Collecting and Using Comparable Corpora for Statistical Machine Translation

Inguna Skadiņa[1], Ahmet Aker[2], Nikos Mastropavlos[3], Fangzhong Su[4], Dan Tufis[5], Mateja Verlic[6], Andrejs Vasiļjevs[1], Bogdan Babych[4], Paul Clough, Robert Gaizauskas[2], Nikos Glaros[3], Monica Lestari Paramita[2], Mārcis Pinnis[1]

Tilde[1], University of Sheffield[2], Institute for Language and Speech Processing[3], University of Leeds[4], Research Institute for Artificial Intelligence, Romanian Academy [5], Zemanta[6]

E-mail: accurat@tilde.lv

Abstract

Lack of sufficient parallel data for many languages and domains is currently one of the major obstacles to further advancement of automated translation. The ACCURAT project is addressing this issue by researching methods how to improve machine translation systems by using comparable corpora. In this paper we present tools and techniques developed in the ACCURAT project that allow additional data needed for statistical machine translation to be extracted from comparable corpora. We present methods and tools for acquisition of comparable corpora from the Web and other sources, for evaluation of the comparability of collected corpora, for multi-level alignment of comparable corpora and for extraction of lexical and terminological data for machine translation. Finally, we present initial evaluation results on the utility of collected corpora in domain-adapted machine translation and real-life applications.

Keywords: comparable corpora, under-resourced languages, machine translation

1. Introduction

In recent decades data-driven approaches have led to significant advances in machine translation (MT). However, the applicability of current data-driven methods depends on the availability of very large quantities of parallel data.

The problem of availability of such linguistic resources is especially acute for under-resourced languages and narrow domains. For many languages only a few parallel corpora of reasonable size are available. Statistical machine translation (SMT) systems trained on these corpora perform well on texts which are from the same domain, but are almost unusable for other domains.

At the same time, for many languages multilingual resources, such as news feeds or multilingual Web pages, which share a lot of common paragraphs, sentences, phrases, terms and named entities, are widely available. These data extracted from comparable resources (corpora) can be useful for both statistical and rule-based MT.

A comparable corpus is a relatively recent concept in MT. While methods on how to use parallel corpora in MT are well studied (e.g. Koehn, 2010), methods and techniques for comparable corpora have not been thoroughly investigated. However, latest research has shown that adding extracted parallel lexical data from comparable corpora to the training data of an SMT system improves the system's performance by reducing the number of un-translated words (Hewavitharana and Vogel, 2008). It has been also demonstrated that language pairs and domains with little parallel data can benefit from usage of comparable corpora (Munteanu and Marcu, 2005; Lu et al., 2010; Abdul-Rauf and Schwenk, 2009 and 2011).

Methods and techniques that exploit multilingual comparable corpora to overcome the bottleneck of insufficient parallel data for under-resourced languages are researched in the FP7 project ACCURAT – *Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation* (Skadiņa et al., 2010a). The project aims to find, analyse and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources and improve MT quality for under-resourced languages and narrow domains. The ACCURAT particularly targets a number of under-resourced languages, such as Croatian (HR), Estonian (ET), Greek (EL), Latvian (LV), Lithuanian (LT) and Romanian (RO), as well as evaluates applicability of comparable corpora for adapting MT to specific narrow domains.

In this paper we present tools and techniques developed in the ACCURAT project that allow the additional data needed for SMT to be extracted from comparable corpora. We present methods and tools for:

- acquisition of comparable corpora from the Web,
- evaluation of the comparability of collected corpora,
- multi-level alignment and extraction of lexical and terminological data from collected corpora for improvement of machine translation.

We also present an initial evaluation of the utility of collected corpora in applications.

2. Tools for building comparable corpora

Comparable corpora are potentially easier to build than parallel corpora for a large variety of languages and for many specific thematic areas. Although bilingual corpora can be comparable at a variety of levels and in various aspects, they are only able to improve MT system performance when they contain a good number of parallel textual segments. Therefore, we focus on gathering bilingual comparable text corpora containing a significant amount of mappable textual data.

