that'.
Evans (2001) proposes a machine learning based system for the automatic classification of *it*, which attempts to classify *it* for different usages such as nominal anaphoric, clause anaphoric, idiomatic, pleonastic and others. However, the system reports a high accuracy only on classifying pleonastic and nominal anaphoric *it*. The reason is simple, the features used in the training process are most appropriate for classification of pleonastic instances, and other types of *it* occur quite rare.
In recent years the study of pleonastic *it* identification has shifted toward different machine learning methods such as using support vector machines in (Litrán et al., 2004) or using a Bayesian network in (Hammami et al., 2010). Charniak and Elsner (2009) detect non-referential *it* in a unsupervised generative model. The detection of non-referential pronouns using counts from web-scale N-gram data is described in (Bergsma and Yarowsky, 2011).
For a task related to ours, a parallel corpus is used in (Camargo de Souza and Orăsan, 2011). Camargo de Souza and Orasan present a coreference resolution system for Portuguese trained on an English-Portuguese parallel corpus. The noun phrase coreference chains are identified thanks to the projected English coreference chains, which have been obtained from an English coreference resolver. Mitkov and Barbu (2002) develop a bilingual pronoun resolution system for English and French using an English-French parallel corpus, which benefits from the gender distinction of *it* in French and from the performance of the English algorithm.

4. Annotated Data

PCEDT 2.0 contains 2312 documents annotated at the tectogrammatical layer of Czech and English. Altogether, they consist of 49 208 pairs of sentences. Personal pronoun *it* has been annotated manually in all this data, independently in Czech and English part of the corpus, with the automatic word-alignment done afterwards (Mareček et al., 2008), including the alignment between nodes of the tectogrammatical layer.

4.1. Layers of Annotation

The PCEDT 2.0 annotation consists of multiple linguistically motivated layers:
The **m-layer** (morphological layer) captures the surface form of the sentence with words automatically part-of-speech tagged and lemmatized.
The **a-layer** (analytical layer) represents the surface syntax (a parse). The syntactic dependencies are provided with labels that carry the usual syntactic information; e.g. 'subject', 'attribute' or 'predicate complement'. Figure 1 presents the visualization of an analytical sentence representation.
The **t-layer** (tectogrammatical layer) is a linguistic representation that combines syntax and, to a certain extent, semantics, in the form of semantic labeling, coreference resolution[2] and argument structure description based on a valency lexicon. This representation draws on the framework of the Functional Generative Description.
The **p-layer** (phrase-structure layer) contains the original Penn Treebank annotation.

[2]Within the theoretical framework of FGD, coreference is divided into two subtypes: grammatical and textual (Panevová, 1991). **Grammatical coreference** occurs if the antecedent can be identified using grammatical rules and sentence syntactic structure (e.g. reflexive pronouns usually refer to the subject of the clause), whereas **textual coreference** is more context-based (e.g. personal pronouns).

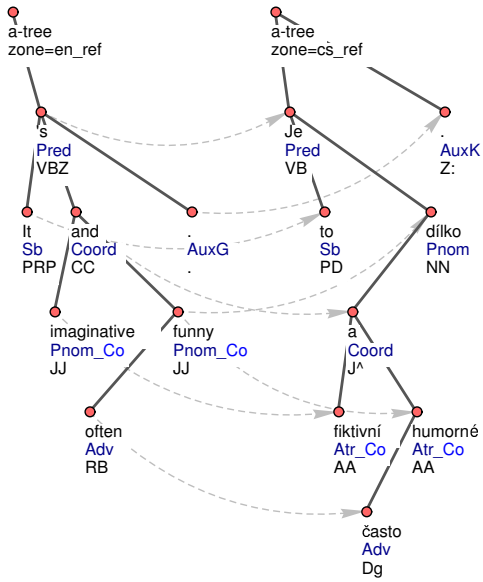


Figure 1: An example of parallel Czech-English a-trees representing sentences *It's imaginative and often funny* and *Je to fiktivní a často humorné dílko*.

4.2. Fully Automatic Annotation

In our study we use both manually annotated PCEDT 2.0 data and the same data automatically analyzed within the Treex framework (Žabokrtský, 2011).
Treex is a multi-purpose open-source framework for developing Natural Language Processing applications, which provides a wide range of integrated modules, such as tools for sentence segmentation, tokenization, morphological analysis, part-of-speech tagging (Spoustová et al., 2007), shallow and deep syntax parsing (McDonald et al., 2005), named entity recognition, anaphora resolution and others.
For our development we have the tokenized plain text from the PCEDT 2.0 of both languages as an input. Then we apply all possible tools in Treex to get them annotated at all layers. After that we used the automatic alignment tool. An example of the final alignment of Czech gold and automatic and English gold and automatic data at t-layer is shown on Figure 2.

4.3. Quantitative Properties

Thanks to the PCEDT 2.0 features mentioned in previous section we could easily distinguish three basic types of *it* in our corpora:

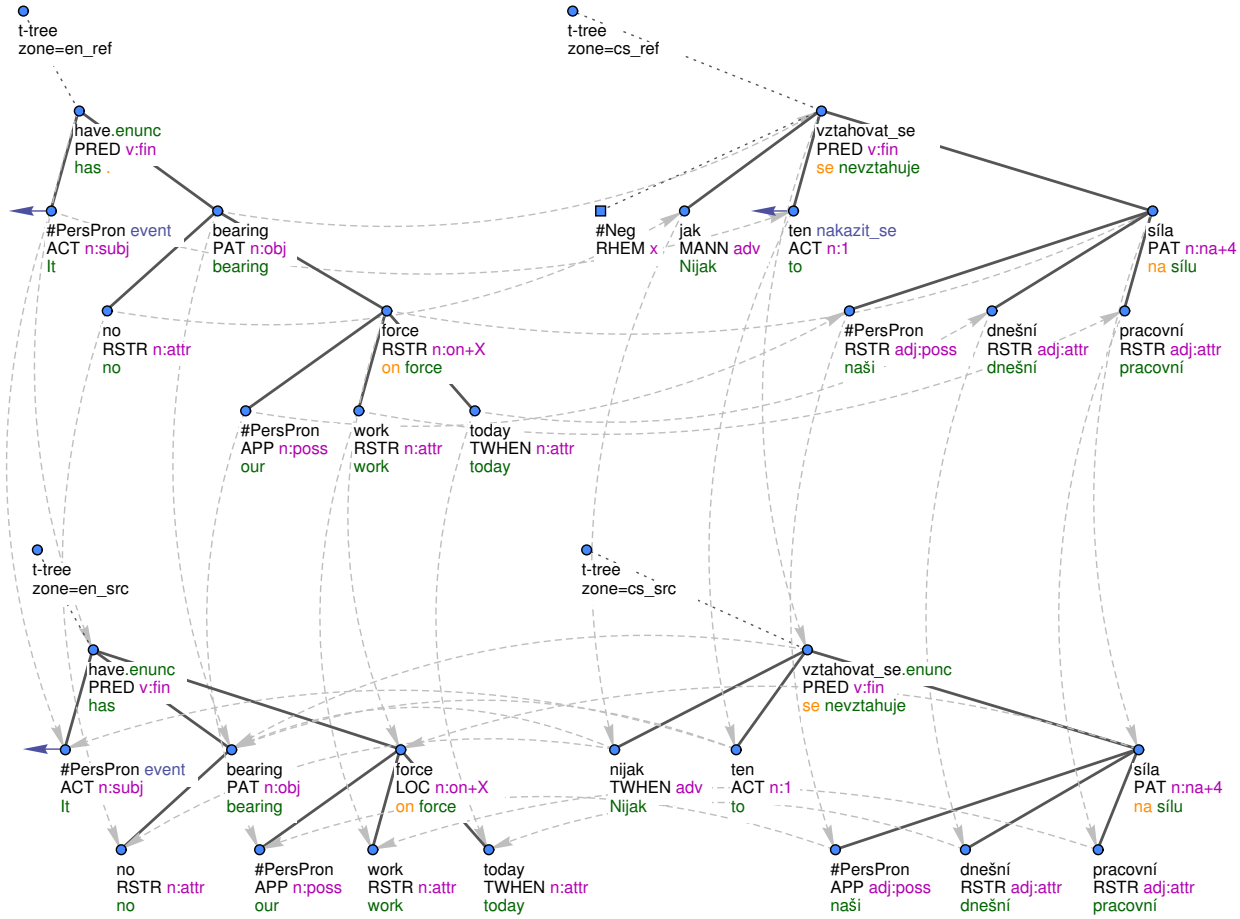


Figure 2: An example of gold parallel Czech-English t-trees aligned with automatic ones ([left to right, top to bottom]: English gold tree, Czech gold tree, English automatic tree, Czech automatic tree) representing sentences *It has no bearing on our work force today* and *Nijak se to nevztahuje na naši dnešní pracovní sílu*.

anaphoric – having a t-lemma substitute `#PersPron` (artificial t-lemma for overt and unexpressed personal pronoun[3]), an a-lemma `it` and a link pointing to its antecedent.

non-anaphoric – having a t-lemma substitute `#PersPron` and an a-lemma `it`, but not having a link pointing to its antecedent.

pleonastic – not having its own t-node on a tectogrammatical layer.

Their Czech equivalents are as follows:

anaphoric – a generated node representing third person singular pronoun having a t-lemma substitute `#PersPron` and a link pointing to its antecedent.

non-anaphoric – a generated node representing third person singular pronoun having a t-lemma substitute `#PersPron`, but not having a link pointing to its antecedent.

pleonastic – a generated node having a t-lemma substitute `#Gen` (artificial t-lemma for grammatical ellipsis of an obligatory argument - general argument) or not having its own t-node on a tectogrammatical layer.

Table 1 shows occurrence frequencies of anaphoric, non-anaphoric and pleonastic pronoun *it* on the English side and its counterparts on the Czech side of the PCEDT 2.0 subsets, we used for experimenting (see the following section).

	Dev data		Eval data	
	English	Czech	English	Czech
anaphoric	2053	4599	1932	3954
non-anaphoric	652	19	425	16
pleonastic	396	349	393	293

Table 1: Personal pronoun *it* number in PCEDT 2.0

We detected 911 occurrences of English anaphoric *it*, which has a Czech equivalent as a demonstrative pronoun *that (to)*; 3085 English non-pleonastic *it* having an equivalent Czech personal pronoun; 11 English pleonastic *it* that has a Czech pleonastic equivalent and 10 Czech pleonastic *it* with an English pleonastic corresponding node; 81 English and

[3] `#PersPron` also stands for textual ellipsis - obligatory arguments of a governing verb / noun.

21 Czech anaphoric *it* that refers to a clause or a sequences of sentences.

4.4. Experimental Data Subsets

In the experiments we used sections 00 – 10 of PCEDT 2.0 as a development data and sections 11 – 19 for final evaluation of proposed methods. The development data were not only aimed to be an inspiration for rules' design but their English side was used for training the bunch of parameters, as well (see Section 5.2.).

5. Resolution in English

For the English part of our work we have developed some hand-written rules on gold data. On automatically analyzed data we have integrated the state-of-the-art system NADA and used it as our baseline. Then we have applied and extended the rules to improve it.

5.1. Experiments on Gold Data

The rules applied on gold data are based on the grammatical, surface and deep syntactic information. Therefore, they are able to detect the pleonastic *it* but they hardly capture non-anaphoric *it*, which commonly requires the wider context or out-of-text information.
Thanks to the tectogrammatical tree structure, the pleonastic *it* identification on gold data is quite simple. In contrast to the Czech task, we do not limit ourselves to the *it*-subjects only, because the corresponding Czech *ono / to*-object is always referential, whereas the English one can be also pleonastic. The proposed algorithm is as follows:
For all personal pronouns *it* having a verb as its parent, **if** one of the following conditions is true:

1. The verb is active and has a predicate of a subordinate subject clause annotated as its Actor.

2. The verb is passive and has a predicate of a subordinate subject clause annotated as its Patient.

3. The verb's lemma is *make* and got a predicate of a subordinate subject clause annotated as its Patient. It is the case of *make it (easy / hard/ etc.) to*.

Then it is a pleonastic instance.

5.2. Experiments on Automatically Analyzed Data

The results of resolving pleonastic *it* on gold data are quite high, but that is only a motivation to improve the deep syntactic parser. Therefore, we have experimented with the NADA system and some other rules on automatically analyzed data.

Rule-based system
Because of the unreliability of automatically annotated actants, we have to change the rules used on gold data. The approach works as follows:
For all personal pronouns *it*, **if** *it* has a verb as its parent **and** one of the following conditions is true:

1. The verb's lemma is *be / become/ make / take* and has an infinitive among its children.

   (17) *It* doesn't take much to provoke an intense debate.

2. The verb's lemma is *be* and there are a subject complement expressed as a predicate nominative or a predicate adjective and a subordinate clause.

   (18) *It* is easy to see why the ancient art is on the ropes.

   (19) *It*'s a shame their meeting never took place.

3. The verb is an active cognitive verb (*appear / follow / matter / mean / seem*) or a passive cognitive verb (believe / expect / note / recommend / say / think) and has a subordinate clause.

   (20) Before the sun sets on the '80s, *it* seems nothing will be left unhocked.

   (21) *It* can be said that the trend of financial improvement has been firmly set.

Then it is a pleonastic instance.
The condition 1 and 2 are further modified to prevent error cases, where *it* has been misannotated to be a child of other node than the verb in condition 1 or the subordinate clause is a subtree of the subject complement instead of the main predicate in condition 2.

NADA system
The NADA system (Bergsma and Yarowsky, 2011) is the state-of-the-art tool for anaphoricity determination of English *it*. Following the lexical and web count features, every occurrence of *it* is assigned a probability of being referential with a previously-mentioned entity. After having set the decision boundary (by default, it is 0.5), the occurrences can be binary classified as anaphoric and non-anaphoric.
The indisputable advantage of NADA is that the input does not have to be linguistically pre-processed at all, it accepts a surface text. Moreover, no linguistic analysis is being performed inside the tool. It makes NADA very simple and quick. On the other hand, if the rich linguistic annotation is available, it cannot exploit it.
As this software is freely available, we were able to integrate it into the Treex framework and combine the tree-oriented rules with the estimates produced by NADA.

Combination of NADA and rules
By combination of the statistical system working on a surface level and tree oriented hand-crafted rules we aimed to extract the best from both approaches. We decided to make a linear interpolation of the features, which consisted of every single rule in the previous approach, their disjunction and quantized values of NADA probability estimates. The parameters have been learnt from the development data using a maximum entropy classifier.[4]

[4]We employed the Perl module `AI::MaxEntropy`

5.3. Evaluation

As we stated in Section 5.2., NADA is a binary classifier distinguishing between anaphoric *it* and the other types. Since PCEDT 2.0 differentiate between 3 types of *it*, in order to successfully combine NADA with the designed rules two of these classes must be merged into one. We conducted experiments with 2 of 3 possible binarizations. The one with a merged class of anaphoric and non-anaphoric was left out as our central target is to be able to distinguish between these two classes.

The binarization with a joint class of non-anaphoric and pleonastic (NON-ANAPH+PLEO) as a class of positive instances accords with the way NADA was meant to be used. The overall results assessed in terms of accuracy as well as precision, recall and F-score measured on the positive class can be seen in Table 2.

NADA alone achieves a score similar to accuracy of 86% reported in (Bergsma and Yarowsky, 2011).[5] In comparison, relying just on the designed rules cannot compete with NADA, suffering mostly from a low coverage of the rules, reflected in a low value of recall. Even on the gold data the rules perform slightly worse mostly because they were tuned to describe just pleonastic occurrences. Combination of the statistical system and rules seemed to be promising. However, we register only a slight improvement of the success rate compared to NADA used separately.

The classes of anaphoric and non-anaphoric (mostly deictic and referring to a larger segment) *it* are alike in terms of referring to something, opposed to its pleonastic usage. Moreover, we constructed the rules to fit the class of pleonastic occurrences mainly, which suggests a better score than in case of the above-mentioned binarization. Following experiments are carried out with pleonastic *it* (PLEO) being a positive class.

The score of NADA alone in this configuration is surprisingly better, even though it was not supposed to be evaluated in this way. The values of precision and recall on a positive class changed, apparently due to changes in the distribution between positive and negative instances. As opposed to the previous configuration, the pure rule-based system outperforms NADA in accuracy here, also reaching a higher precision, which can be justified by the fact that the rules were tailored to recognize the pleonastic occurrences. The combination of both approaches results in the best accuracy of almost 90%, outperforming both of the components if used alone.

6. Resolution in Czech

Because of the Czech phenomena of subject absence, we attempt to identify the instances of predicates, to which a personal pronoun will be generated as a substitution of the unexpressed subject. First we apply hand-written rules on gold data, secondly the same rules in automatic data. Then the rules are improved and added by information from English automatic data (see Section 7.).

[5] Recall that NADA does not require any linguistic annotation, so it achieves the same score for the manually as well as the automatically analyzed data.

6.1. Experiments on Gold Data

Our heuristic procedure for identifying unexpressed implicit subject occurrences (anaphoric and non-anaphoric *it*) is based on constraints. We eliminate cases, where it is an overt subject, an unexpressed general subject or null subject. The procedure works as follows:

For all third person singular verbs, **if** all of the following conditions are true:

1. There is no overt subject, that is:
   (a) There is no overt subject represented by a word.
   (b) There is no subject subordinate clause.
2. There is no unexpressed general subject, that is:
   (a) The verb is not a part of the phrase *Je vidět / slyšet / cítit ((It) is seen / heard / felt)*.
   (b) The verb is not a part of the phrase *Lze / Je možné / Je nutné ((One) can / (It) is possible / (One) needs)*.
   (c) The verb is not a reflexive passive, because a third personal singular reflexive passim often determines a general subject.
   (d) The verb has no an *-o* ending, because the *-o* ending indicates a third personal neuter verb and it seems, a third personal neuter verb often implicates an instance of a general subject.
3. There is no null subject, that is:
   (a) The verb is not an impersonal (weather) verb *jednat se / pršet / zdát se / dařit se / oteplovat se / ochladit se / stát se / záležet (be about / rain / seem / do well / get warmer / get colder / happen / depend)*.
   (b) The verb is not a part of the phrase *Jde o ((It) is about)*.

Then there will be added a generated personal pronoun.

6.2. Experiments on Automatically Analyzed Data

The algorithm for anaphoric and non-anaphoric *it* identification on automatically analyzed data is extended by adding conditions to prevent errors that appear in the automatic annotation.

For all third person singular verbs, **if** all of the following conditions are true:

1. There is no overt subject, that is:
   (a) There is no overt subject represented by a word – *unchanged*.
   (b) There is no subject subordinate clause. The same condition on gold data was true, when the head of the subordinate clause was a finite verb having functor Actor. The new condition was true for finite verbs having functor Actor or Patient, because of the functor misannotation.

	NON-ANAPH+PLEO				PLEO			
	A	P	R	F	A	P	R	F
EN: Majority class	70.30	–	–	–	85.75	–	–	–
EN: Rules-gold	83.76	99.31	39.15	56.16	94.67	90.31	68.68	78.03
EN: Rules-autom	76.31	73.24	31.90	44.44	87.54	56.90	51.66	54.16
EN: NADA	83.86	81.10	59.51	68.65	86.19	51.00	78.01	61.68
EN: NADA + Rules-autom	84.44	78.61	65.40	71.40	89.83	71.88	47.06	56.88

Table 2: The results of evaluation of all tested systems, including two types of evaluation (NON-ANAPH+PLEO and PLEO). Quality of the systems was measured on the Evaluation data in terms of accuracy (A), precision (P), recall (R) and F-score (F). Majority class system corresponds to assigning a majority class to all candidates.

(c) If the verb is active, then it has no Actor among its children. This condition prevents errors in automatic subject annotation in the Czech part, where the overt subject was misannotated as other part-of-speech.

(d) If the verb is passive, then it has no Patient among its children (subject error prevention).