We have investigated efficient methods and developed tools for identifying and gathering large amounts of comparable textual data from the Web for three different types of comparable corpora: (1) corpora consisting of news articles published concurrently; (2) inter-language linked Wikipedia articles and (3) corpora that cover

domain specific language.

2.1 Collecting news corpora

Various attempts at gathering comparable corpora from the Web have been made (e.g., Braschler and Schäuble, 1998; Huang et al., 2010; Talvensaari et al., 2008). The process of obtaining such corpora involves (1) downloading for each language a separate set of documents and (2) matching documents between two languages by comparing the document contents. Useful units for SMT are then extracted from the document pairs by applying extraction methods, such as those described in Section 4 below. To address steps 1 and 2 we developed a novel approach for gathering comparable news texts that uses document titles as surrogates for full document content, motivated by the observation that news titles are a good indicator for the content of the news document (Edmundson, 1969, Lopez et al., 2011). This approach massively reduces processing and data storage requirements as compared with approaches that require full document download. Our tools iteratively download, separately for each language, current news article titles using Google News Search and RSS News feeds. For each language the downloaded titles are split into different bins based on their publication dates, so that each bin contains titles for the same week. For each language pair we take the titles from the corresponding weeks' bins and pair them using different heuristics such as cosine similarity, title length difference, publication date difference. We download the full article contents only of the "good" pairs. Before computing the cosine similarity between the titles in the source and target languages we translate the target titles into the source language using existing MT systems. We also remove stop-words from both titles and perform cosine similarity on the remaining content words.

Working only with titles reduces costs measured in hard disk space and computational power, and also reduces noise in the pairing process by limiting search space to one week. We describe our methodology in more detail in Aker et al. (2012). The comparable news corpora collected using this approach over 10 weeks between 01/12/11 and 12/02/12 are detailed in Table 1.

2.2 Collecting Wikipedia documents

Wikipedia has been viewed as a source of comparable documents due to the existence of inter-language links, which connect Wikipedia articles on the same topic, but written in different languages. However, Filatova (2009) found that these articles may not be comparable to each other; in some cases they may even be contradictory. Therefore, a method is required to filter out such non-comparable documents. We developed an approach to measure the similarity of document pairs by performing cross-lingual sentence alignment.

Our approach, which is language-independent, is based on the method proposed by Adafre & de Rijke (2006). This approach uses anchor text information from the Wikipedia articles to identify parallel sentences. First, a bilingual lexicon is constructed by extracting all document titles which are connected by the Wikipedia inter-language links. This lexicon is then used to translate all anchor texts found in the non-English articles into English. We then calculate the Jaccard coefficient to measure the similarity of sentences, pairing each sentence in the shorter document to the highest scoring sentence in the longer document. Finally, a measure of document-level similarity is computed based on averaging the scores of the paired sentences. Document pairs whose scores fall above a pre-defined minimum threshold are considered to be comparable; those below are filtered out.

We find this approach correlates with human cross-language similarity judgments (more details can be found in Paramita et al., 2012). Comparable data collected using this method are described in Table 1.

Lang Pair (Source – Target)	News Corpora			Wikipedia Corpora		
	# Doc Pair	# Words (Source)	# Words (Target)	# Doc Pair	# Words (Source)	# Words (Target)
EN-SL	33,561	187 K	284 K	20,351	15 M	2.6 M
EN-RO	30,761	307K	207K	48,880	27.2 M	4.8 M
EN-LV	39,942	316 K	138 K	4,273	5.9 M	627 K
EN-LT	38,878	289 K	240 K	10,308	10.6 M	1.5 M
EN-HR	19,246	238 K	193 K	14,147	13.7 M	3.4 M
EN-ET	16,144	192 K	133 K	14,112	16.8 M	1.7 M
EN-EL	76,838	450 K	265 K	3,668	4 M	1.1 M
EN-DE	129,341	840 K	510K	149,891	66.7 M	52.9 M

Table 1: News and Wikipedia comparable corpora collected.

Narrow Domain \ Language	EN	LV	LT	HR	EL	RO	DE
Renewable Energy	18.92	0.45	0.61	0.44	0.91	1.14	-
Topical News Political	24.93	6.61	1.57	9.65	25.80	5.43	-
Topical News Sports	8.81	2.57	3.48	4.51	13.57	3.96	-
Topical News Technological	25.77	3.12	2.56	-	10.29	4.05	-
Topical News Disasters	25.80	1.81	2.74	2.37	8.75	4.37	-
Automotive engineering	6.12	-	-	-	-	-	8.28
Assistive technologies	29.14	-	-	-	-	-	38.65
Software localization	4.20	1.54	-	-	-	-	-

Table 2: Narrow domain comparable corpora collected (token counts in millions).

2.3 Collecting domain specific corpora

For collecting domain-specific corpora from the Web, a highly configurable Focused Monolingual Crawler (FMC) has been developed, based on the Bixo[3] open-source Web mining toolkit. Given a narrow domain (topic) and a language, FMC has to be fed with two input datasets: (i) a list of topic definition multi-word term expressions and (ii) a list of topic-related seed URLs. The user can configure FMC in a variety of ways, e.g. set file types to download, domain filtering options, self-terminating conditions, crawling politeness parameters, etc..

Crawling starts from the seed URLs and expands dynamically to other URLs, while lightweight text classification is performed on the Web pages being visited, so as to retrieve only those Web documents that are relevant to the chosen topic. Operations such as boilerplate removal, text normalisation and cleaning, language identification, etc. are done during runtime; post-crawling processing steps (including de-duplication, post-classification and filtering) are also implemented.

The FMC output consists of the collected Web documents in both HTML and text format as well as their metadata. The metadata are stored in XML using a cesDOC format that can be validated against XCES standard schemas.

To collect a pair of bilingual comparable corpora two separate crawls are required (one per language). The comparability of the bi(multi)lingual documents retrieved is achieved by ensuring that, for each language, the FMC tool is made to return Web documents that are close to the same topic.

By using FMC, 28 comparable corpora on 8 narrow domains and in 6 language pairs (EN-LV, EN-LT, EN-HR, EN-RO, EN-EL and EN-DE) amounting to a total of more than 148M tokens have been constructed. Corpora-specific information is given in Table 2.

3. Evaluation of comparable corpora and the comparability metric

Successful detection of translation equivalents from comparable corpora very much depends on the quality of these corpora, specifically – on the degree of their textual equivalence and successful alignment on various text units. Thus a metric for measuring comparability of pairs of documents in different languages performs two main functions:

- *evaluates the quality* of the collected comparable corpora,
- *enhances* the corpora by ranking pairs of documents by their comparability, which indicates the likelihood of retrieving good-quality translation equivalents from the aligned document pairs.

Evaluation of the quality of automatically collected corpora characterises them in terms of broad comparability categories. These categories are defined by the provenance and alignability of the collected text units, i.e.,

- parallel texts, which can be aligned at the word level,
- strongly comparable texts, which describe the same event or a phenomenon and can be aligned on the document level,
- weakly comparable corpora, which are within the same narrow domain and can only be aligned on the corpus level.

These categories were calibrated on a smaller manually collected set of bilingual corpora, the Initial Comparable Corpora (ICC) (Skadiņa et al, 2010b) which includes corpora for 10 language pairs: ET-EN, LV-EN, LT-EN, EL-EN, RO-EL, HR-EN, RO-EN, RO-DE, LV-LT and SL-EN. Every corpus, except RO-EL and LV-LT, consists of approximately one million words for under-resourced language. Taken together the collected corpora consist of 12.5 million words for Croatian, Estonian, Greek, Latvian, Lithuanian, Romanian and Slovenian. ICC also include narrow domain EN-DE corpora for automotive, medicine, assistive technology and software domains.

There have been several studies dealing with comparability for comparable corpora. Some studies (Sharoff, 2007; Maia, 2003; McEnery and Xiao, 2007, Resnik and Smith, 2003) analyse comparability by assessing corpus composition, such as structural criteria (e.g., format and size), and linguistic criteria (e.g., topic, domain, and genre). Munteanu and Marcu (2005) rank comparability of article pairs by making use of a cross-lingual information retrieval approach and the information of article publication date. Smith et al. (2010) use "interwiki" links to identify aligned comparable

[3] http://bixo.101tec.com

Wikipedia documents by treating Wikipedia as a comparable corpus. Li and Gaussier (2010) determine corpus comparability by measuring the proportion of lexical overlapping between comparable documents with bilingual dictionary.