2. There is no unexpressed general subject – *unchanged*.

3. There is no null subject – *unchanged*.

Then there will be added a generated personal pronoun.

6.3. Evaluation

Contrary to the English task, where all personal pronouns *it* are presented on the surface sentence and we attempt to identify occurrences to be hidden on the tectogrammatical layer, the Czech target is detecting dropped third person singular pronouns in the subject position in order to express it on the tectogrammatical layer.
We use the binary classification of unexpressed third pronominal singular subject:

- referential – anaphoric and non-anaphoric dropped pronoun in the subject position having a generated node and being a child of the predicate.
- non-referential – pleonastic pronoun not being expressed either on the surface sentence or on the tectogrammatical layer.

There is another difference between the English task and the Czech task. Whereas a non-pleonastic pronoun for the English part means an anaphoric or non-anaphoric *it* only, a non-pleonastic pronoun for Czech is an anaphoric or non-anaphoric *he / she / it*. The reason lies on the gender differentiation of non-animal nouns and the use of gender differentiated pronouns to refer to them in Czech.
The rules on Czech data were implemented to suit the task: looking for a referential/implicit unexpressed subject and generating a tectogrammatical node for it. The scores of both systems are shown in Table 3.
Applying the rules on automatically analyzed data gives a perceptibly lower result than the rules on gold data. It is not surprising because on automatically analyzed data the overt subject is often misannotated as an object or other part-of-speech and vice versa. The subject subordinate clause is not straightforwardly recognizable, too.

7. Exploiting the Parallel Corpus

In the experiments so far, the proposed rules have employed just that language side of the corpus, which they were constructed for. We attempted to exploit the parallel nature of the PCEDT 2.0 corpus by designing rules that look also at the other side.
In general, information from the English side of automatically analyzed trees tends to be more reliable than the one from the Czech side. Particularly, it confirmed to be true for English rules, which used the Czech data. Such rules had no effect when they were combined with other rules for English.
On the other hand, in the opposite direction we designed the following rules:
For all third person singular verbs, **if** all of the following conditions is true:

1. The corresponding English verb has no non-pronominal subject. This condition prevents errors in automatic subject annotation in the Czech part, where the overt subject was misannotated as other part-of-speech.

2. There may be an unexpressed implicit subject, that is one of the following conditions is true:

   (a) Conditions 1 – 3 on automatically analyzed data are true.

   (b) The corresponding English verb has a *he / she* subject. This condition helps to detect cases, where the Czech conditions wrongly identified the existence of an overt subject. See error examples below:

   (22) Na noc se vrací do opuštěné
   At night RFLX returns to condemned
   budovy, kterou nazývá domovem.
   building, which **calls** home$_{ACT.error}$.
   'At night he returns to the condemned building **he calls** home.'

   (23) Banka First Union, říká,
   Bank$_{Sb-of-says.error}$ First Union, **says**,
   má nyní balíčky pro sedm skupin
   has now packages for seven groups
   zákazníků.
   of customers.

'First Union, **he says**, now has packages for seven customer groups.'

Then there will be added a generated personal pronoun.

These turned out to substantially contribute on the final quality of the whole rule-based system thanks to the information about English corresponding personal pronouns *he / she* that are expressed on the surface sentence and subjects, because the subject of an English clause can be also detected easier. Table 3 shows that if we include these inter-language rules, the accuracy increases by almost 3.5% absolute.

	ANAPH+NON-ANAPH			
	A	P	R	F
CZ: Majority class	86.58	–	–	–
CZ: Rules-gold	98.79	92.89	98.39	95.56
CZ: Rules-autom	87.68	52.97	73.34	61.51
CZ: Rules-autom+EN	91.08	64.20	75.87	69.55

Table 3: The results of evaluation of rule-based systems for Czech. Configuration "Rules-autom+EN" shows an impact of adding rules that use the English side

8. Conclusion

In this paper we have presented the annotation of personal pronoun *it* in the recently released Prague Czech-English Dependency Treebank 2.0. We have analyzed its occurrences in both languages and developed rule-based approaches to automatically identify the Czech and English *it* types. On the English side we also combined these tree-oriented rules with the statistical state-of-the-art system for this task, which improved the success rate on resolution of pleonastic occurrences.

Furthermore, we successfully exploited the parallel nature of the PCEDT 2.0 corpus and employed the English data in the task of Czech *it* identification.

In the future work, we plan to develop new rules and integrate machine learning methods in a greater extent. In addition, we would like to apply such system along with a coreference resolver to the much larger automatically analyzed parallel corpus CzEng 1.0 (Bojar et al., 2011). We hope the self-training on larger data together with a richer rule-/feature-set to increase the quality of coreference resolution.

9. Acknowledgments

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). This work has been supported by the Czech Science Foundation under the contract 201/09/H057 and by the grant GAUK 4226/2011. The authors would like to thank prof. Eva Hajičová, assoc. prof. Zdeněk Žabokrtský and the anonymous reviewers for their valuable comments and suggestions to improve the paper.

10. References

Shane Bergsma and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2011. Czeng 1.0.

José Guilherme Camargo de Souza and Constantin Orăsan. 2011. Can Projected Chains in Parallel Corpora Help Coreference Resolution? In *Anaphora Processing and Applications*, pages 59–69. Springer.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece, March. Association for Computational Linguistics.

Michael Denber. 1998. Automatic Resolution of Anaphora in English. Technical report, Eastman Kodak Co, Imaging Science Division.

Richard Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45 – 57.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2011. Prague Czech-English Dependency Treebank 2.0.

Souha Mezghani Hammami, Rahma Sallemi, and Lamia Hadrich Belguith. 2010. A bayesian classifier for the identification of non-referential pronouns in arabic. In *In Proceedings of the 7th International Conference on Informatics and Systems- INFOS 2010*, pages 1–6.

Sandra M. Harabagiu and Steven J. Maiorano. 1999. Knowledge-Lean Coreference Resolution and its Relation to Textual Cohesion and Coherence. In *The Relation of Discourse/Dialog Structure and Reference*.

Lynette Hirschman. 1997. MUC-7 Coreference Task Definition.

Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, and Oliver Čulo. 2003. Anotování koreference v pražském závislostním korpusu. Technical Report TR-2003-19, ÚFAL MFF UK, Prague, Prague.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Comput. Linguist.*, 20(4):535–561, dec.

José Carlos Clemente Litrán, Kenji Satou, and Kentaro Torisawa. 2004. Improving the identification of non-anaphoric it using support vector machines. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 58–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proceedings of the Twelfth EAMT Conference*, pages 102–111.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver.

Ruslan Mitkov and Catalina Barbu. 2002. Using bilingual corpora to improve pronoun resolution. *Languages in Contrast*, 4(2):201–211.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.

Giang Linh Ngụy and Magda Ševčíková. 2011. Unstated Subject Identification in Czech. In *WDS'11 Proceedings of Contributed Papers, Part I*, pages 149–154.

Chris D. Paice and Gareth D. Husk. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun *it*. *Computer Speech and Language*, 2.

Jarmila Panevová. 1991. Koreference gramatická nebo textová? In *Etudes de linguistique romane et slave*. Krakow.

Jesús Peral, Manuel Palomar, and Antonio Ferrández. 1999. Coreference-oriented interlingual slot structure machine translation. In *In Proceedings of the ACL Workshop Coreference and its Applications*, pages 69–76.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Petr Sgall, Alla Goralčíková, Eva Hajičová, and Ladislav Nebeský. 1967. *Generativní popis jazyka a česká deklinace*. Prague:Academia.

Petr Sgall. 1969. *A Functional approach to syntax in generative description of language*. American Elsevier Pub. Co.

John M. Sinclair. 1995. *English Grammar*. Harper Collins Publisher, UK.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, pages 67–74, Stroudsburg, PA. Association for Computational Linguistics.

Michael Swan. 1995. *Practical English Usage*. Oxford University Press, UK.

Zdeněk Žabokrtský. 2011. Treex – an open-source framework for natural language processing. In *Information Technologies – Applications and Theory*, pages 7–14.

CLIR- and ontology-based approach for bilingual extraction of comparable documents

Manuela Yapomo[1], Gloria Corpas[2], Ruslan Mitkov[3]
[1]Evaluations and Language resources Distribution Agency (ELDA)
[2]University of Malaga
[3]University of Wolverhampton
manuyap@yahoo.fr, gcorpas@uma.es, R.Mitkov@wlv.ac.uk

Abstract

The exploitation of comparable corpora has proven to be a valuable alternative to rare parallel corpora in various Natural Language Processing tasks. Therefore many researchers have stressed the need for large quantities of such corpora and the scarcity of works on their compilation. This paper describes a CLIR-based method for automatic extraction of French-English comparable documents. At the start of the process, source documents are translated and most representative terms are extracted. The resulting keyword list is further enlarged with synonyms on the assumption that keyword expansion might improve the retrieval of such documents. Retrieval is performed on the indexed target collection and a further filtering step based mainly on temporal information and document length takes place. Preliminary results suggest that the employment of ontology could improve the performance of the system.

Keywords: Comparable documents, comparable corpora; Cross-Language Information Retrieval (CLIR); ontology; similarity measurement.

1. Introduction and Previous Work

Comparable corpora are referred to as collections of documents in the same or in different languages made up of similar texts. Using snippets of several definitions, Skadina, et al. (2010a, p.7) came up with a more elaborate description which is the following: "a collection of similar documents that are collected according to a set of criteria, e.g. the same proportions of texts of the same genre in the same domain from the same period (McEnery and Xiao, 2007) in more than one language or variety of languages (EAGLES, 1996) that contain overlapping information (Munteanu and Marcu, 2005; Hewavitharana and Vogel, 2008)".

The present work focusing on the collection of comparable documents discusses the development of a tool based on cross-language retrieval which given an input of source collection, outputs a target collection of the 'most comparable' texts to the given source documents. This tool is cross-lingual in its nature as the source and target collections can be in two different languages. In this particular project, we have experimented with English and French.

Comparable corpora have enjoyed an increasing importance in recent years as their exploitation was found to be a productive alternative to parallel corpora in several fields of Natural Language Processing (NLP) and beyond. Several works on terminology extraction (Gamallo, 2007; Saralegi, San Vicente and Gurrutxaga, 2008), Machine Translation (MT) (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009), Cross-Language Information Retrieval (CLIR) (Talvensaari et al., 2007), etc. relying on comparable corpora provide empirical evidence for this view. They play an important role for translation and terminology as well (Bowker and Corpas, forthcoming).

Comparable documents are traditionally acquired from the web or from existing research corpora and different approaches have been proposed to perform this task. To mine English-German-Spanish comparable documents from the Internet, Talvensaari et al. (2008) employ focused crawling. Domain specific vocabulary is collected separately in all three languages and used to acquire relevant seed URLs. The selected URLs are then employed in the crawling phase to identify relevant pages from which text paragraphs are extracted. Leturia, San Vicente and Saralegi (2009) present a search engine-based approach for acquiring specialised Basque-English comparable corpora from the web. The tool takes as input a mini-corpus from which most relevant words are extracted and used as seeds to retrieve relevant web pages. Relying on two newspaper subcorpora, Bekavac et al. (2004) describe the collection of Bulgarian-Croatian comparable documents by mapping common vocabulary and publication dates in documents of the two corpora. Talvensaari et al. (2007) introduce the CLIR-based approach in gathering comparable Swedish-English documents from two newspaper collections. They extract good keys with RAFT (Relative Average Term Frequency). The resulting keys are translated and ran against the target collection with Lemur retrieval system (www.lemurproject.org).

Our work takes the CLIR-based approach further. In this study, we perform ontology based-query expansion thus exploiting the synonymy relation in WordNet with a view to achieving better efficiency in the retrieval procedure. This novel approach is applied to the bilingual compilation of comparable documents in English and French. The general idea of our methodology is, given K source documents and M target documents, to extract the N (<=M) target documents most comparable to the source documents. Applying this methodology in an incremental fashion would be the basis of compiling comparable corpora.

The paper is organised as follows: Section 2 describes our methodology and outlines the system architecture. Section 3 reports the evaluation results obtained so far with regard to the performance of the system. Finally, section 4 offers concluding remarks.

2. Methodology and Architecture of the System

The source documents are first translated into the target language. They then undergo preprocessing prior to keyword extraction. The list of keywords obtained is further expanded with synonyms. After the phases of document translation, keyword extraction and expansion, document retrieval and filtering are undertaken. The pipeline of the system is illustrated in Figure 1:

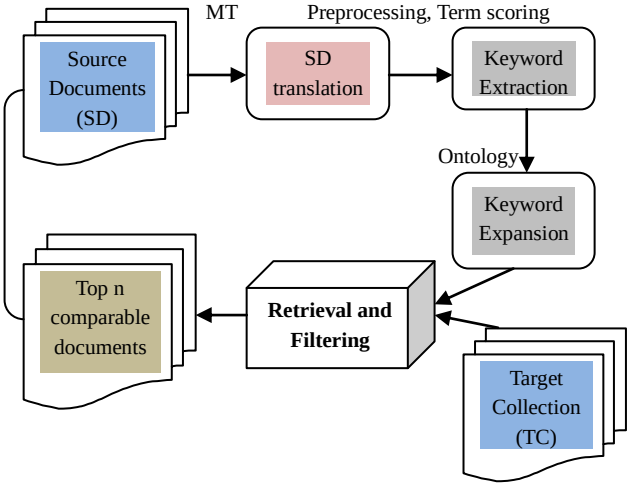


Figure 1: General architecture of the system

2.1 Document Translation

Cross-language retrieval research so far has exploited either dictionary translation (Pirkola et al., 2001) or Machine Translation (Huang et al., 2010). Each translation approach has its advantages and disadvantages. For queries -which are list of words,- dictionary translation appears to be more appropriate. In multilingual dictionaries however, words carry usually more than one translation, and thus ambiguity is carried over to the target language.

In general, MT usually produces a better translation than dictionary-based translation as syntax and other factors are usually taken into account (depending on the MT system). As a result, there is less ambiguity in a translation performed by an MT system. However, the performance of an MT system may not always be of acceptable quality. In general, there is consensus that MT is more suitable for document translation than for keywords translation. However, as in dictionaries, OOV (Out Of Vocabulary) words are encountered with MT tools which also often miss domain-specific terminology. In this work we employ MT based on the premise that it works better for document translation and helps avoiding the problem of ambiguity occurring with dictionaries. Microsoft Translator has been selected as an MT system for this study. The output of the MT system is subject to further processing, namely keywords extraction.

2.2 Keyword Extraction

Prior to performing keywords extraction, the system performs (i) preprocessing of data and (ii) term weighting. Preprocessing in the present study consists in lemmatisation and POS-tagging using the TreeTagger (Schmid, 1994), a tool for annotating texts with part-of-speech and lemma information. Lemmatisation is performed to transform inflected forms into their base forms. POS-tagging is a better alternative to stop words removal as only content words, which are nouns, proper nouns, adjectives and verbs are taken into account. Lemmatisation is a further advantage for languages such as French, which has a rich flexive system. It helps avoiding incorrect count of a term frequency for words which have more than 1 part-of-speech tag. POS-tagging is also helpful in decreasing ambiguity of multi-category words in WordNet.

The next step of term weighting consists in assigning a relevance value to content-bearing words in the source collection. A number of approaches have been proposed to this end. They can be grouped as supervised and unsupervised methods. Supervised methods involve machine learning (Zhang et al., 2006). They are quite stable but demand much effort, since training annotated corpus and a classifier are required. In this work, unsupervised methods are preferred to supervised ones. Following this approach, several formulae have been proposed.

Word frequency or term frequency (TF) was introduced by Luhn (1957) but is quite basic. More robust term weighting methods are preferable. Matsuo and Ishizuka (2004) used word co-occurrence to identify keywords from a unique document. TF-IDF is a standard relevance measure used in several studies (Ramos, 2003; Li, Fan and Zhang, 2007). A limitation of TF-IDF is that it does not necessarily show the goodness of relevant keys that may occur just once or twice in some important documents. Furthermore, the collection should be large enough to yield a reliable IDF. Since our source documents meet the previous requirement for IDF, we will adopt TF-IDF as relevance measure in this work.

After weight is assigned to all the content bearing words in our source documents set, we can move on to keyword extraction. This will be done by selecting the top n keys with higher TF-IDF values. We can proceed to keyword expansion, which we believe might increase the performance of the system.

2.3 Keyword Expansion

Keyword expansion consists in enlarging a keyword list. This is done by adding to the list of initial keywords, words with which they share some semantic relations. Approaches to keyword expansion are based on

probabilistic and ontology-based methods. Probabilistic query expansion consists in extracting terms that are most related to query keys based on co-occurrences of terms in documents. The ontology-based method, on the other hand, makes use of semantic relations already established in ontologies to select terms. In this work, we are interested in this latter approach to keywords expansion. We exploit synonymy in Wordnet (Miller et al., 1993).
How to expand queries automatically is not a trivial task because one has to avoid the problem of ambiguity. When integrating WordNet in our system, we attempt to resolve this problem by POS-tagging our source collection. In this way, the POS-tag could help discarding other categories of a polysemous word. In other to further reduce ambiguity, we will select only the first synset (synonym set) of a word. The choice of the first synset is quite simplistic but will work in most cases for it is the most general sense. We also limit ourselves to the two first lemma-names of the first synset in other to avoid proliferation of keywords.

2.4 Retrieval and Filtering

Document retrieval can be referred to as the matching of some query against a collection of texts with the purpose of obtaining documents relevant to the query only. In line with the definition of comparable corpora in section 1, not only similarity of target documents to the query will be taken into account but also temporal information and size of related documents in our objective to retrieve comparable documents .
In this work, the Opensource toolkit Indri is used to carry out the retrieval process. Indri is part of the Lemur project. Prior to document retrieval, all the target documents were indexed with Lemur. Date normalisation is equally performed according to a specific date format understandable by Indri toolkit. After indexing, proper retrieval can be undertaken. To do filtering based on extralinguistic criteria (date of publication and document length), the corresponding feature-intervals should be defined so as to select only documents that meet the filtering constraints adopted. Since this tool should work with any linguistic data, time span will be extracted from the source documents to ensure that all filtered documents fall within the same time-period and have a text-length ranging from 1,000 to 50,000 characters. This interval is mainly chosen to filter out too small and too large documents.

3. Evaluation

In this part of the paper, we first describe the data that will be used for tests. Experiments and results are then reported with observations.

3.1 Data

To carry out experiments, we use two sets of source and target documents made up of news articles, randomly collected from different news websites.
Our source collection contains 38 selected articles in French. The criteria to meet when selecting the texts are that they should be about the same or closely related topic. The total number of words contained in our source set is of 25,047 with an average number of 659 words in each document. The domain of selected documents was economy and they were all more or less related to the topic of "2008 economic crisis" Documents were taken from news websites lemonde.fr, lepoint.fr, etc.

As regards the target document set we selected 280 which we classified. We opted for a modified version of Braschler and Schäuble (1998)'s relevance scheme as comparability metric for annotation and evaluation purposes. Table 1 illustrates our modification of Braschler and Schäuble's relevance scale:

Classes in this study	**Equivalent classes according to Braschler and Schäuble (1998)**	**Comments**
Class 1	(1) Same story	The two documents deal with the same event.
Class 2	(2) Related story	The two documents deal with the same event or topic from a slightly different viewpoint. Alternatively, the other document may concern the same event or topic, but the topic is only a part of a broader story or the article is comprised of multiple stories.
Class 3	(4) Common terminology	The events or topics are not directly related, but the documents share a considerable amount of terminology.
Class 4	(5) Unrelated	The similarities between the documents are slight or nonexistent.

Table 1: Modification of Braschler and Schäuble 's guidelines for classifying target documents

Our modification of Braschler and Schäuble's scheme consists in the deletion of the third class (shared aspects) on the grounds that named entities are not taken into account in our study. Retrieved documents belonging to Class 1 and 2 are considered good alignments whereas retrieval of documents from class 3 and 4 is not.
To classify documents at hand, precisions were added as regards the theme of the documents collection for our experiments:

(1) *Same story* in this context contains texts that are about *the Great Recession*. This includes texts

about causes, manifestations and effects; descriptive, explanatory texts, etc.

(2) *Related story* involves documents reporting financial crisis. It includes articles about financial crises in general or specific ones, different from that of the first category. Examples are *the Great Depression* or *Inflation in Zimbabwe.*

(3) *Common terminology* comprises documents sharing vocabulary. These are documents which are about finances in general.

The documents collected were distributed in each class as illustrated in Table 2 below:

Collection	# of documents	Class	Time Span
Source set (Fr)	38	Class 1	2007 – 2011
Target set (En) (280)	69	Class 1	2007 – 2011
	63	Class 2	No date and size restriction
	81	Class 3	
	67	Class 4	

Table 2: Description of source and target data

3.2 Experiments

We evaluated the performance of our tool on the data described in the previous section. To achieve the retrieval of comparable documents, we had to extract keywords from a translation of source documents using TF-IDF. We further exploited WordNet to enlarge the keyword list with synonyms. The resulting translated keys were used as queries and run against the target language data with Lemur retrieval system. Date of publication and size are used to further filter out less relevant documents.

Experiments were carried out with different configurations to find out which one gives the best results. Different options were tried at the levels of (i) keyword extraction and (ii) keyword expansion. Our experiments can be split in two groups. The purpose of our first group of experiments was to determine which portion of most relevant keys (k) was to be used for retrieval. We carried out experiments with k=10, k=15 and k=20 respectively. Keyword extraction performed with average success. Among the extracted keys, good ones perfectly matching the topic were *recession, subprime*. Relatively good keys were *bankruptcy, mortgage, price, lending, bank*. Many irrelevant keys such as *institution, country, recover, down* were extracted which would negatively affect retrieval. Relevant words such as *crisis, economy, deflation*, etc were not extracted.

In the second set of experiments, we tested the effect of WordNet as described in section 2.3. After expansion of keywords lists k=10, k=15 and k=20, we respectively obtained the following expanded lists k1=14, k2=24 and k3=31 terms. Most of the words in the initial keyword list did not find synonyms in WordNet and most of those that were assigned synonyms were not good keys. Some are *institution (establishment), country (state, land), recover* (*regain, find*).

In the two different groups of experiments, time span and size are used to further filter out documents. As mentioned in section 2.4, temporal information is extracted from source data if available and a size interval of 1,000 to 50,000 characters of texts always applies.

3.3 Results

To carry out evaluation of the efficiency of the system designed, we analyse results of retrieval carried out in the two sets of experiments described in the previous section.

Table 3 shows results of retrieval using different sets of significant terms.

	k=10		k=15		k=20	
	#	%	#	%	#	%
Class 1	25	35,7	21	30	18	25,7
Class 2	11	15,7	23	32,8	15	21,4
Class 3	32	45,7	26	37,1	29	41,4
Class 4	2	2,8	0	00	8	11,4
Total	70	100	70	100	70	100

Table 3: Results of retrieval with different sets of relevant keys

The shaded areas in Table 3 and Table 4 below show the best retrieval performances for classes 1 and 2. Results of retrieval show that most of the documents retrieved belong to class 3. This can be explained by the fact that keys extracted are very general words in the semantic field of finance.

Few documents of the second class were retrieved contrarily to documents of the third class which are less comparable. This may be due to the presence of very general words in the keywords list. Around 30% of retrieved documents fall within class 1. We can observe than the first and second sets of keywords, k=10 and k=15 perform better for retrieval of class 1 documents. The second set of keys (k=15) allows retrieval of the largest amount of documents in class 2.

Table 4 shows results of retrieval with the same set of words as those in Table 3 with the difference that keywords are now expanded with synonyms in WordNet.

	k1=14		k2=24		k3=31	
	#	%	#	%	#	%
Class 1	20	28,5	21	30	15	21,4
Class 2	13	18,5	24	34,2	12	17,1
Class 3	33	47,1	23	32,8	36	51,1
Class 4	4	5,7	2	2,8	7	10
Total	70	100	70	100	70	100

Table 4: Results of retrieval with different sets of relevant keys and WordNet

With keyword expansion, retrieval appears to be less efficient for documents of class 1. Similarly to the previous group of experiments, more documents from the third class are extracted. The experiment with k2 performs best. Indeed, with this scheme, fewer documents from the third class are extracted and more documents from the second class are obtained.

Though we cannot formulate general conclusions based on these results from our small set of data, we observe that the best results were obtained using the top 15 keys with synonyms in WordNet. WordNet therefore seems to have a positive impact on the retrieval.

4. Conclusion

This work describes a bilingual approach for extracting comparable documents to a specific set of documents. Given K source documents, the N (<=M) most comparable documents to the source documents are extracted from an M target set. Applying this methodology in an incremental fashion would be the basis of compiling comparable corpora.

Our work takes the CLIR-based approach further. In this study we perform ontology-based query expansion of the most relevant terms thus exploiting the synonymy relation in WordNet with a view to achieving better efficiency in the retrieval procedure. The evaluation of the tool that we developed shows that the best results obtained are after expanding a set to 24 keywords.

5. References

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of Comparable Corpora to improve SMT performance. *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens , pp.16–23.

Bekavac, B., Osenova, P., Simov, K. and Tadić, M. (2004). Making monolingual corpora comparable: a case study of Bulgarian and Croatian. *Proceedings of the 4th Language Resources and Evaluation Conference: LREC04*, Lisbon, pp. 1187-1190.

Bowker, L. and Corpas, G. Translation Technology. In Mitkov, R. *The Oxford Handbook of Computational Linguistics*. Second, substantially revised edition. Oxford Unversity Press,

Braschler, M. and Schäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*. Berlin: Springer- Verlag, pp.183–197.

Gamallo, P. (2007). Learning bilingual lexicons from comparable English and Spanish corpora. *Proceedings of Machine Translation Summit XI*, Copenhagen, pp. 191-198.

Huang, D., Zhao, L., Li, L. and Yu, H. (2010). Mining large-scale comparable corpora from Chinese-English news collections. *Proceedings of the 22th International Conference on Computational Linguistics:* Coling 2010, Beijing, August 2010, pp. 472–480;

Leturia, I., San Vicente, I. and Saralegi, X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the internet. *5th International Web as Corpus (WAC5)*. Donostia-San Sebastian.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller K. (1993). Introduction to WordNet: An on-line lexical database. Cambridge: MIT Press.

Munteanu, D. and Marcu, D. (2005). Improving Machine Translation performance by Exploiting non-parallel corpora. *Journal Computational Linguistics, 31(4)*. Cambridge: MIT Press, pp.477-504.

Pirkola, A., Hedlund, T., Keskustalo, H. and Järvelin, K. (2001). Dictionary-based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval, 4(3-4)*, pp.209-230.

Saralegi, X., San Vicente, I. and Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *Proceedings of the Workshop on Comparable Corpora, LREC'08*, Basque Country, pp.27-32.

Skadina, I. et al. (2010a). *Analysis and evaluation of comparable corpora for under resourced areas of Machine Translation. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora.* European Language Resources Association (ELRA), La Valletta, Malta, pp.6-1.

Schmid, H. (1994). Part-of-Speech tagging with Neural Networks. *Proceedings of the 15th International Conference on Computational Linguistics: COLING-94*.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M. and Keskustalo, H. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval 11*, pp.427-445.

_____et al. (2007). Creating and exploiting a comparable corpus in Cross-Language Information Retrieval. *ACM Transactions on Information Systems, 25(1).*

[1] **www.lemurproject.org** (Accessed February 17, 2012).

ICA for Bilingual Lexicon Extraction from Comparable Corpora

Amir HAZEM and Emmanuel MORIN

Laboratoire d'Informatique de Nantes-Atlantique (LINA)
Université de Nantes, 44322 Nantes Cedex 3, France
Amir.Hazem@univ-nantes.fr, Emmanuel.Morin@univ-nantes.fr

Abstract

Independent component analysis (ICA) is a statistical method used to discover hidden features from a set of measurements or observed data so that the sources are maximally independent. This paper reports the first results on using ICA for the task of bilingual lexicon extraction from comparable corpora. We introduce two representations of data using ICA. The first one is called global ICA (GICA) used to design a global representation of a context according to all the target entries of the bilingual lexicon, the second one is called local ICA (LICA) and is used to capture local information according to target bilingual lexicon entries that only appear in the context vector of the candidate to translate. Then, we merge both GICA and LICA to obtain our final model (GLICA). The experiments are conducted on two different corpora. The French-English specialised corpus 'breast cancer' of 1 million words and the French-English general corpus 'Le Monde / New-York Times' of 10 million words. We show that the empirical results obtained with GLICA are competitive with the standard approach traditionally dedicated to this task.

1. Introduction

The use of comparable corpora for the task of bilingual lexicon extraction has received great interest since the beginning of 1990. It was introduced by Rapp (1995) as an alternative to the inconvenience of parallel corpora, which are not always available and are also difficult to collect especially for language pairs not involving English and for specific domains, despite many previous efforts in compiling parallel corpora (Church and Mercer, 1993). According to Rapp (1995, p320): *<...The availability of a large enough parallel corpus in a specific field and for a given pair of languages will always be the exception, not the rule.>*

The standard approach proposed by Rapp (1995) for aligning words from comparable corpora is, without doubt, the gold standard and the main state of the art in this domain based on a word space model. Words are represented by context vectors in high dimensional vector spaces by using distributional statistics. Contextual information has been widely used in statistical analysis of natural language corpora (Deerwester et al., 1990), (Honkela et al., 1995), (Ritter and Kohonen, 1989). Words are represented by the contexts in which they occur. This representation is motivated by the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts. Many investigations and a number of studies have emerged, (Fung, 1995; Fung, 1998; Fung and Lo, 1998; Peters and Picchi, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Gaussier et al., 2004; Morin et al., 2007; Laroche and Langlais, 2010, among others).

Word space models, are not specific to bilingual lexicon extraction. Considerable attention is given to it in current research on semantic indexing (Sahlgren and Karlgren, 2005). Many different applications use word space models, including information retrieval (Dumais et al., 1988), word sense disambiguation (Schütze, 1992), (Hanson et al., 1993), various semantic knowledge tests (Lund et al., 1995), (Karlgren and Sahlgren, 2001), and text categorisation (Sahlgren and Coster, 2004).

In the standard word space methodology, for bilingual lexicon extraction from comparable corpora, each word is represented by its context vector for both source and target languages. For a word to be translated in the source language, its context vector is first translated using a bilingual lexicon, then, a similarity measure is used between the translated context vector and all the target context vectors. Finally, The target words are ranked according to their similarity scores. It is worth noticing that context vectors which are the basis of the word space model, may contain information redundancy, and suffer from data sparseness. We believe that a better representation of context vectors, by using a subspace in which vectors are orthogonal and data is maximally independent, should provide a better representation of data and thus reach a better accuracy for word alignment. In this paper, we propose to apply the independent component analysis (ICA) transform, which is basically an extension of the principal component analysis (PCA) transform. Both have proven their efficiency in data representation in many fields such as face recognition, data compression, etc. The remainder of this paper is organised as follows: Section 2. presents the standard approach based on lexical context vectors dedicated to word alignment from comparable corpora. Section 3. describes ICA technique. Section 4. describes our approach. Section 5. describes the different linguistic resources used in our experiments. Section 6. evaluates the contribution of the standard and ICA approaches to the quality of bilingual terminology extraction through different experiments. Section 7. presents our discussion and finally, Section 8. presents our conclusion and some perspectives.

2. Standard Approach

The main work in bilingual lexicon extraction from comparable corpora is based on lexical context analysis and relies on the simple observation that a word and its translation tend to appear in the same lexical contexts. The basis of this observation consists in the identification of first-order affinities for each source and target language: "*First-order affinities describe what other words are likely to be*

found in the immediate vicinity of a given word" (Grefenstette, 1994a, p. 279). These affinities can be represented by context vectors, and each vector element represents a word which occurs within the window of the word to be translated (for instance a seven-word window approximates syntactical dependencies).
The implementation of this approach can be carried out by applying the following four steps (Rapp, 1995; Fung and McKeown, 1997):

Context Characterisation

Let us denote, by $\mathbf{i}$ the context vector of the word i [1]. All the words in the context of each word i are collected, and their frequency in a window of n words around i extracted. For each word i of the source and the target languages, we obtain a context vector $\mathbf{i}$ where each entry $\mathbf{i}_j$, of the vector is given by a function of the co-occurrences of words j and i. Usually, association measures such as mutual information (Fano, 1961) or the log-likelihood (Dunning, 1993) are used to define vector entries.

Vector Transfer

The words of the context vector $\mathbf{i}$ are translated using a bilingual dictionary. Whenever the bilingual dictionary provides several translations for a word, all the entries are considered but weighted according to their frequency in the target language. Words with no entry in the dictionary are discarded.

Target Language Vector Matching

A similarity measure, $\mathrm{sim}(\bar{\mathbf{i}}, \mathbf{t})$, is used to score each word, t, in the target language with respect to the translated context vector, $\bar{\mathbf{i}}$. Usual measures of vector similarity include the cosine similarity (Salton and Lesk, 1968) or the weighted Jaccard index (WJ) (Grefenstette, 1994b) for instance.

Candidate Translation

The candidate translations of a word are the target words ranked following the similarity score.

The translation of the words of the context vectors, which depends on the coverage of the bilingual dictionary vis-à-vis the corpus, is an important step of the standard approach; as more elements of the context vector are translated, the context vector will be more discriminating in selecting translations in the target language. This drawback can be partially circumvented by combining a general bilingual dictionary with a specialised bilingual dictionary or a multilingual thesaurus (Chiao and Zweigenbaum, 2003; Déjean et al., 2002). Moreover, this approach is sensitive to the choice of parameters such as the size of the context, the choice of the association and similarity measures. The most complete study about the influence of these parameters on the quality of bilingual alignment has been carried out by Laroche and Langlais (2010).

[1] Generally, bold lower case letters indicate vectors and bold upper case letters indicate matrices.

3. Independent Component Analysis

In the classic version of the linear ICA model (Jutten and Hérault, 1991), (Comon, 1994), (Hyvarinen et al., 2001), each observed random $x = (x_1, x_2, ..., x_n)^T$ is represented as a weighted sum of independent random variables $s = (s_1, ..., s_k, ..., x_n)^T$, such as:

$$x = As \tag{1}$$

where A is the mixing matrix that contains the weights which are assumed to be different for each observed variable and s is the vector of the independent components. If we denote the columns of matrix A by a_i the model can be written as:

$$x = \sum_{i=1}^{D} a_i s_i \tag{2}$$

The statistical model in equation 1 is called the ICA model which describes how the observed data are generated by a process of mixing the components s_i. Both the mixing matrix A and the independent components s are learned in an unsupervised manner from the observed data x.
The starting point for ICA is the assumption that the components s_i are statistically independent. ICA can be seen as an extension to principal component analysis (PCA) and factor analysis. The main difference between ICA and PCA is, while PCA finds projections which have maximum variance, ICA finds projections which are maximally non-Gaussian. PCA is useful as a pre-processing technique that can reduce the dimension of the data with minimum mean-squares error. In contrast, the purpose of ICA is not dimension reduction. For our analysis we applied the FastICA (Hyvarinen, 1999) algorithm where the data matrix X is considered to be a linear combination of independent components:

$$X = AS \tag{3}$$

where columns of S contain the independent components and A is a linear mixing matrix. The dimension of the data was first reduced by PCA in order to decorrelate the data, to reduce over-learning and to get the square mixing matrix A. After variance normalisation (the whitened data), n independent components which create a feature representation in the component space were extracted with ICA.

4. Method

Our method consists in building a discriminating subspace using ICA which represents a double interest. Indeed, the mathematical properties of ICA ensure a better data representation, and using PCA as a pre- processing step, provides a dimension reduction which can be very useful when using large comparable corpora.

Data Representation

In our case, the observed data x is an N-by-N word-word matrix where columns represent contexts and rows represent words. The N words of the target language that appear in the bilingual dictionary are retained for constructing matrix X. Each column of X represents a context vector of a word i with $i \in N$. For a given element X_{cr} of matrix X, X_{cr} denotes the association measure of the r:th analysed word with the c:th context word. The chosen association measures are mutual information and the log likelihood.

GICA Representation

Data representation in GICA consists in building a whole component space s that represents a global view of words in the target corpus. Each component s_k encodes some interesting features extracted from the N target words. Here, we can analyse how the positions of the words in the target language are related according to the general representation of data which gives a global view of the distribution of words by considering contexts of all the words of the corpus that appear in the bilingual lexicon.

LICA Representation

Data representation in LICA consists in building a partial component space s that represents a local view of words in the target corpus according to the translated context vector of the candidate. Each component s_k encodes some interesting features extracted from the M target words that are part of the translated context vector of the candidate. Here, we can analyse how the positions of the words in the target language are related according to the partial representation of data by considering only the contexts of the candidate. The aim of this specific representation is to capture information related to the candidate only. This can be seen as a local or a specific representation.

For each method GICA and LICA, we use the same context characterisation and vector transfer in the same way that the standard approach. Context vectors of source and target words are computed and the words of the context vector of the candidate are translated using a bilingual dictionary. The main difference of our method resides in building a new vector space using ICA that transforms matrix X into a new component space $s = (s_1, ..., s_k, ..., x_n)^T$. Matrix X can be seen as the concatenation of N context vectors of the target words that appear in the bilingual lexicon.

4.1. Words Projection

Once the new component space s is built, The translated context vector of the candidate and all the context vectors of the target words are projected into the new subspace.

Let us denote $\mathbf{i}$ a context vector of a given word i. The projection of the context vector of i in the new subspace and noted $\mathbf{i}_p$ is shown in equation 4.

$$\mathbf{i}_p = \mathbf{i}^T \times S \tag{4}$$

4.2. Distance Measure

As in the standard approach, the candidate translations of a word are the target words ranked following the similarity score or dissimilarities (proximities). Here we only deal with dissimilarity that can often be understood as distance. We use a normalised Euclidean distance also called Chord distance (Korenius et al., 2006) as shown in equation 5.

$$d(\mathbf{i}, \mathbf{j}) = \sqrt{\sum_{k=1}^{n} (\frac{\mathbf{i_k}}{||\mathbf{i}||} - \frac{\mathbf{j_k}}{||\mathbf{j}||})^2} \tag{5}$$

4.3. GLICA Model

Let us denote $d_{GL}(i,j)$, ($d_G(i,j)$ and $d_L(i,j)$), the GLICA, GICA and LICA distances. GLICA is merely a weighted sum of GICA and LICA as given by the following equation:

$$d_{GL}(i,j) = \lambda \times d_G(i,j) + (1-\lambda) \times d_L(i,j) \tag{6}$$

Although the representation of GLICA is simple, it is important to highlight the fact that this model retains only candidates that appear in both GICA and LICA. That is to say, all the target words that are not present in the local or the global independent component space are discarded.

5. Linguistic Resources

The experiments have been carried out on two different French-English corpora: a specialised corpus from the medical domain within the sub-domain of 'breast cancer' and a general corpus from newspapers 'LeMonde/New-York Times '. Due to the small size of the specialised corpus we wanted to conduct additional experiments on a large corpus to have a better idea of the behaviour of our approach. Both corpora have been normalised through the following linguistic pre-processing steps: tokenization, part-of-speech tagging, and lemmatisation. The function words have been removed and the words occurring less than twice (i.e. hapax) in the French and the English parts have been discarded.

5.1. Specialised Corpus

We have selected the documents from the Elsevier website[2] in order to obtain a French-English specialised comparable corpus. We have automatically selected the documents published between 2001 and 2008 where the title or the keywords contain the term 'cancer du sein' in French and 'breast cancer' in English. We collected 130 documents in French and 118 in English and about 530,000 words for each language. The comparable corpus comprised about 7,400 distinct words in French and 8,200 in English.

In bilingual terminology extraction from specialised comparable corpora, the terminology reference list required to evaluate the performance of the alignment programs is often composed of 100 single-word terms (SWTs) (180 SWTs in (Déjean and Gaussier, 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002), and 100 SWTs in (Daille and Morin, 2005)). To build our reference list, we selected 400 French/English SWTs from the UMLS[3] meta-thesaurus and the *Grand dictionnaire terminologique*[4]. We kept only

[2] www.elsevier.com

[3] www.nlm.nih.gov/research/umls

[4] www.granddictionnaire.com/

the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 122 French/English SWTs were extracted.

5.2. General Corpus

We chose newspapers as they offer a large amount of data. We selected the documents from the French newspaper 'Le Monde' and the English newspaper 'The New-York Times '. We automatically selected the documents published between 2004 and 2007 and obtained 5 million words for each language. The comparable corpus comprised about 41,390 distinct words in French and 44,311 in English.
The terminology reference list is much more consequential and contains 500 SWTs. It has been extracted from ELRA-M0033 randomly.

5.3. Bilingual Dictionary

The French-English bilingual dictionary required for the translation phase was the ELRA-M0033 dictionary. It contains, after projection in the 'breast cancer 'corpus and linguistic pre-processing steps, 3600 English single words and 3550 french single. And contains after projection in the corpus 'Le Monde/New-York Times' and linguistic pre-processing steps, 17.100 English single words and 16600 french single words belonging to the general language.

6. Experiments and Results

In this section, we first give the parameters of the standard and ICA based approaches, than we present the results conducted on the two corpora presented above: 'Breast cancer' and 'LeMonde/New-YorkTimes'.

6.1. Experimental Setup

Three major parameters need to be set to the standard approach and the ICA based approaches (LICA, GICA and GLICA), namely the similarity measure, the association measure defining the entry vectors and the size of the window used to build the context vectors. Laroche and Langlais (2010) carried out a complete study of the influence of these parameters on the quality of bilingual alignment. As a similarity measure, we chose to use the Cosine (Salton and Lesk, 1968) and the Weighted Jaccard Index (Grefenstette, 1994b) for the standard approach, while for ICA approaches, we chose the Euclidean distance which is the standard measure for PCA and ICA transforms. The entries of the context vectors were determined by the mutual information (Fano, 1961) and the log-likelihood (Dunning, 1993), and we used a seven-word window since it approximates syntactic dependencies. Other combinations of parameters were assessed but the previous parameters turned out to give the best performance.

6.2. Evaluation on the Breast Cancer Corpus

We investigated the performance of the standard approach (SA) and ICA based approaches (GICA, LICA and GLICA) on the 'Breast Cancer' corpus, using the evaluation list of 122 words.

We evaluate the accuracy by using the term : "top k" which means that the correct translation was found in the first k words presented by a given approach.

Evaluation Using Mutual Information

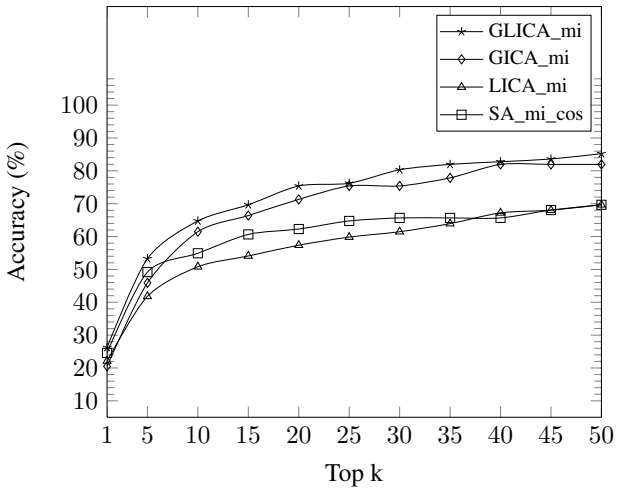


Figure 1: Accuracy at top k for the breast cancer corpus using mutual information.

We can see in Figure 1 that GLICA_mi approach always outperforms the standard approach for all values of k. The accuracy at the top 20 for SA_mi_cos is 62.29% while GLICA_mi approach gives 75.40%. We can also notice that GICA_mi outperforms SA_mi_cos from $k = 5$. Even if LICA_mi is almost always under the other approaches, according to Figure 1, it remains useful for GLICA.

Evaluation Using Log-Likelihood

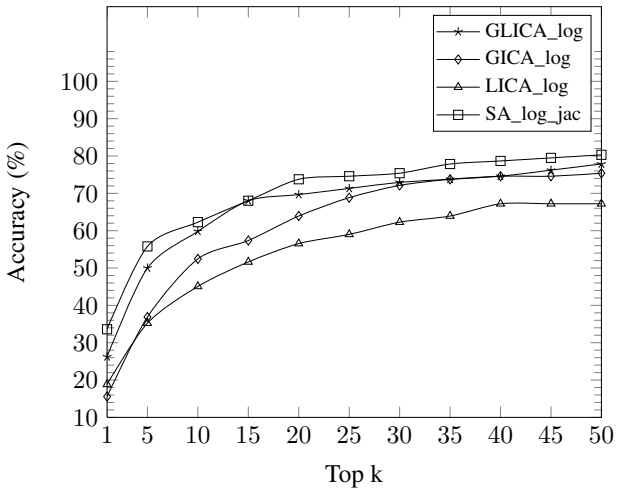


Figure 2: Accuracy at top k for the breast cancer corpus using log-likelihood.

We can see in Figure 2 that GLICA_log approach is under the standard approach for almost all values of k (except at $k = 15$). The accuracy at the top 20 for SA_log_jac is 73.77% while GLICA_mi approach gives 69.67%. Both, GICA_log and LICA_log are also under the baseline.

According to Figure 1 and Figure 2, we can notice that the best configuration for the standard approach is SA_log_jac with an accuracy of 73.77% for the top 20, while for our approach, the best configuration is GLICA_mi with an accuracy of 75.40% for the top 20. It is worth to notice that the merging process of the local and the global ICA plays an important role for improving the accuracy of our final model GLICA.

Evaluation on the best configuration of the Standard and GLICA approaches

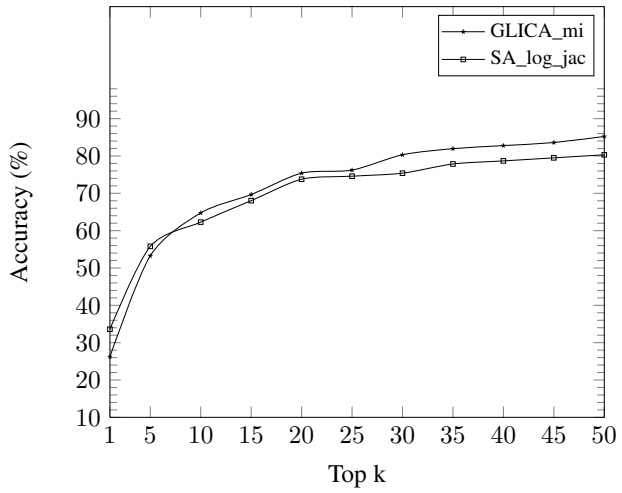


Figure 3: Accuracy at top k for the breast cancer corpus using the best parameters configuration of the standard and GLICA approaches.

Figure 3 presents the best performance of the standard and GLICA approaches. We can see that our approach outperforms the standard approach from $k > 5$. GLICA_mi reaches an accuracy of 64.75% at $k = 10$ and 75.40% at $k = 20$ while the standard approaches reaches an accuracy of 62.29% at $k = 10$ and 73.77% at $k = 20$. We can also notice that the standard approach outperforms our approach for both $k = 1$ and $k = 5$. GLICA_mi reaches an accuracy of 26.22% at $k = 1$ and 53.27% at $k = 5$ while the standard approach reaches an accuracy of 33.60% at $k = 1$ and 55.79% at $k = 5$.

Evaluation of the GLICA approach according to λ

Figure 4 shows how the GLICA (GLICA_mi) approach can be sensitive to the variations of the parameter λ. It seems that our approach is more accurate for $0.5 < \lambda < 0.9$ which means that the merging process gives more importance to the global ICA (GICA) than to the local ICA (LICA).

Evaluation on the LeMonde/New-YorkTimes Corpus

We then investigate the performance of the standard approach (SA) and ICA based approaches (GICA, LICA and GLICA) on 'LeMonde/New-YorkTimes' corpus, using an evaluation list of 500 words.

Evaluation Using Mutual Information

We can see in Figure 5 that GICA_mi LICA_mi and GLICA_mi approaches always outperform the standard approach for all values of k. The accuracy for the top 20

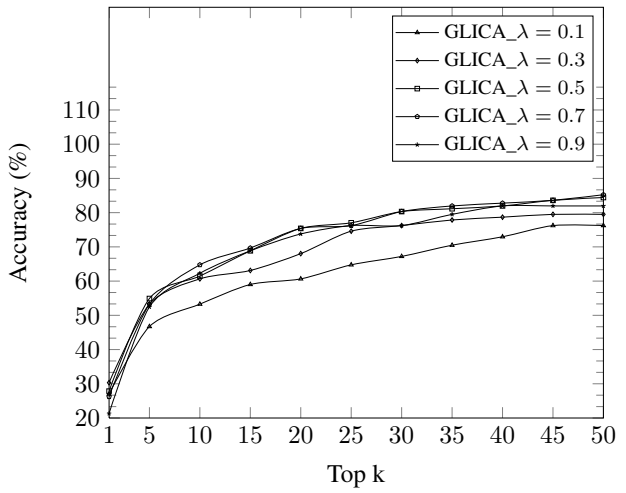


Figure 4: Accuracy at top k for the breast cancer corpus according to λ.

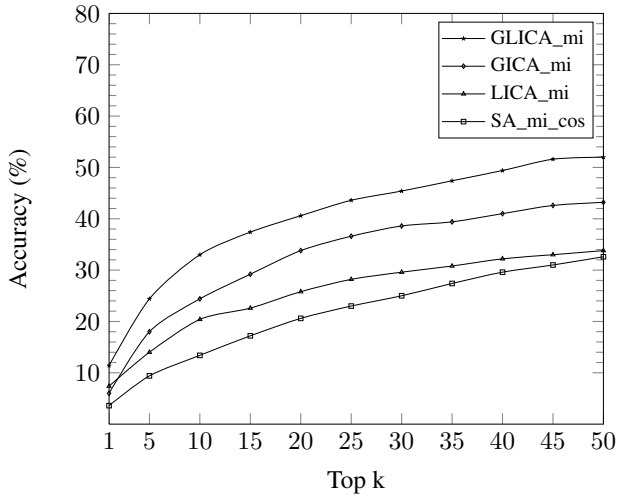


Figure 5: Accuracy at top k for LeMonde/NewYorkTimes using mutual information

for SA_mi_cos is 20.6% while GICA_mi approach gives 33.8%, LICA_mi approach gives 25.8% and GLICA_mi approach gives 40.6%. According to Figure 5 All the ICA models outperform the standard approach for this configuration (using mutual information as the association measure).

Evaluation Using Log-Likelihood

We can see in Figure 6 that the GLICA_log is slightly better than the standard approach. The accuracy for the top 20 for SA_log_jac is 38.8% while GLICA_mi approach gives 39.4%. Both, GICA_log and LICA_log are under the baseline.

According to Figure 5 and Figure 6, we can notice that the best configuration for the standard approach is SA_log_jac with an accuracy of 38.8% at the top 20, while for our approach, the best configuration is GLICA_mi with an accuracy of 40.6% at the top 20. It is also interesting to notice that GLICA_log outperforms SA_log_jac with an accuracy of 39.4% for $k = 20$.

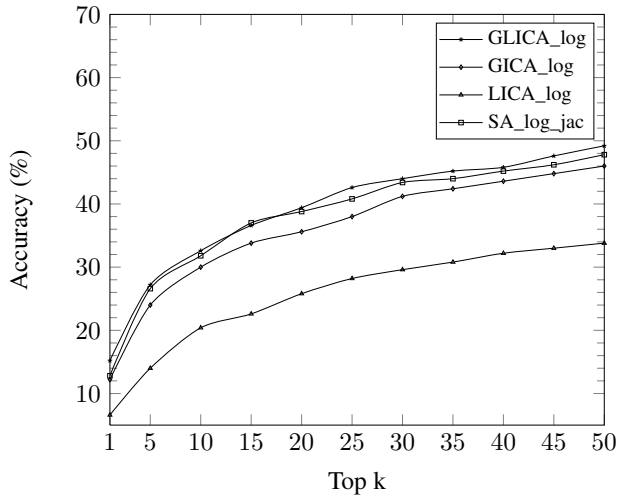


Figure 6: Accuracy at top k for LeMonde/NewYorkTimes using log-likelihood

Evaluation on the best configuration of the Standard and GLICA approachs

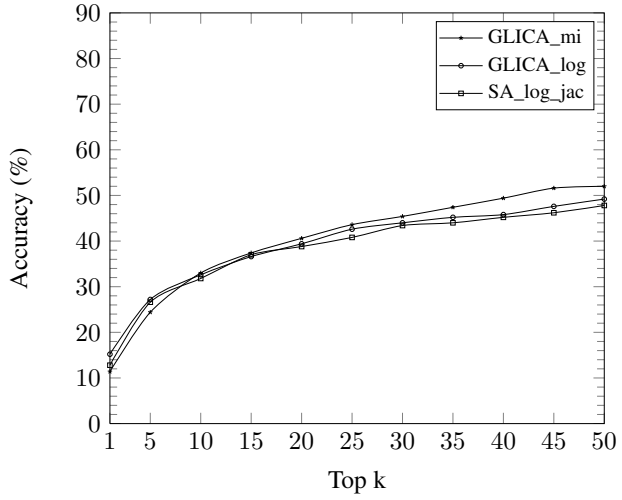


Figure 7: Accuracy at top k for LeMonde/NewYorkTimes corpus using the best parameters configuration of the standard and GLICA approaches.

Figure 7 presents the best performance of the standard and GLICA approaches. We can see that our approach outperforms the standard approach from $k > 5$. GLICA_mi reaches an accuracy of 33% at $k = 10$ and 40.6% at $k = 20$ while the standard approach reaches an accuracy of 31.8% at $k = 10$ and 38.8% at $k = 20$. We can also notice that the standard approach outperforms our approach for both $k = 1$ and $k = 5$. GLICA_mi reaches an accuracy of 11.4% at $k = 1$ and 24.4% at $k = 5$ while the standard approach reaches an accuracy of 12.8% at $k = 1$ and 26.6% at $k = 5$. On the contrary, GLICA_log outperforms the standard approach for both $k = 1$ with an accuracy of 15.2% and $k = 5$ with an accuracy of 27.2%.

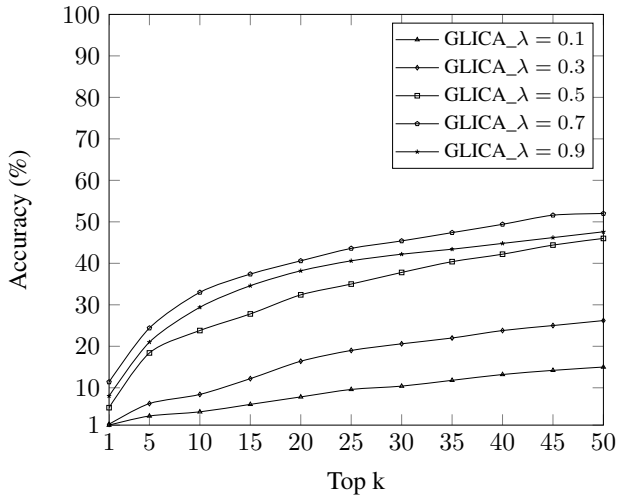


Figure 8: Accuracy at top k for LeMonde/NewYorkTimes according to λ.

Evaluation of the GLICA approach according to λ

Figure 8 shows how the GLICA (GLICA_mi) approach can be sensitive to the variations of the parameter λ. It seems that our approach is more accurate for $0.7 < \lambda < 0.9$ which means that the merging process gives more importance to the global ICA (GICA) than to the local ICA (LICA).

7. Discussion

The purpose of our experiments was to compare the proposed method with the baseline not only according to the best parameters configuration of each method, but also, in terms of behaviour according to the two main association measures that have proven their efficiency in thise domain (Rapp, 1999), and by choosing two different comparable corpora, a domain specific and a general one. The main interest of using two different comparable corpora is to test and validate our method according to the size and the type of the corpus.

For the 'breast cancer' corpus, the experiments based on mutual information, have shown that GLICA_mi and GICA_mi outperform SA_mi_cos while LICA_mi is slightly under SA_mi_cos. On the contrary, the use of the log-likelihood on the same corpus have shown that SA_log_jac outperforms LICA_log , GICA_log and GLICA_log. For the best configuration of each method, GLICA_mi shows better results than SA_log_jac. We can conclude from this first set of experiments on the breast cancer corpus that the standard approach reaches its best accuracy with log-likelihood while GLICA reaches its best performance with mutual information and for the best configuration of each method, GLICA_mi outperforms Sa_log_jac (except for $k = 1$ and $k = 5$).

For the 'LeMonde/New-YorkTimes' corpus, the results have also shown that GLICA_mi, GICA_mi and LICA_mi outperform SA_mi_cos. And that GLICA_log outperforms SA_log_jac while LICA_log, GICA_log were under the baseline (SA_log_jac). For the best configuration, GLICA_mi outperforms SA_log_jac (except for $k = 1$ and $k = 5$). This second set of experiments allows us to confirm

that both ICA-based methods and the standard method have the same behaviour on two different comparable corpora, and that the best association measure for the standard approach is the log-likelihood while for the ICA-based methods mutual information performs better.

According to the results stated previously, it is rightful to try to understand the reasons why GLICA accuracy is better using mutual information than log-likelihood on the 'Breast Cancer' corpus, while conversely, GLICA_log performs better than GLICA_mi on the 'LeMonde/New-YorkTimes' corpus for $k = 1$ and $k = 5$. Is it a matter of corpus size? or is it a matter of data representation? Further experiments need to be conducted in this direction.

In the GLICA approach, the parameter λ was fixed at 0.7, which means that we gave an advantage to GICA in the merging process. In fact, it was not our aim in this paper to deal with the parameter λ. We believe that in an appropriate environment, with an optimal data representation for both local and global component spaces, λ should be fixed at 0.5, so we consider GICA and LICA with the same importance. It is our hope for future work to carry out an in-depth study on this parameter, in addition to other merging techniques other than the one used for GLICA.

The GLICA method shows two advantages : (1) it is a merger of GICA which captures global context information of words, and LICA which captures local context information. Thus, GLICA has both global and local views on context representation. (2) Thanks to PCA pre-processing, GLICA offers a dimension reduction which enables a faster computation. As a comparison, the context vector size of a given word in the standard approach varies between the frequency of the word to its frequency multiplied by the size of the context window, which can easily reach thousands of words for frequent words and hundreds for less frequent words. For GLICA, the size of the context vectors in the ICA subspace is fixed to one hundred, it is independent from word frequency.

Finally, GLICA can be considered as promising for future work. The GLICA model does not take into account any linguistic or semantic information, it is just based on bag of words context. Many improvements need to be done especially for context representation.

8. Conclusion

In this paper, we have described and compared two techniques which focus on bilingual lexicon extraction from comparable corpora. The standard method considered as the state of the art and our method based on independent component analysis transform. This work represents, to the best of our knowledge, the first application of ICA to the task of bilingual lexicon extraction from comparable corpora. We have shown that a GLICA-based model can significantly outperform the standard approach model, for both the specialised and the general comparable corpora. The fact that our GLICA-based model outperforms the standard approach indicates that independent component analysis deserves more attention and can be considered as an alternative to the standard approach. It is our hope that this work will encourage further exploration of the potential of ICA modeling within alignment based on comparable corpora.

9. Acknowledgement

The research leading to these results has received funding from the French National Research Agency under grant ANR-08-CORD-013 and from the European Communitys Seventh Framework Programme (*/*FP7/2007-2013*/*) under Grant Agreement no 248005.

10. References

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. IOS Press.

Kenneth Ward Church and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

P. Comon. 1994. Independent component analysisa new concept? *Signal Processing*, 36:287314.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS*, pages 281–285. ACM.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

Pascale Fung and Yuen Yee Lo. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, pages 414–420.

Pascale Fung and Kathleen McKeown. 1997. Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, pages 192–202, Hong Kong.

Pascale Fung. 1995. Compiling Bilingual Lexicon Entries From a non-Parallel English-Chinese Corpus. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'95)*, pages 1–16, Langhorne, PA, USA.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From ParallelCorpora to Non-parallel Corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Éric Gaussier, Jean-Michel Renders, Irena Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Gregory Grefenstette. 1994a. Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, pages 279–290, Amsterdam, The Netherlands.

Gregory Grefenstette. 1994b. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles, editors. 1993. *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*. Morgan Kaufmann.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. 1995. Contextual relations of words in grimm tales analyzed by self-organizing map. In *ICANN*, pages 3–7.

A. Hyvarinen, J. Karhunen, and E Oja. 2001. Independent component analysis. *New York:a John Wiley Sons.*

A. Hyvarinen. 1999. Fast and robust fixed-point algorithms for independent component a analysis. *IEEE Transactions on Neural Networks*, 10(3):626634.

C Jutten and J. Hérault. 1991. Blind separation of sources. part i. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:110.

J. Karlgren and M. Sahlgren. 2001. From words to understanding. In *Foundations of Real-World Intelligence*, pages 294–308.

Tuomo Korenius, Jorma Laurikkala, Martti Juhola, and Kalervo Järvelin. 2006. Hierarchical clustering of a finnish newspaper article collection with graded relevance assessments. *Inf. Retr.*, 9(1):33–53.

Audrey Laroche and Philippe Langlais. 2010. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Kevin Lund, Curt Burgess, and Ruth Ann Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Carol Peters and Eugenio Picchi. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Gregory Grefenstette, editor, *Cross-language information retrieval*, chapter 7, pages 81–90. Kluwer Academic Publishers.

Reinhard Rapp. 1995. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

Helge Ritter and Teuvo Kohonen. 1989. Self-organizing semantic maps. *biological Cybernetics*, 4(64):241–254.

M. Sahlgren and R Coster. 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING, August 23-27, Geneva, Switzerland*, pages 487–493.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341.

Gerard Salton and Michael E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.

Hinrich Schütze. 1992. Word space. In *NIPS*, pages 895–902.

Improving Compositional Translation with Comparable Corpora

Hiroyuki Kaji, Takashi Tsunakawa, Yoshihiro Komatsubara
Department of Computer Science, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu-shi, 432-8011, Japan
E-mail: {kaji, tuna}@inf.shizuoka.ac.jp, gs10017@s.inf.shizuoka.ac.jp

Abstract

We improved the compositional term translation method by using comparable corpora. A bilingual lexicon consisting of pairs of word sequences within terms and their correlations is derived from a bilingual document-aligned corpus. Then, for an input term, compositional translations are produced together with their confidence scores by consulting the corpus-derived bilingual lexicon. Thus, we can select the correct translation for the input term from among as many candidate ones as possible. An experiment with a comparable corpus of Japanese and English scientific-paper abstracts demonstrated that compositional translation using the corpus-derived bilingual lexicon outperforms that using an ordinary bilingual lexicon. Future work includes the incremental improvement of the bilingual lexicon with correlations, the refinement of the confidence score, and the extension of the compositional translation model to allow word order to be changed.

Keywords: term translation, comparable corpus, bilingual lexicon

1. Introduction

Technical term translation is one of the key issues in document translation as well as crosslingual information retrieval. Obviously, no existing bilingual lexicon covers all of the terms in a domain. However, most technical terms are compound words and 88% of Japanese technical terms in some domains have compositional English translations (Tonoike, et al. 2006). Thus, the compositional translation method plays an essential role in translating technical terms.

The performance of the compositional translation method naturally depends on the bilingual lexicon it consults. It cannot produce a correct translation for a term unless the lexicon provides appropriate translations for the constituent words of the term. At the same time, it is difficult to select a correct translation for the term from among many candidate translations produced compositionally when the bilingual lexicon provides as many translations as possible for each of the constituent words. It should be noted that the latter problem may become more serious if we improved the coverage of the bilingual lexicon to overcome the former problem.

We propose improving the compositional translation method by using a bilingual corpus. A wide-coverage bilingual lexicon, which consists of word sequence pairs in two languages together with their correlations, is acquired from a bilingual corpus. Then, a ranked list of translations is produced for an input term by compositionally generating candidate translations together with their confidence scores based on the correlations between the constituent words and their translations. Our contribution is not bilingual lexicon acquisition from a bilingual corpus but an improved compositional translation method with confidence scores.

Our proposed framework is compatible with both parallel and comparable corpora. Parallel corpora generally produce bilingual lexicons with more reliable correlations than comparable corpora (Och and Ney 2003; Koehn et al. 2003). However, there are few domains in which large parallel corpora are available. Therefore, we assume that the input corpus is a comparable corpus, more specifically a document-aligned corpus. Use of weakly comparable corpora, which are much more widely available but may produce bilingual lexicons with less reliable correlations, is beyond the scope of this paper (Fung and Yee 1998; Rapp 1999; Andrade et al. 2010; Ismail and Manandhar 2010; Morin and Prochasson 2011).

There have been many studies on bilingual lexicon acquisition from parallel or comparable corpora, where the task is usually to find translations for terms occurring in the input corpus. Bilingual lexicon acquisition methods have usually been evaluated in terms of recall and precision of target language translations acquired for source language terms occurring in the input corpus (Fung and Yee 1998; Rapp 1999; Cao and Li 2002; Tanaka 2002). In contrast, our task is to translate a term even when it does not occur in the input corpus; therefore, we evaluated our framework in terms of precision of translations produced for a test set of input terms collected independently of the input corpus. This task setting is natural when we assume practical applications of bilingual lexicons such as document translation and crosslingual information retrieval.

2. Problems and our framework

Consider the pair of a Japanese term “光通信<*HIKARI TSUUSHIN*>” and its English translation “optical communication.” We humans can recognize the correspondence between “光<*HIKARI*>” and “optical” as well as that between “通信<*TSUUSHIN*>” and “communication.” In other words, the translation from “光通信” to “optical communication” is compositional. However, few electronic Japanese-English lexicons provide the correspondence between a Japanese noun, e.g.,

"光," and an English adjective, e.g., "optical." Therefore, the automatic compositional translation method usually fails to produce the correct translation "optical communication" for the input term "光通信."

Assume that a pair of a Japanese noun "光" and an English adjective "optical" has been registered in a bilingual lexicon. It would provide possible translations, such as "light," "ray," and "beam," as well as "optical" for "光." Likewise, it would provide possible translations, such as "communication," "correspondence," and "report," for "通信." Thus, the compositional translation method may produce many candidate translations including "optical communication," "optical correspondence," "optical report," "light communication," "light correspondence," and others from which it must select the correct one.

As exemplified above, the compositional translation method exhibits two problems, incomplete bilingual lexicons and many candidate translations most of which are spurious. To overcome these problems, we propose a framework consisting of the following two steps: (1) acquiring a bilingual lexicon with correlations from a bilingual corpus, and (2) producing compositional translations together with confidence scores.

(1) Acquiring a bilingual lexicon with correlations from a bilingual corpus

We assume that a comparable corpus consisting of pairs of relevant documents is available and we use the method for calculating pairwise correlations between words in two languages based on co-occurrence statistics in aligned sentences (Matsumoto and Utsuro 2000). This method, which is originally intended for parallel corpora, is applicable to comparable corpora by treating document pairs as sentence pairs (Utsuro et al. 2003). It seems workable as long as the documents are small. It has an advantage in that it does not require a seed bilingual lexicon unlike other methods applicable to comparable corpora.

Our purpose was to construct a wide-coverage bilingual lexicon of term constituents rather than the actual terms. Most of the correspondences between constituents are those between simple words, e.g., "光" and "optical," but some are those between a simple word and a compound word, e.g., "薄膜<*HAKUMAKU*>" and "thin film," and vice versa, e.g., "移動体<*IDOU TAI*>" and "mobile." Therefore, we need to extract not only pairs of simple words but also mixed pairs of simple and compound words. However, it is not necessarily easy to identify compound words. Moreover, from a practical point of view, it is preferable that the bilingual lexicon provides possible translations for any word sequence included in a term; translation pairs of longer word sequences would increase the possibility of correct translations being produced for a term. Therefore, we consider any word sequence included in a term as its constituent and calculate pairwise correlations between those in the source and target languages.

(2) Producing compositional translations together with confidence scores

To select a correct translation from among many candidate translations produced compositionally, we calculate a confidence score for each of the candidate translations. Note that constituent translation pairs have been acquired together with their correlations. We regard the correlations as the confidence scores for the constituent translations and define the confidence score for a compositional translation based on the scores for its constituent translations.

As mentioned in Step 1, the bilingual lexicon provides translations not only for a word but also for a word sequence. However, their correlations or confidence scores are not so reliable. Therefore, we re-evaluate the translations the bilingual lexicon provides for a word sequence: namely, we produce compositional translations for a word sequence even when it is included in the bilingual lexicon and combine the two confidence scores, one provided by the bilingual lexicon and the other calculated compositionally.

The following two sections describe the two steps of our framework in some detail, where the source and target languages are assumed as Japanese and English, respectively. Our framework can be applied to any language pairs with some modifications in language-specific issues such as treatment of morphology.

3. Acquiring bilingual lexicon for compositional translation

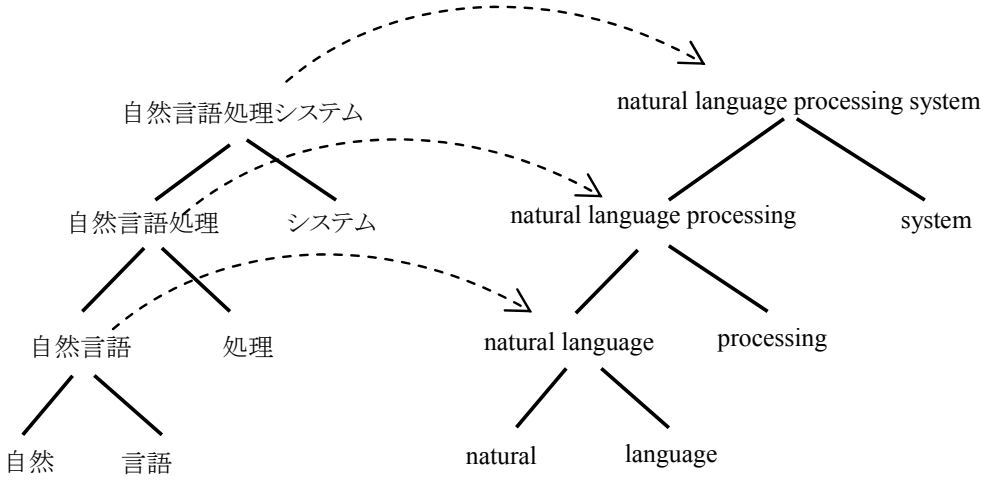
We extract every word sequence included in terms from both Japanese and English documents. Most Japanese terms are Noun+, i.e., sequences of one or more nouns, and most English terms are Adjective*Noun+, i.e., sequences of one or more nouns optionally preceded by one or more adjectives, where adjectives include present participles and past participles of verbs. At present, we do not deal with terms with more complicated structures, e.g., those including prepositional phrases. Therefore, we extract every Japanese word sequence consisting of nouns and every English word sequence consisting of nouns and adjectives.

We define the correlation of a Japanese word sequence J and an English word sequence E by using Dice's coefficient. That is,

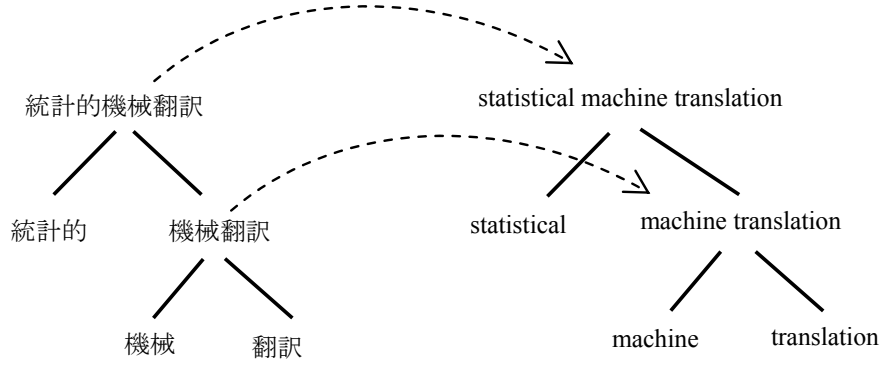
$$C(J,E) = \frac{2 \cdot g(J,E)}{f(J) + f(E)}, \qquad [1]$$

where $f(J)$ and $f(E)$ are the number of Japanese documents and that of English documents in which J and E occur, respectively, and $g(J, E)$ is the number of pairs of Japanese and English documents in which J and E co-occur.

We ignore the frequencies of word sequences occurring in a document. This is because we intend to apply our framework to nonparallel corpora: the frequency of a Japanese word sequence occurring in a Japanese document is not necessarily comparable to that of the corresponding English word sequence occurring in



(a) Example 1



(b) Example 2

Fig. 1: Structure of terms and compositional translation

the English document aligned with the Japanese document. We also ignore the lengths of word sequences because they are not necessarily maintained across languages, as exemplified by the pairs (移動体<*IDOU TAI*>, mobile) and (薄膜<*HAKUMAKU*>, thin film).

It should be noted that we distinguish between maximal word sequences, which are not subsequences of longer word sequences, and non-maximal word sequences. Japanese maximal and non-maximal word sequences tend to correspond to English maximal and non-maximal word sequences, respectively, in a pair of aligned documents. Accordingly, a document pair is counted as 0.5 for a pair of maximal and non-maximal word sequences co-occurring in the document pair, while it is counted as 1.0 for a pair of maximal word sequences co-occurring in it as well as for a pair of non-maximal word sequences co-occurring in it. Assume that "光通信" and "optical communication" co-occur as maximal word sequences in a pair of aligned documents. This document pair is counted as 0.5 for pairs (光, optical communication), (通信, optical communication), (光通信, optical), and (光通信, communication), while it is counted as 1.0 for pairs (光通信, optical communication), (光, optical), and (通信, communication) (Note that it is also counted as 1.0 for incorrect pairs (光, communication) and (通信, optical)). Thus, we reduce the confusion between a compound word and its constituent words.

Since the correlations are unreliable for a word sequence infrequently occurring in the input corpus, we set a threshold θ_f for the number of documents in which a word sequence occurs. We calculate correlations for every pair of Japanese and English word sequences both of which occur in θ_f or more documents. Since we intend to translate Japanese terms into English, we select the top N_1 English word sequences in descending order of correlation for each Japanese word sequence (In the experiment described in Sec. 5, we set θ_f and N_1 to 10 and 20, respectively.).

4. Compositional translation with confidence score

Note that a term can be represented with a binary tree according to its head-modifier relations, as exemplified in Fig. 1. We assume that Japanese term J can be compositionally translated into English term E if and only if J and E are isomorphic or represented with the same

Input term: $jw_1^N (= jw_1 jw_2 \cdots jw_N)$

Bilingual lexicon: L

Triangular matrix: $A(i,j)$ (i=1,…,N; j=i,…,N)

Algorithm:

1) For i:=1 to N do

For j:=i to N do

$A(i,j) \leftarrow \phi$.

2) For i:=1 to N do

$A(i,i) \leftarrow \{\langle ew, S(jw_i, ew)\rangle \mid (jw_i, ew) \in L\}$

3) For r:=1 to N–1 do

For i:=1 to N–r do

For k:=i to i+r–1 do

$$A(i,j) \leftarrow A(i,j) \cup \left\{ \langle e_1 e_2, S(jw_i^j, e_1 e_2)\rangle \mid \begin{matrix} \langle e_1, S(jw_i^k, e_1)\rangle \in A(i,k), \\ \langle e_2, S(jw_{k+1}^j, e_2)\rangle \in A(k+1, j) \end{matrix} \right\}$$

Fig. 2: Compositional translation algorithm

binary tree. Based on this assumption, we define the confidence score $S(J,E)$ for the compositional translation from a Japanese term or word sequence J to an English word sequence E as follows:

$$S(J,E) = \begin{cases} \lambda \cdot S'(J,E) + (1-\lambda) \cdot C(J,E) \\ \qquad (|J| \geq 2, |E| \geq 2) \\ C(J,E) \\ \qquad (\min\{|J|, |E|\} = 1) \end{cases}, \quad [2]$$

where $S'(J,E)$ is a confidence score based on compositionality, $C(J,E)$ is the correlation based on co-occurrence in pairs of aligned documents, λ is a parameter adjusting the weights for $S'(J,E)$ and $C(J,E)$, and $|J|$ and $|E|$ denote the lengths of word sequences J and E, respectively.

We define the confidence score based on compositionality as follows:

$$S'(J,E) = \max_{\substack{1 \leq i < p \\ 1 \leq j < q}} \frac{2 \cdot S(jw_1^i, ew_1^j) \cdot S(jw_{i+1}^p, ew_{j+1}^q)}{S(jw_1^i, ew_1^j) + S(jw_{i+1}^p, ew_{j+1}^q)}, \quad [3]$$

where $J = jw_1 jw_2 \cdots jw_p (= jw_1^p)$ and $E = ew_1 ew_2 \cdots ew_q (= ew_1^q)$. This formula is based on the following idea. We define the confidence score based on compositionality as the harmonic mean of the confidence scores for the two constituent translations. However, we do not know the correct structures for J and for E. Therefore, we calculate the confidence score for every combination of possible decompositions of J and E and select the maximum confidence scores based on the assumption that the combination of correct structures maximizes the confidence score.

Formula [3] shows that we assume the coincidence of word order between a Japanese term and its English translation. This is generally not the case. It is not difficult to modify the formula to deal with the change in word order. Moreover, this formula does not contain the factor representing the compatibility of the two constituent translations. It should be noted that correlation $C(J,E)$ reflects to some extent the compatibility of the constituent translations.

Next, we describe a dynamic programming algorithm for compositionally producing translations. It is similar to the CKY parsing algorithm for context-free grammars, as shown in Fig. 2. It constructs a triangular matrix $A(i,j)$ consisting of cells each of which corresponds to a subsequence jw_i^j in the input term and contains translation candidates and their confidence scores for its corresponding subsequence. To prevent combinatorial explosion, we restrict candidate translations contained in each cell to those with N_2 highest confidence scores (In the experiment described in Sec. 5, we set N_2 to 100.).

5. Experiment

5.1 Experimental settings

本報告では，振幅スペクトルからの音声合成法を利用した音質劣化の少ないピッチ変換法を提案する。本方式ではまず，サンプリング変換を用いて周波数スケーリングを行なうことでピッチ周波数を変更する。その後，サンプリング変換した音声に対して，振幅スペクトル系列のスペクトル包絡特性を原音声の特性に復元する。最後に，変更された振幅スペクトル系列から，原音声の音声速度と等しくなるように音声を合成する。本方式で得られるピッチ変換音声は，変換倍率が約0．8倍から2．0倍の範囲では，かなり原音声の音質を保存している。	This paper proposes a method of pitch modification using the speech synthesis method from short‐time Fourier transform (STFT) magnitude. The method modifies first the pitch frequency by frequency scaling using sampling rate conversion. For the speech whose sampling rate is converted, spectral envelopes of STFTs magnitude are restored to the ones of original speech. Finally, a speech is synthesized from the modified STFT magnitude but the frame shift rate for synthesis is set so that the synthesis speech rate equals to the original one. The resulting pitch modified speech can preserve very well the quality of original speech over the range 0.8‐2.0 of rate conversion.

(a) Almost parallel

鳥沢のHPSGパージングアルゴリズムは，HPSGの辞書項目からコンパイルされたCFG（文脈自由文法）を用いるフェーズ1と，それだけではカバーしきれない制約を素性構造を用いて計算するフェーズ2からなる。本稿ではフェーズ1の並列化アルゴリズムを提案した。超並列計算機AP1000＋上で並列オブジェクト指向言語ABCL／fを用いて実装した。新聞を例題として50語以下の文（平均19語）をパージングし，構文木をすべて数え上げるのに要した時間は一文当たり98ミリ秒であった。	This paper describes an attempt to develop a parallel parsing algorithm for Torisawa's parsing algorithm for HPSG. Torisawa's algorithm consists of two phases. At Phase 1, a parser enumerates possible parse trees using CFG rules compiled from lexical entries in HPSG. The constraints uncovered by the CFG are solved at Phase 2, using feature structures and a variant of unification, partial unification. We realized a parallel parsing algorithm for Phase 1, on a highly parallel computer AP1000+ (256 Super Sparc 50Mhz) with concurrent object‐oriented programming language ABCL/f. The average parsing time for the sentences consisting of less than 50 words was 98msec.

(b) Totally comparable but organized differently

1台のカメラとターンテーブルを用い，さまざまな角度から撮影した物体の2次元画像から3次元形状を構築する手法を開発した。ターンテーブルの分割角度 θ 毎に仰角 ϕ で対象の2次元画像を撮影し，3次元モデルにより3次元形状モデルを構築する。モデルから復元した2次元画像と元の2次元画像を比較し，復元精度によって3次元モデルの評価解析を行った結果，各種誤差要因のほか形状の複雑さの影響が判明した。形状の複雑度を定義し，複雑度に基づいて精度指標を修正すことで，複雑さの影響を減少した。	Using one CCD camera and the turn table, we propose a method to construct three dimensional object shape from two dimensional images. By comparing the two dimensional image obtained from three dimensional object shape constructed by our proposed method, and original image, we find that three dimensional object shape is restored precisely.

(c) Partially comparable

Fig. 3: Example pairs of Japanese and English abstracts

We conducted an experiment using the Japan Science and Technology Agency (JST) corpus of Japanese and English scientific-paper abstracts. It consists of pairs of Japanese and English abstracts with varying comparability, as exemplified in Fig. 3. The lengths of the Japanese abstracts range from 200 to 500 characters and those of the English abstracts range from 50 to 300 words. We used 107,979 pairs of abstracts in the field of information engineering, published in 1980 through 2004, to derive a bilingual lexicon with correlations. We used a Japanese morphological analyzer Mecab [1] and a language independent part-of-speech tagger TreeTagger [2] to segment the Japanese and English texts into words, respectively.

We prepared two test sets; AI test set consisting of 1,094 Japanese terms with reference English translations from the Japanese-English Index in the Encyclopedia of Artificial Intelligence (JSAI 2008) and NLP test set consisting of 1,661 Japanese terms with reference English translations from the Japanese-English Index in the Encyclopedia of Natural Language Processing (ANLP 2010).

We used the compositional translation method with each of the following three bilingual lexicons to produce a ranked list of English translations for a Japanese term in the two test sets.

(1) Corpus-derived lexicon + ordinary lexicon

The bilingual lexicon derived from the JST corpus was

[1] http://mecab.sourceforge.net/

[2] http://www.ims.stuttgart.de/projekte/corplex/TreeTagger/

merged with the EDR[3] Japanese-English, EDICT[4] Japanese-English, and Eijiro[5] English-Japanese Dictionaries. Since these ordinary lexicons do not contain correlations, a uniform correlation value of 0.1 was given to all pairs of Japanese and English words in them, and the maximum of the two values was selected for a pair of Japanese and English words contained in both the corpus-derived lexicon and the ordinary lexicons.

(2) Corpus-derived lexicon

The bilingual lexicon derived from the JST corpus only

(3) Ordinary lexicon

The EDR Japanese-English, EDICT Japanese-English, and Eijiro English-Japanese Dictionaries were merged into one and, then, augmented so that a ranked list of translation candidates could be output for an input term; namely, each pair of Japanese and English words was given a correlation value proportional to the number of pairs of aligned documents in which they co-occur.

For each bilingual lexicon, λ was adjusted using another set of Japanese terms and their English translations from the Japanese-English Index in the Encyclopedia of Artificial Intelligence. This set was disjoint with the above-mentioned AI test set. The value of λ was 0.40, 0.43, and 0.33 for (1) corpus-derived lexicon + ordinary lexicon, (2) corpus-derived lexicon, and (3) ordinary lexicon, respectively.

5.2 Experimental results

Table 1 lists the mean reciprocal rank (MRR) of the correct translations and Top *k* precision (*k*=1, 3, and 10), i.e., the percentage of input terms whose correct translations were included in those with *k* highest confidence scores, for the compositional translation with each of the three bilingual lexicons, where we judged only the reference translations as correct. The data suggest that the proposed framework is promising; not only the corpus-derived lexicon + ordinary lexicon but also the corpus-derived lexicon outperformed the ordinary lexicon. In Table 1, correct translations are broken down into two categories: translations the bilingual lexicon provides and translations produced compositionally. When the corpus-derived lexicon + ordinary lexicon and the corpus-derived lexicon were used, about 30% of the correct translations were those produced compositionally. This demonstrates the necessity and effectiveness of on-the-fly compositional translation.

Top *k* precisions of at most 50% were very low compared with those reported in previous literature on bilingual lexicon acquisition from parallel or comparable corpora. One of the reasons for the low precision is the test sets prepared independently of the corpus from which the bilingual lexicon derived. In fact, 11% of the Japanese terms in the AI test set and 12% of those in the NLP test set included word sequences not covered by the corpus-derived lexicon. Most of such terms were unpopular transliterated ones, e.g., "タクタイルボコーダ <*TAKUTAIRU BOKOODA*>" (*tactile vocoder*), those including proper nouns, e.g., "ボールドウィン効果 <*BOORUDOUIN KOOKA*>" (*Boldwin effect*), and scarcely used terms, e.g., "ブラーフミ文字 <*BURAAHUMI MOJI*>" (*Brahmi script*).

The data in Table 1 is rather singular; for almost 80% of the test terms whose correct translations were in top 10, the top ranked ones were actually correct. We can say that the proposed method is reliable for a term occurring rather frequently in the corpus, while it is unreliable for a term occurring infrequently in the corpus. The performance for the NLP test set was much worse than that for the AI test set. This is probably because the JST corpus contains a relatively small number of paper abstracts on natural language processing.

Table 2 lists the results of compositional translation with the corpus-derived lexicon + ordinary lexicon and that with the ordinary lexicon for several input terms. These results suggest the effectiveness of the proposed method as well as room for improvement.

6. Discussion

The compositional translation method has been widely used to extract a pair of a word and its translation from corpora, although it is restricted to extracting a pair of compound words. It usually consults an existing bilingual lexicon to generate candidate translations, which then are validated by using a corpus (Cao and Li 2002; Tanaka 2002; Baldwin and Tanaka 2004; Tonoike et al. 2006). In contrast, we proposed consulting a bilingual lexicon derived from a corpus. The experiment demonstrated that our framework improved the possibility of producing a correct translation. Note that unless a correct translation was produced, the validation procedure would be useless. A distinguishing feature of our improved compositional translation method is that it estimates confidence scores for candidate translations. Although there has been work investigating score functions for compositional translation (Tonoike et al. 2006), our score is unique in that it is based on a comparable corpus.

The method described in Sec. 3 is not the only way to derive a bilingual lexicon from a comparable corpus. Alternatively, we can extract parallel sentence pairs from a comparable corpus to acquire a bilingual lexicon with a statistical machine translation tool. This is a common way to exploit comparable corpora for SMT (Fung and Cheung 2004; Munteanu and Marcu 2005; Abdul-Rauf and Schwenk 2009). It is also applicable to augmenting a seed bilingual lexicon for contextual similarity-based bilingual lexicon acquisition from a comparable corpus (Morin and Prochasson 2011). Our method based on co-occurrence statistics in pairs of aligned documents should be evaluated comparatively with this alternative. Our method would be better for very nonparallel corpora, while the alternative would be better for comparable

[3] http://www2.nict.go.jp/r/r312/EDR/index.html
[4] http://www.csse.monash.edu.au/~jwb/edict.html
[5] http://www.alc.co.jp/

Table 1: Summary of experimental results

(a) Artificial Intelligence domain (# of test terms: 1,094)

Bilingual Lexicon	Corpus-derived + ordinary	Corpus-derived	Ordinary
MRR	0.44	0.4	0.22
Top 1 precision	0.402	0.370	0.197
(Bilingual lexicon)	(0.289)	(0.263)	(0.089)
(Compositional translation)	(0.113)	(0.107)	(0.108)
Top 3 precision	0.464	0.428	0.238
(Bilingual lexicon)	(0.326)	(0.297)	(0.112)
(Compositional translation)	(0.138)	(0.131)	(0.125)
Top 10 precision	0.510	0.473	0.351
(Bilingual lexicon)	(0.351)	(0.320)	(0.135)
(Compositional translation)	(0.169)	(0.153)	(0.144)

(b) Natural Language Processing domain (# of test terms: 1,661)

Bilingual Lexicon	Corpus-derived + ordinary	Corpus-derived	Ordinary
MRR	0.35	0.31	0.20
Top 1 precision	0.314	0.282	0.167
(Bilingual lexicon)	(0.231)	(0.202)	(0.102)
(Compositional translation)	(0.083)	(0.081)	(0.066)
Top 3 precision	0.377	0.331	0.217
(Bilingual lexicon)	(0.272)	(0.229)	(0.143)
(Compositional translation)	(0.105)	(0.102)	(0.074)
Top 10 precision	0.415	0.362	0.271
(Bilingual lexicon)	(0.296)	(0.246)	(0,178)
(Compositional translation)	(0.120)	(0.117)	(0.093)

corpora from which many parallel sentence pairs can be extracted.

The followings are the directions for improving our framework. First, we need to improve the bilingual lexicon with correlations. The present corpus-derived bilingual lexicon contains too many spurious pairs. The examples in Table 2 imply that it contains such pairs as (属性<*ZOKUSEI*> (*property*), decision tree), (ネットワーク <*NETTOWAAKU*> (*network*), service), and (反駁 <*HANBAKU*> (*refutation*), PAC learning model). This may be unavoidable as we do not have a seed lexicon. However, once a bilingual lexicon is acquired, we can use it to acquire a less noisy bilingual lexicon. In other words, we can refine our bilingual lexicon incrementally.

Second, there is room for refining the confidence score. Currently, we do not consider the relation or compatibility between constituent translations. A possible refinement of the confidence score is to multiply the harmonic means of the confidence scores for constituent translations by the correlation between the constituent translations, which can be estimated from a target-language monolingual corpus. We have an alternative to this refinement. That is, producing unlikely translations as candidates and validating the candidates by using a target-language monolingual corpus or the Web may not be problematic (Dagan and Itai 1994; Grefenstette 1999; Way and Gough 2003).

Third, we need to extend our compositional translation model to allow word order to be changed. For example, while a Japanese term is a noun sequence, its English translation can include a prepositional phrase. A factor of structural transfer should be incorporated into our confidence score. Some previous work has addressed compositional translation involving changes in word order (Baldwin and Tanaka 2004).

Table 2: Example of compositional translation results

#	Input term	Rank	Corpus-derived + ordinary		Ordinary	Reference translation
			Translation	Score	Translation	
1	属性継承 <*ZOKUSEI KEISHOU*>	1	attribute inheritance	0.060	attribute inheritance	property inheritance
		2	attribute succession	0.023	***property inheritance***	
		3	decision tree inheritance	0.021	characteristic inheritance	
2	単純再帰ネットワーク <*TANJUN SAIKI NETTOWAKU*>	1	simple recursive network	0.021	-	simple recurrent network
		2	simple recursion network	0.018	-	
		3	simple recursive service	0.017	-	
3	統合データベース <*TOUGOU DETABESU*>	1	***integrated database***	0.188	integration data base	integrated database
		2	intermolecular	0.069	synthesis data base	
		3	information database	0.058	fusion data base	
4	統計的機械翻訳 <*TOUKEI TEKI KIKAI HONYAKU*>	1	***statistical machine translation***	0.062	statistic object machine translation	statistical machine translation
		2	statistical method machine translation	0.047	statistic target machine translation	
		3	statistical machine translation system	0.046	statistic aim machine translation	
5	統計的統語解析 <*TOUKEI TEKI TOUGO KAISEKI*>	1	statistical syntactic analysis	0.040	-	statistical parsing
		2	statistical method syntactic analysis	0.033	-	
		3	statistical syntactic structure	0.032	-	
6	反駁 <*HANBAKU*>	1	PAC learning model	0.089	counterblast	refutation
		2	・ F・	0.067	negation	
		3	***refutation***	0.062	rebuttal	
7	ベイズ決定理論 <*BEIZU KETTEI RIRON*>	1	***Bayes decision theory***	0.056	-	Bayes decision theory
		2	unknown datum theory	0.034	-	
		3	Bayesian decision theory	0.034	-	
8	命題様相論理 <*MEIDAI YOUSOU ROMMRI*>	1	proposition modal logic	0.062	proposition aspect logic	propositional modal logic
		2	***propositional modal logic***	0.036	problem aspect logic	
		3	proposition modal	0.032	proposition state logic	

[Note] Bold and Italicized translations were judged as correct.

7. Conclusion

We improved the compositional term translation method with comparable corpora. A bilingual lexicon consisting of word sequence pairs within terms and their correlations is acquired from a document-aligned corpus. The correlations between word sequences in two languages are calculated based on their co-occurrence in aligned document pairs. Then, for an input term, candidate translations are compositionally produced together with their confidence scores, which are defined based on the correlations between the constituents. Thus, the correct translation for the input term can be selected from among as many candidate ones as possible.

An experiment with a comparable corpus consisting of Japanese and English scientific-paper abstracts demonstrated that compositional translation with the corpus-derived bilingual lexicon outperformed that with an ordinary bilingual lexicon. Future work includes the incremental improvement of the bilingual lexicon with correlations, the refinement of the confidence score, and the extension of the compositional translation model to allow word order to be changed.

8. Acknowledgements

We thank the Japan Science and Technology Agency for permitting us to use the JST Japanese and English scientific-paper abstracts. This work was partly supported

by Grant-in-Aid for Scientific Research, MEXT (22300032).

9. References

Abdul-Rauf, Sadaf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 16-23.

Andrade, Daniel, Tetsuya Nasukawa, and Jun'ichi Tsujii. 2010. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 19-27.

ANLP (Association for Natural Language Processing). 2010. Gengo Shori Gaku Jiten (Encyclopedia of Natural Language Processing). Kyoritsu Publishing Co. (Tokyo).

Baldwin, Timothy and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrated Processing*, pages 24-31.

Cao, Yunbo and Hang Li. 2002. Base noun translation using Web data and the EM algorithm. In *Proceedings of the 19h International Conference on Computational Linguistics*, pp. 127-133.

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, vol. 20, No. 4, pp. 563-596.

Fung, Pascale and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 57-63.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp. 414-420.

Grefenstette, Gregory. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, Vol. 21.

Ismail, Azniah and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Poster Volume, pages 481-489.

JSAI (Japanese Society for Artificial Intelligence). 2008. Jinkou Chiou Gaku Jiten (Encyclopedia of Artificial Intelligence). Kyoritsu Publishing Co. (Tokyo).

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the ACL*, pp. 48-54.

Matsumoto, Yuji and Takehito Utsuro. 2000. Lexical knowledge acquisition. In *R. Dale, H. Moisl, and H. L. Somers (ed.). Handbook of Natural Language Processing*, Ch. 24, pp. 563-610 (Marcel Dekker Inc.).

Morin, Emmanuel and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, ACL 2011*, pages 27-34.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, vol. 31, No. 4, pp. 477-504.

Och, Franz Josef and Hermann Ney. 2003. A Systematic comparison of various statistical alignment models. *Computational Linguistics*, vol. 29, No. 1, pp. 19-51.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Tanaka, Takaaki. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 981-987.

Tonoike, Masatsugu, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2006. Comparative Study on Compositional Translation Estimation using a Domain/Topic-Specific Corpus collected from the Web. In *Proceedings of the 2nd International Workshop on Web as Corpus*, pp. 11-18.

Utsuro, Takehito, Takashi Horiuchi, Kohei Hino, Takeshi Hamamoto, and Takeaki Nakayama. 2003. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In *Proceedings of the 10th Conference of the European Chapter of the ACL*, pp. 355-362.

Way, Andy and Nano Gough. 2003. *wEBMT*: Developing and validating an example-based machine translation system using the World Wide Web. *Computational Linguistics*, vol. 29, No. 3, pp. 421-457.

Multi-word term extraction from comparable corpora by combining contextual and constituent clues

Nikola Ljubešić[1], **Špela Vintar**[2], **Darja Fišer**[2]

[1]Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
[2]Department of Translation, Faculty of Arts, University of Ljubljana

Abstract

In this paper we present an approach to automatically extract and align multi-word terms from an English-Slovene comparable health corpus. First, the terms are extracted from the corpus for each language separately using a list of user-adjustable morphosyntactic patterns and a term weighting measure. Then, the extracted terms are aligned in a bag-of-equivalents fashion with a seed bilingual lexicon. In the extension of the approach we also show that the small general seed lexicon can be enriched with domain-specific vocabulary by harvesting it directly from the comparable corpus, which significantly improves the results of multi-word term mapping. While most previous efforts in bilingual lexicon extraction from comparable corpora have focused on mapping of single words, the proposed technique successfully augments them in that it is able to deal with multi-word terms as well. Since the proposed approach requires minimal knowledge resources, it is easily adaptable for a new language pair or domain, which is one of its biggest advantages.

1. Introduction

Resource-poor language pairs and domains can benefit greatly from the increasingly popular field of bilingual lexicon extraction from comparable corpora. The approaches bootstrap lexica of general as well as domain-specific vocabulary from large, usually web-based collections of texts in two languages that are not translations of each other but rather share common properties, such as subject field, time of publication, target audience etc.

Term extraction from comparable corpora is usually understood as a task that combines monolingual term recognition in each of the languages and cross-lingual term alignment using various techniques. Fung and McKeown (1997) and Rapp (1995) are considered the beginners of the alignment approach based on the hypothesis that two terms are likely to be translations of each other if they occur in similar contexts. Several authors experiment with different measures of context similarity (Chiao and Zweigenbaum, 2002; Morin et al., 2007) and report up to 80% accuracy in finding the correct translation among the 20 best candidates. Some approaches extend the bilingual mapping through cognate detection (Saralegi et al., 2008), while Lee et al. (2010) propose an EM-based hybrid model for term alignment.

It should be noted that these early approaches deal almost exclusively with single-word terms, and also that nearly all authors conclude that the size and comparability of the corpora play a key role in achieving good performance. In our previous work we too have shown a strong positive correlation of the degree of corpus comparability and size (Ljubešić et al., 2011). In addition, we have established that good coverage of the seed lexicon that is used to translate the features in the context vectors plays a much bigger role than its size, and that the seed lexicon can be built completely automatically provided that there is a lexical overlap between two closely-related languages (Fišer and Ljubešić, 2011). However, in all our previous experiments in lexicon extraction, we, just like most related work, have not tackled multi-word expressions, which are very important in natural language processing and for which there are even fewer already existing resources, especially because a number of domains evolve at a great speed, making the static resources obsolete very quickly.

The bag-of-equivalents term alignment approach is an effective method of finding multi-word-to-multi-word term equivalents. It is similar to the compositional approach used by Morin and Daille (2010) or to the abduction method described by Carl et al. (2004), however both of the above use predefined lexico-syntactic patterns to predict term variations. Our approach is more robust, however it requires a domain-specific translation lexicon, ideally with several translation possibilities, and this may not be readily available (Vintar, 2010). The main goal of this paper is to show that by enriching the lexicon with automatically extracted domain-specific single-word terms the overall performance of multi-word term extraction from a comparable corpus can be significantly improved.

This paper is structured as follows: in the next section we present all the resources and tools that were used in the experiment. The experimental setup is described in detail in Section 3. The results are evaluated and discussed in Section 4, and the paper is concluded with some final remarks and ideas for future work.

2. Resources and tools used

2.1. Comparable corpus

The main source of lexical knowledge in this experiment was the English-Slovene comparable corpus of on-line articles on health and lifestyle, which had already been used successfully in our previous research (Fišer et al., 2011). Health-related documents were extracted from the ukWaC (Baroni et al., 2009) and slWaC (Ljubešić and Erjavec, 2011) web corpora by the criterion that the cosine similarity to a domain model had to be higher than 0.25. The domain model was built on documents from the two main health-related Internet domains. It is based on content words as features and TF-IDF feature weights where the IDF weights were calculated on a news-domain corpus.

The subset of the constructed domain corpus we used in this experiment contains 1.5 million tokens for each language.

2.2. Seed lexicon

The seed lexicon used as an anchor between the two languages was constructed from the freely available Slovene-English and English-Slovene Wiktionaries that cover mostly general vocabulary. The entries from both Wiktionaries were merged and if the same pair of words was found in both resources, they were given a higher probability. The seed lexicon constructed in this way contains 6.094 entries.

2.3. LUIZ

LUIZ is a hybrid bilingual term extractor that uses parallel or comparable corpora as input and outputs mono- and bilingual lists of term candidates (Vintar, 2010).
Term recognition is performed on the basis of user-adjustable morphosyntactic patterns provided for each language. Then the extracted candidate phrases are assigned a term-hood value by comparing the frequency of each word to a reference corpus. Term alignment is performed using the bag-of-equivalents approach (Vintar, 2010), which presupposes a probabilistic bilingual lexicon as input. A list of possible translation candidates for a source multiword term is proposed by comparing each target term candidate to a bag of potential translation equivalents provided by the lexicon and computing an equivalence score.

2.4. ccExtractor

ccExtractor is a context-based bilingual lexicon extraction tool that was built during our previous experiments (Ljubešić et al., 2011; Fišer et al., 2011; Ljubešić and Fišer, 2011). It consists of a series of scripts that enable:

- building context vectors for a list of headwords from each corpus,
- translating features of context vectors from source language to target language via an existing seed lexicon and
- calculating the best translation candidates between headwords in the source language and the target language.

In this research the tool is used to enhance the general small seed lexicon used for multi-word term alignment with LUIZ.

3. Experimental setup

The main task in the experiment was to find translation candidates for multi-word terms from the health comparable corpus. In order to achieve this, the experiment was divided into three parts.
In the first part of the experiment we used LUIZ to extract multiword term candidates from both corpora. The result is a list of 25,865 English and 27,102 Slovene multiword term candidates.
In the second part of the experiment we aligned the extracted multiword term candidates between English and Slovene with LUIZ via our seed lexicon.
In the third part of the experiment we tried to improve the results by enhancing the seed lexicon used by LUIZ with 412 translation equivalents of the domain-specific vocabulary in the corpus that is not covered in the seed lexicon, which we obtained with ccExtractor. Term extraction and alignment were then repeated with the same settings, the only difference being the extended seed lexicon.
With this step we combined contextual information obtained from ccExtractor with the constituent information provided by LUIZ.

3.1. Term extraction

Term recognition in each part of the corpus was performed with the help of a predefined set of morphosyntactic patterns for each language. These patterns describe part-of-speech sequences of mainly noun phrases up to 5 words in length. Once candidate phrases were extracted from the corpora, a term weighting measure was used to assign a termhood value to each phrase. This measure computes single-word termhood by comparing the frequency of each word ($f_{n,D}$) to a reference, non-specialized corpus ($f_{n,R}$), and then combines the termhood scores of all constituent words with the frequency (f_a) and length (n) of the entire candidate phrase.

$$W(a) = \frac{f_a^2}{n} * \sum_{1}^{n} (\log \frac{f_{n,D}}{N_D} - \log \frac{F_{n,R}}{N_R}) \qquad (1)$$

3.2. Term alignment

The extracted multi-word terms were then aligned in the bag-of-equivalents fashion (see section 2.3) using the seed bilingual lexicon. For a given source multi-word term each target term candidate is compared to a bag of potential translation equivalents provided by the lexicon and an equivalence score is computed, thus generating a ranked list of possible translation candidates. If, for example, the bilingual lexicon contains the English-Slovene entries

```
blood  kri     1.0
flow   pretok  0.66  tok  0.33
```

the bag-of-equivalents for the English term candidate *blood flow* will contain all three equivalents, *kri*, *pretok* and *tok*. We now compare the Slovene term candidates to the bag and compute the equivalence score as the sum of the translation probabilities found in the target term, normalized by term length. Thus, for the above English term we extract

```
pretok krvi      0.83
tok krvi         0.66
šibak tok krvi   0.43
```

This approach is able to identify several good translation equivalents for a source term, which is especially valuable in domains with less standardized terminology and a lot of term variation. Furthermore, this approach is also able to find translation equivalents for the terms for which seed lexicon entries are missing or faulty.
In our current setting we are able to identify multi-word-to-multi-word equivalents of different lengths, but we do not identify single-word-to-multi-word term pairs.

3.3. Extension of the seed lexicon

In the third part of the experiment the idea was to extend the alignment of the extracted multi-word terms with the extension of the seed lexicon by adding the most relevant vocabulary from the corpus. Using the ccExtractor, we extracted three most probable Slovene translations for all English lemmas that were not already included in the initial seed lexicon.

The headwords in both parts of the corpus had to satisfy the minimum frequency constraint of 50 occurrences which is the most reasonable frequency threshold as proven in our previous experiments (Ljubešić et al., 2011). When building context vectors, a window of three lemmas on both sides of the headword was used and the collected features were weighted by the TF-IDF score. Context similarity was calculated with the Dice similarity metric. The probabilities of the translation candidates were calculated as their context similarity weights scaled to a probability distribution.

There were 412 English lemmas in the corpus that had not been present in the seed lexicon already and that satisfied the occurrence frequency criterion. Therefore, our extended seed lexicon contains 6.506 entries. This lexicon was used in the second run of the experiment in which all the other settings were the same as in the first run.

4. Evaluation of the results

In this section we report the results of manual evaluation of term extraction in both languages as well as the quality of term alignment. We focus here on measuring the accuracy of term extraction and alignment and while recall would be interesting to study more closely as well, we were not able to do it in this experiment because in order to measure it, we would need either a comprehensive terminological dictionary of this area for measuring absolute recall or a manually annotated corpus with multi-word terms in both languages for measuring recall relative to the terms used in the corpus.

4.1. Evaluation of term extraction

In total, 25,865 term candidates were extracted from the English part of the corpus and 27,102 from the Slovene part. The extracted term candidates were assigned a termhood score and in order to evaluate the quality of the extracted terms, we manually evaluated 100 highest-ranked term candidates for each language.

In the evaluation scheme, each candidate was categorized into one of three possible categories:

- the candidate was a correctly extracted multi-word term from the health domain;
- the candidate was a correctly extracted multi-word term but did not belong to the health domain;
- the candidate was not correctly extracted (a part of a multi-word term) or the multi-word expression was not a term.

The results of manual evaluation are shown in Table 1. Among the English candidates, 76 were correctly extracted health terms (e.g. *blood test*), 5 were terms but belonged to some other domain (e.g. *primary school*) and 19 of the candidates were either incorrectly extracted multi-word terms or multi-word expressions that belong to the general vocabulary (e.g. *next year*). The results for Slovene are slightly better: 86 of the candidates were correct, 3 were terms from a different domain and 11 were incorrectly extracted multi-word terms or other multi-word combinations. The reason for better results in Slovene is probably a cleaner, less noisy corpus, both in terms of domain-specific documents and in terms of corpus annotation because slWaC was built much more conservatively than ukWaC.

Term quality	English	Slovene
good term	76%	86%
term from a different domain	5%	3%
not a term	19%	11%

Table 1: Evaluation of term extraction on 100 highest ranked term candidates

An interesting characteristic in the highest-ranking term candidates is their length. In both languages, two-word terms are by far the most frequent, with only 4 English and 6 Slovene candidates that are longer than two words. On the one hand, this is to be expected because the longer the term, the less frequent it is in the corpus. But it also must be noted that the corpus does not contain expert medical texts but mostly magazine articles with health issues and lifestyle advice for the general public that contain fewer complex medical terms.

4.2. Evaluation of term alignment

The quality of term alignment was evaluated for each run of the experiment, with the original and the extended seed lexicon, in order to evaluate the impact of seed lexicon extension.

The extension of the seed lexicon was evaluated in our previous work (Fišer et al., 2011). It has a correct translation in the first position in 45% of cases while in additional 11% of cases there is a correct translation among the first ten candidates. We did not measure specifically the percentage of correct translations on the first three positions used in this research.

In this part of evaluation we checked the proposed term pairs and measured the accuracy of term alignment by manually inspecting the list of 477 multi-word term pairs that received an equivalence score higher than 0.5 in either run of the experiment. In the list 380 of these pairs were identical in both runs of the experiment while translation suggestions for 97 of the source terms were different with the two different seed lexicons. First we evaluate the termhood of the source language candidates and then, in case the candidates are considered a term, we evaluate the accuracy of its translation.

The evaluation schema used when evaluating termhood is:

- good term;
- term from a different domain;

- not a term,

while the evaluation schema used for evaluating the translation quality is:

- correct translation;
- close translation;
- incorrect translation.

Score	Percentage
good term	43.6%
term from a different domain	12.6%
not a term	43.8%

Table 2: Evaluation of term extraction on the 477 source language term candidates with equivalence score higher than 0.5

As Table 2 shows, source language term candidates that have good probable translation equivalents (equivalence score higher than 0.5) are partial or full terms in 56% of the cases. This is much lower than when evaluating the top ranked term candidates. In our opinion, there are two reasons for that:

- these are the terms with a high equivalence score, not a high termhood score;
- term candidates with a high equivalence score consist of constituents found in the general seed lexicon from which terms are rarely built.

The quality of term alignment is shown in Figure 1. We stress once again that term alignment evaluation was performed only on those pairs that were good terms in the source language. When using the original seed lexicon, translations for 41.5% of the terms are correct or close to correct, while, when using the extended seed lexicon, 52.2% of translations are correct or close to correct. It is interesting to note that there is an increase of almost 8% of the correctly aligned terms while the number of close to correct terms goes up by 3%. At the same time, the number of incorrectly aligned terms goes down by almost 11%. This can be considered a very big improvement and clearly shows that it is very beneficial to add the most relevant vocabulary for the particular domain or corpus to the seed lexicon, even if the equivalents are extracted automatically and are therefore somewhat noisy.
Another interesting observation is the fact that the pairs that were shared among the two seed lexicons are of a relatively high quality already and that the extension of the seed lexicon helped in exactly those cases that the original lexicon was not able to handle well at all, either because it was too small in size or too general for this particular domain. This shows that the already existing resources can easily and successfully be complemented with a simple and fully automatic technique such as ours, giving a big boost to the quality of term alignment.

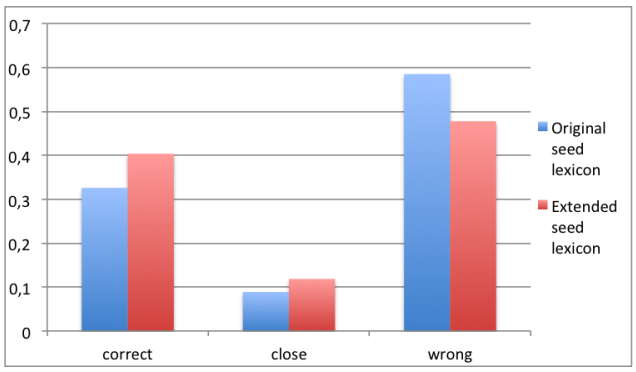


Figure 1: Evaluation of term alignment on terms with the equivalence score higher than 0.5 using the original and extended seed lexicon

5. Conclusions and future work

In this paper we presented an approach to extract translations of multi-word terms from domain-specific comparable corpora, a problem which has so far been largely neglected by most of the related work. We used LUIZ, a hybrid tool for bilingual multi-word term extraction and alignment. In addition, we used ccExtractor, a statistical tool for finding translation equivalents for single-word terms in comparable corpora in order to extend the seed lexicon with the most relevant terms in the corpus, which improved the results of multi-word term alignment by almost 11%. Additionally, this is the first extrinsic evaluation of context-based single-word lexicon extraction from comparable corpora.
While these results do not outperform the benchmark results achieved by LUIZ when aligning multi-word terms in parallel corpora, this is understandable because looking for MWT equivalents in comparable corpora is a much more difficult task. In addition, although the number of resulting MWTs obtained in this experimental setting is not very large, their precision is much higher than in the regular SWT extraction and alignment approach. With this in mind, the results we obtained with the extended seed lexicon are very encouraging and can already be very useful as a time-saving aid to terminologists who no longer have to look for the terms and their equivalents themselves but merely validate/correct the proposed ones.
Further improvements are possible by increasing the corpus size, which would, to start with, yield more single-word term candidates. This would improve the coverage of MWTs but could possibly have an adverse effect as well if a larger amount of noisy data in the lexicon would decrease the precision of the alignment. Finally, the term extraction procedure would benefit from more data as well.
In the future we plan to use the approach on a more scientifically-oriented medical domain corpus where complex terms play an even bigger role and there is less general language. Currently, we are also working on the adaptation of LUIZ to handle new languages, such as Croatian, which will enable the creation of multilingual terminological resources from web-based domain-specific comparable corpora.

6. Acknowledgement

Research reported in this paper has been supported by the ACCURAT project within the EU 7th Framework Programme (FP7/2007-2013), grant agreement no. 248347, and by the Slovene national postdoctoral grant no. Z6-3668.

7. References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, pages 209–226.

Michael Carl, Ecaterina Rascu, and Johann Haller. 2004. Using weighted abduction to align term variant translations in bilingual texts. In *Proceedings of LREC 2004*.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 125–131, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Darja Fišer, Nikola Ljubešić, Špela Vintar, and Senja Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 19–26, Portland. Association for Computational Linguistics.

P. Fung and K. McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Lianhau Lee, Aiti Aw, Min Zhang, and Haizhou Li. 2010. Em-based hybrid model for bilingual terminology extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 639–646, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrwac and slwac: Compiling web corpora for croatian and slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 395–402. Springer.

Nikola Ljubešić and Darja Fišer. 2011. Bootstrapping bilingual lexicons from comparable corpora for closely related languages. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 91–98. Springer.

Nikola Ljubešić, Darja Fišer, Špela Vintar, and Senja Pollak. 2011. Bilingual lexicon extraction from comparable corpora: A comparative study. In *First International Workshop on Lexical Resources, An ESSLLI 2011 Workshop, Ljubljana, Slovenia - August 1-5, 2011*.

Emmanuel Morin and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1):79–95.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671, Prague, Czech Republic, June. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, MA, USA.

X. Saralegi, I. San Vicente, and A. Gurrutxaga. 2008. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *6th International Conference on Language Resources and Evaluations (LREC'08) - Building and using Comparable Corpora workshop*, pages 27–32, Marrakech, Morocco.

Špela Vintar. 2010. Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2):141—158.

Textual Characteristics of Different-sized Corpora

Robert Remus[†], **Mathias Bank**[‡]

†Natural Language Processing Group, University of Leipzig, Germany
‡Pattern Science AG, 63579 Freigericht, Germany
rremus@informatik.uni-leipzig.de, m.bank@cid.biz

Abstract

Recently, textual characteristics, i.e. certain language statistics, have been proposed to compare corpora originating from different genres and domains, to give guidance in language engineering processes and to estimate the transferability of natural language processing algorithms from one corpus to another. However, until now it is unclear how these textual characteristics behave for different-sized corpora. We monitor the behavior of 7 textual characteristics across 4 genres – news articles, Wikipedia articles, general web text and fora posts – and 10 corpus sizes, ranging from 100 to 3,000,000 sentences. Thereby we show, certain textual characteristics are almost constant across corpus sizes and thus might be used to reliably compare different-sized corpora, while others are highly corpus size-dependent and thus may only be used to compare similar- or same-sized corpora. Moreover we find, although textual characteristics vary from genre to genre, their behavior for increasing corpus size is quite similar.

Keywords: Textual Characteristics, Language Statistics, Corpus Comparison

1. Introduction

With the continuous development of natural language processing methods and machine learning algorithms, more and more approaches become available to assess various aspects of natural language text. Among these methods, many are highly text type-dependent and hence not easily transferable from one genre or domain to another, e.g. parsing (Sekine, 1997), word sense disambiguation (Escudero et al., 2000) and sentiment analysis (Aue and Gamon, 2005; Blitzer et al., 2007; Wang and Liu, 2011). Therefore, Bank et al. (2012) recently proposed to estimate the transferability of natural language processing methods from one genre or domain to another via the *textual characteristics* of the respective corpora. They found textual characteristics to vary greatly for different genres and pose the hypothesis, if textual characteristics of one corpus are similar to those of another, it is likely that algorithms working well on the former corpus also work well on the latter.

However, Bank et al. (2012) do not study the behavior of textual characteristics of *different-sized corpora*. Their analysis requires corpora of the same size in order to provide reliable results. As this requirement might not be applicable to real world scenarios, where one wants to compare different-sized corpora, we will address the following questions: Do textual characteristics vary not only across genres, but also across corpus sizes? If so, which textual characteristics are corpus size-dependent and which are not? Put differently, which textual characteristics may be used to compare both different-sized and same-sized corpora and which might only be used to compare similar- or same-sized corpora?

1.1. Related Work

To our knowledge, there has been only very little general work on comparing corpora based on their textual characteristics, and almost no work regarding potential corpus size-dependences of textual characteristics. Kilgarriff (2001) surveys several language statistics to measure corpus similarity and corpus homogeneity based on words and their distributions. Rayson and Garside (2000) propose to compare corpora using "frequency profiles" of words as well as syntactic and semantic tags.

With a specific goal in mind, several studies on textual characteristics have been carried out: Suzuki and Kageura (2007) explore Japanese prime ministers' Diet addresses, by focusing on the "quantity and diversity of nouns", to develop an understanding of changes in political content and the differences in 2 types of Diet addresses. Verspoor et al. (2009) investigate surface linguistic structures, sentence length distributions and term probability distributions in traditional and Open Access scientific journals to proof their similarity in order to ultimately be able to re-use previously proposed natural language processing algorithms. Na et al. (2010) analyze movie reviews from 4 online genres: critic reviews, user reviews, posts to discussion boards and blog posts. They analyze their vocabulary, average number of words, sentences and paragraphs, part of speech distributions, various movie aspects, as well as opinions expressed in the texts, partly automatically, partly manually. Goeuriot et al. (2011) analyze textual characteristics, e.g. posting lengths and part of speech distributions, of posts to 3 different drug review fora. Ghose and Ipeirotis (2011) measure amongst others readability and spelling accuracy of reviews to assess their helpfulness to other users and the reviews' economic impact. All studies mentioned above compare different-sized corpora, however without implicitly or explicitly addressing the potential difficulties these comparisons pose.

1.2. Outline

This paper is structured as follows: In the next Section, we describe the textual characteristics introduced in Bank et al. (2012). In Section 3. we apply them to corpora from different genres and monitor their behavior for different corpus sizes. Finally, we draw conclusions and point out possible directions for future work in Section 4.

2. Textual Characteristics

Bank et al. (2012) use only textual characteristics, i.e. language statistics, that can be easily and quickly calculated, without the need for advanced language processing modules, e.g. part of speech taggers or syntax parsers. This enables them to directly apply all measures to any corpus and ensures comparable results among them, without having to adapt those text type-dependent modules to previously unknown language properties. These textual characteristics are:

1. Shannon's *entropy* H measures the average amount of information in an underlying data structure. Applied in the field of language engineering, the mean amount of information of a token t_i can be calculated by approximating its probability $p(t_i)$ via its frequency in a given corpus. The entropy as given in Formula 1 is normalized to the vocabulary size $|V|$, i.e. the number of types in the corpus:

$$H = -\sum_{t_i \in V} p(t_i) \log_{|V|} p(t_i) \quad (1)$$

2. The *relative vocabulary size* R_{Voc} (Těšitelová, 1992, chapter 1.2.3.3) is given by the ratio of the vocabulary size $|V|$ and the total number of tokens N_m with respect to "meaningful" words. These are defined as words, that are not function words[1] ($N_m = \{t \mid t \notin N_f\}$), e.g. nouns, adjectives and verbs:

$$R_{\text{Voc}} = \frac{|V|}{N_m} \quad (2)$$

3. The *vocabulary concentration* C_{Voc}(Těšitelová, 1992, chapter 1.2.3.3) is defined by the ratio of the total number of tokens N_{top} with respect to the most frequent terms in the vocabulary V ($V_{top} = \{t \mid t \in V \wedge r(t) \leq 10\}$) and the total number of tokens N in a corpus

$$C_{\text{Voc}} = \frac{N_{top}}{N} \quad (3)$$

where rank $r(t)$ is defined as the position of a token t in a frequency-ordered list.

4. The *vocabulary dispersion* D_{Voc} expresses the relative amount of low frequency tokens ($V_{\text{low}} = \{t \mid t \in V \wedge f(t) \leq 10\}$) in the vocabulary V:

$$D_{\text{Voc}} = \frac{|V_{\text{low}}|}{|V|} \quad (4)$$

where frequency $f(t)$ is defined as the number of occurrences of the token t in a corpus.

5. The *corpus predictability* CP expresses the transition probabilities between tokens. For this, we need to calculate the entropy of a first-order Markov source $\mathcal{S}$ of two tokens t_i, t_j as given in Formula 5

$$H(\mathcal{S}) = -\sum_{t_i} p(t_i) \sum_{t_j} p_{t_i}(t_j) \log p_{t_i}(t_j) \quad (5)$$

where $p_{t_i}(t_j)$ denotes the probability of t_j given that it is preceded by t_i. CP is then calculated by normalizing the entropy of a first-order Markov source by its maximum possible entropy and subtracting it from 1:

$$CP = 1 - \frac{H(\mathcal{S})}{H_{max}(\mathcal{S})} \quad (6)$$

6. A rudimentary *grammatical complexity* GC can be calculated by the ratio of the number of function words N_f to the number of meaningful words N_m:

$$GC = \frac{N_f}{N_m} \quad (7)$$

Although this rather basic approach cannot state a real level of grammatical structure of a corpus, it still provides evidence for the amount of effort put into expressing syntax.

7. The *average sentence length* L_S influences parsing, relation extraction etc. The length $|s|$ of a sentence s is defined by the amount of tokens it contains, and the average sentence length of all sentences S is defined as in Formula 8:

$$L_S = \frac{1}{|S|} \sum_{s \in S} |s| \quad (8)$$

Additionally, Bank et al. (2012) measure *spelling accuracy* and *information density*. As both textual characteristics require manual intervention, we only compute the 7 measures described above.

3. Experiments

We now construct different-sized corpora and apply the textual characteristics described in Section 2. to them.

3.1. Constructing Different-sized Corpora

For our experiments we use 3 large English-language corpora provided by the *Wortschatz* project[2] (Quasthoff et al., 2006), each originating from a different genre: news articles, Wikipedia articles and general web text. To ensure comparability, all Wortschatz corpora are built in a standardized fashion (Quasthoff and Eckart, 2009). Their intended use is statistical corpus and language comparison (Eckart and Quasthoff, 2010). As an additional genre, we use a corpus of posts to the automotive web forum `benzworld.org`. Due to copyright reasons, this corpus is not publicly available.

To study the behavior of textual characteristics for different corpus sizes we construct sub-corpora C_k^g containing $k \in \{100, 300, 1000, 3000, \ldots, 3000000\}$ sentences for each genre $g \in \{\text{news}, \text{wikipedia}, \text{web}, \text{fora\_posts}\}$ so that

$$\forall l < m : C_l^g \subset C_m^g$$

i.e. any smaller corpus is always a real subset of any larger corpus. Table 1 provides an overview of the resulting news article, Wikipedia article, web text and fora post corpora.

[1] As function words N_f Bank et al. (2012) defined: *the, a, an, he, him, she, her, they, us, we, them, it, his, to, on, above, below, before, from, in, for, after, of, with, at, and, or, but, nor, yet, so either, neither, both, whether*

[2] `http://wortschatz.uni-leipzig.de/`

3.2. Results

Applying the textual characteristics described in Section 2. to these corpora leads to the results presented Figure 1(a), 1(b), 1(c) and 1(d). Interestingly, although the actual textual characteristics vary from genre to genre as expected (cf. also Table 1) and as it has been shown before (Bank et al., 2012), their behavior for different-sized corpora is very similar across all 4 genres: Not surprisingly, vocabulary concentration C_{Voc}, grammatical complexity GC and average sentence length L_S are almost "constant" for sufficiently large corpora, i.e. $k > 1000$. We note however, the larger the corpus, i.e. the larger k,

1. the lower its entropy H,
2. the lower its relative vocabulary size R_{Voc},
3. the lower its vocabulary dispersion D_{Voc} and
4. the higher its corpus predictability CP.

Across the 4 genres all pairwise *correlations* of H, R_{Voc}, D_{Voc} and CP are greater than 0.99 (significant at level $\alpha = 0.001$). Although this behavior may need further clarification in more experiments, it seems to signify invariant language properties, irrespective of the considered genres.

3.3. Discussion

The intuition behind the observed behavior of entropy, relative vocabulary size, vocabulary dispersion and corpus predictability is as follows: The entropy H is known to be dependent on the "message" length N (Manning and Schütze, 1999). The longer the message, i.e. the larger the corpus, the more redundant information it contains and hence the entropy decreases. The relative vocabulary size R_{Voc} decreases with increasing corpus size as the growth rate of "meaningful" tokens N_m is linear to the corpus size, whereas the growth rate of vocabulary size $|V|$ drops off for larger and larger corpora. As a result, the relative vocabulary size is almost zero for very large corpora. The vocabulary dispersion D_{Voc} also decreases with increasing corpus size, but with a lower rate than R_{Voc}. Its functional form is almost "s-shaped". This may be because the growth rate of low frequency terms $|V_{\text{low}}|$, e.g. spelling errors, is typically smaller then the growth of vocabulary size $|V|$. However, $|V|$'s growth rate drops off for larger and larger corpora and thus vocabulary dispersion decreases non-linearly. As text in a corpus typically follows language-internal rules, e.g. a grammar, and the vocabulary size $|V|$'s growth rate is smaller than the number of tokens N's growth rate, the number of possible term combinations is limited. Consequently, corpus predictability CP increases with increasing corpus size.

Coming back to our initial questions, we conclude: Vocabulary concentration, grammatical complexity and average sentence length are not corpus size-dependent given a sufficiently large corpus, i.e. more than 1000 sentences in our case. They may reliably be used to compare both same- and different-sized corpora. In contrast, entropy, relative vocabulary size, vocabulary dispersion and corpus predictability are corpus size-dependent and thus may *not* be reliably used to compare different-sized corpora.

To still compare different-sized corpora based on entropy, relative vocabulary size, vocabulary dispersion and corpus predictability, we suggest to *(under)sample* corpora to a common size and then apply the aforementioned textual characteristics. Alternatively, vocabulary dispersion may be used cautiously for corpora of a similar size and instead of entropy H we might calculate the *entropy rate* H_{rate} as shown in Formula 9:

$$H_{\text{rate}} = -\frac{1}{N} \sum_{t_i \in V} p(t_i) \log_{|V|} p(t_i) \tag{9}$$

Additionally to Formula 1's normalization to $|V|$, Formula 9 is also normalized to the number of tokens N and converges for $N \to \infty$ (Manning and Schütze, 1999).

4. Conclusions & Future Work

We studied the behavior of 7 textual characteristics for different-sized corpora in 4 genres. Although the actual textual characteristics vary from genre to genre as expected, we have shown their behavior for different-sized corpora is very similar across all 4 genres. We observed vocabulary concentration, grammatical complexity and average sentence length are not corpus size-dependent, whereas entropy, relative vocabulary size, vocabulary dispersion and corpus predictability are. Therefore, we suggest the former may reliably used to compare both same- and different-sized corpora and the latter may only be used to compare same- or similar-sized corpora.

Future research avenues include exploring the possibilities of fitting appropriate functions to the textual characteristics curves in order to interpolate between different-sized corpora and thereby avoid sampling. Additionally, we like to extend our study to more genres, e.g. novels, scientific essays, tweets and blog posts.

5. References

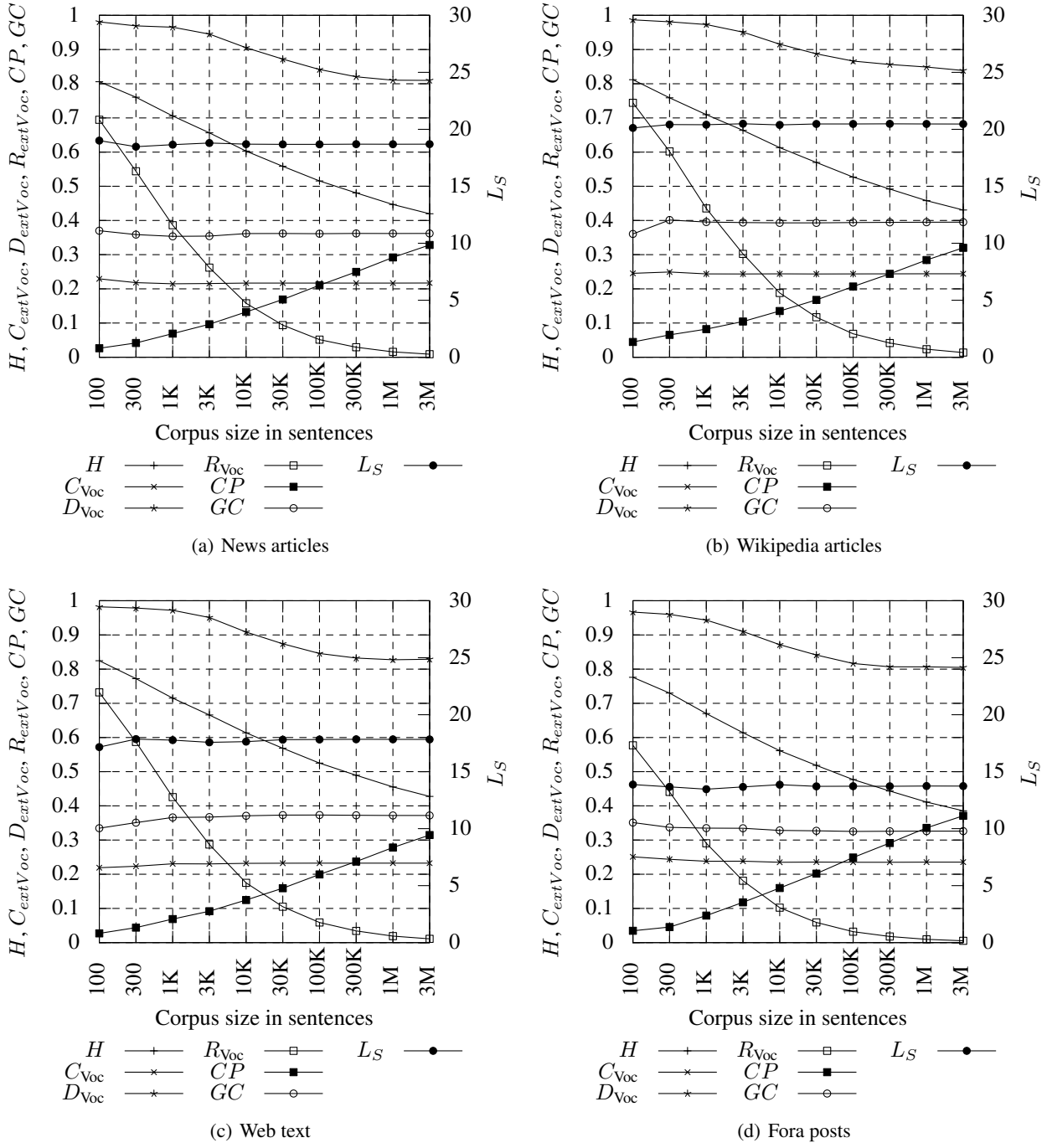
A. Aue and M. Gamon. 2005. Customizing Sentiment Classifiers to New Domains: a Case Study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

M. Bank, R. Remus, and M. Schierle. 2012. Textual Characteristics for Language Engineering. In *8th International Conference on Language Resources and Evaluation (LREC'12), to appear*.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447.

T. Eckart and U. Quasthoff. 2010. Statistical Corpus and Language Comparison using Comparable Corpora. In *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 15–20.

Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation Systems. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 172–180.



(a) News articles

(b) Wikipedia articles

(c) Web text

(d) Fora posts

Figure 1: Behavior of textual characteristics of English-language corpora of increasing size.

A. Ghose and P. Ipeirotis. 2011. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23:1498–1512.

L. Goeuriot, J.C. Na, W. Y. Min Kyaing, S. Foo, C. Khoo, Y.-L. Theng, and Y.K. Chang. 2011. Textual and Informational Characteristics of Health-related Social Media Content: A Study of Drug Review Forums. In *Proceedings of Asia-Pacific Conference on Library & Information Education & Practice (A-LIEP)*.

A. Kilgarriff. 2001. Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

C.D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

J.C. Na, T.T. Thet, and C. Khoo. 2010. Comparing Sentiment Expression in Movie Reviews from four online Genres. *Online Information Review*, 34(2):317–338.

Uwe Quasthoff and Thomas Eckart. 2009. Corpus Building Process of the Project "Deutscher Wortschatz". In *Proeedings of the Workshop on Lingustic Processing Pipelines*.

U. Quasthoff, M. Richter, and C. Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. In *Proceed-*

Corpus	Tokens	Types	H	R_{Voc}	C_{Voc}	D_{Voc}	CP	GC	L_S
news_100	1,866	946	0.8060	0.6946	0.2294	0.9810	0.0264	0.3700	19.01
news_300	5,428	2,173	0.7604	0.5439	0.2181	0.9692	0.0417	0.3587	18.48
news_1K	18,271	5,212	0.7058	0.3861	0.2149	0.9653	0.0691	0.3535	18.65
news_3K	55,178	10,676	0.6564	0.2620	0.2156	0.9452	0.0965	0.3543	18.80
news_10K	182,943	21,196	0.6028	0.1578	0.2170	0.9054	0.1324	0.3617	18.70
news_30K	548,395	37,806	0.5584	0.0939	0.2169	0.8720	0.1689	0.3619	18.68
news_100K	1,828,304	69,172	0.5157	0.0515	0.2168	0.8417	0.2106	0.3613	18.68
news_300K	5,491,076	117,964	0.4806	0.0293	0.2169	0.8203	0.2497	0.3622	18.70
news_1M	18,296,680	211,254	0.4468	0.0157	0.2170	0.8103	0.2919	0.3619	18.69
news_3M	54,903,309	357,955	0.4196	0.0089	0.2170	0.8098	0.3283	0.3620	18.70
wikipedia_100	1,947	1,065	0.8117	0.7442	0.2455	0.9869	0.0446	0.3606	20.12
wikipedia_300	5,943	2,553	0.7588	0.6020	0.2494	0.9812	0.0653	0.4013	20.42
wikipedia_1K	19,814	6,182	0.7096	0.4354	0.2438	0.9731	0.0823	0.3954	20.40
wikipedia_3K	59,699	12,944	0.6636	0.3022	0.2435	0.9509	0.1044	0.3940	20.49
wikipedia_10K	198,069	26,724	0.6130	0.1880	0.2440	0.9157	0.1353	0.3931	20.39
wikipedia_30K	597,049	50,208	0.5697	0.1171	0.2435	0.8873	0.1677	0.3930	20.40
wikipedia_100K	1,990,411	97,721	0.5269	0.0684	0.2437	0.8663	0.2068	0.3940	20.47
wikipedia_300K	5,975,787	177,640	0.4919	0.0415	0.2439	0.8563	0.2439	0.3945	20.48
wikipedia_1M	19,910,567	335,409	0.4580	0.0235	0.2441	0.8488	0.2844	0.3949	20.47
wikipedia_3M	59,719,241	588,673	0.4306	0.0138	0.2441	0.8386	0.3200	0.3950	20.47
web_100	1,643	901	0.8246	0.7319	0.2191	0.9822	0.0270	0.3347	17.16
web_300	5,192	2,256	0.7722	0.5870	0.2234	0.9787	0.0438	0.3510	17.86
web_1K	17,286	5,386	0.7152	0.4257	0.2305	0.9720	0.0689	0.3663	17.77
web_3K	51,253	10,754	0.6658	0.2868	0.2302	0.9509	0.0918	0.3671	17.57
web_10K	171,432	21,867	0.6136	0.1749	0.2320	0.9084	0.1247	0.3710	17.64
web_30K	519,253	39,720	0.5686	0.1050	0.2323	0.8741	0.1591	0.3730	17.81
web_100K	1,732,458	74,203	0.5251	0.0588	0.2325	0.8457	0.1994	0.3731	17.83
web_300K	5,204,182	129,709	0.4896	0.0342	0.2325	0.8324	0.2370	0.3726	17.84
web_1M	17,343,098	240,133	0.4554	0.0190	0.2324	0.8274	0.2783	0.3719	17.84
web_3M	52,014,020	421,318	0.4277	0.0111	0.2324	0.8288	0.3147	0.3720	17.83
fora_posts_100	1,339	572	0.7761	0.5772	0.2509	0.9668	0.0346	0.3512	13.87
fora_posts_300	3,958	1,305	0.7299	0.4409	0.2438	0.9602	0.0457	0.3372	13.67
fora_posts_1K	13,035	2,838	0.6699	0.2906	0.2380	0.9433	0.0791	0.3349	13.46
fora_posts_3K	39,714	5,383	0.6140	0.1809	0.2382	0.9099	0.1179	0.3346	13.66
fora_posts_10K	134,078	10,314	0.5614	0.1022	0.2351	0.8714	0.1596	0.3282	13.86
fora_posts_30K	397,647	17,729	0.5185	0.0592	0.2356	0.8412	0.2019	0.3271	13.71
fora_posts_100K	1,327,165	32,014	0.4769	0.0320	0.2350	0.8165	0.2489	0.3252	13.73
fora_posts_300K	3,979,848	53,994	0.4433	0.0180	0.2351	0.8068	0.2914	0.3258	13.72
fora_posts_1M	13,290,428	96,495	0.4112	0.0096	0.2354	0.8065	0.3355	0.3261	13.74
fora_posts_3M	39,851,933	160,608	0.3856	0.0053	0.2354	0.8050	0.3710	0.3262	13.74

Table 1: Textual characteristics of English-language corpora of increasing size.

ings of the 5th International Conference on Language Resources and Evaluation (LREC'06), pages 1799–1802.

P. Rayson and R. Garside. 2000. Comparing Corpora using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6.

S. Sekine. 1997. The Domain Dependence of Parsing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*, pages 96–102.

T. Suzuki and K. Kageura. 2007. Exploring the Microscopic Textual Characteristics of Japanese Prime Ministers' Diet Addresses by Measuring the Quantity and Diversity of Nouns. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 459–470.

Marie Těšitelová. 1992. *Quantitive Linguistics*. John Benjamins Publishing Company.

K. Verspoor, K.B. Cohen, and L. Hunter. 2009. The Textual Characteristics of Traditional and Open Access Scientific Journals are Similar. *BMC bioinformatics*, 10(1):183.

D. Wang and Y. Liu. 2011. A Cross-corpus Study of Unsupervised Subjectivity Identification based on Calibrated EM. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 161–167.