In contrast to the above previous work, our comparability metric takes several features into account, including lexical information and document structure. These features are then combined via a weighted average strategy and compared in terms of cosine similarity between the feature vectors (Su and Babych, 2012), resulting in a comparability score in the range [0,1], with higher values corresponding to greater comparability. For the calibration task (Table 3) we experimentally established that the combination of these internal features could accurately predict comparability categories as externally defined by human judgment.

More specifically, we mapped the categories into a numeric scale: Parallel = 3; Strongly comparable = 2; Weakly comparable = 1. Then we computed the Pearson's *r* correlation between the range of these calibrated values and the average comparability scores of the corresponding comparability levels for nine of the language pairs in the ICC corpora. The results are presented in Table 3.

Language pair	Parallel	Strongly Comparable	Weakly comparable	Pearson's r correlation
DE-EN	0.912	0.622	0.326	**0.99998**
EL-EN	0.841	0.635	0.250	**0.98505**
ET-EN	0.765	0.547	0.310	**0.99971**
LT-EN	0.755	0.613	0.308	**0.97855**
LV-EN	0.770	0.627	0.236	**0.96588**
RO-EN	0.782	0.614	0.311	**0.98658**
SL-EN	0.779	0.582	0.373	**0.99985**
EL-RO	0.863	0.446	0.214	**0.98672**
RO-DE	0.717	0.573	0.469	**0.99569**

Table 3: Average metric values for ICC comparability categories and their correlation with the numeric values for categories.

It can be seen from Table 3 that there is a strong positive correlation between the values predicted by the metric and the ICC manual annotation values, which indicates a strong link between our internally identified textual features and the external provenance and alignability of comparable corpora.

However, note that the absolute values of the comparability scores substantially vary for different language pairs. Normally these values depend on the quality of mapping resources (the size of bilingual dictionaries), and also – for non-lemmatised texts – the degree of morphological variation in source and target languages, which can be measured as data sparseness, e.g., the type/token ratio, of the source and target languages. It is difficult to predict what the effect of the combination of these parameters will be on the absolute value of the comparability scores. Therefore, if we need to predict the comparability labels of a new set of documents for a new language pair, then the absolute values of the comparability score need to be calibrated on some annotated resource, such as the ICC. But when calibration is completed the prediction of comparability categories can become very accurate.

Current experiments have focused on the total group of documents labelled as Parallel, Strongly and Weakly comparable in the ICC. In future experiments we will also address the relation between the size of the corpus to be labelled and the accuracy of the prediction. We expect that with smaller size the correlation will go down, and there will be a certain minimum corpus size for which the accuracy of the prediction will be sufficiently accurate.

After calibration we applied the comparability metric to the news corpus described above for the same language pairs, which allows us to understand the nature of the collected documents in terms of comparability categories and estimate their usefulness for extraction of translation equivalents.

Application of the metric also enhances the collected comparable corpora by ranking pairs of documents across languages by their comparability scores making application of phrase alignment tools much more efficient by allowing them to focus on high ranking pairs.

4. Extraction of MT-related data from comparable corpora

By "MT-related data" extracted from comparable corpora we understand collections of translation equivalent chunks of text. Such a chunk may contain a pair of terminological expressions, a pair of named entities, a pair of regular phrases or even a pair of sentences or paragraphs. Parallel data that may be extracted from comparable corpora differ both in quantity and quality depending on the comparability degree of the documents in the corpora.

4.1 Extraction of named entities and terminological units

Even weakly comparable corpora can contain useful translation equivalences for named entities or terminological units. The ACCURAT project has developed tools for extraction of such translation equivalents[4].

[4] These tools and documentation (ACCURAT D2.6) can be downloaded from http://www.accurat-project.eu.

Our general approach is to tag named entities and terms monolingually and then to map them cross-lingually. For monolingual named entity recognition (NER) and term extraction new tools were developed for project's languages that did not have such tools available before. For NER, for instance, TildeNER (Pinnis, 2012) for Latvian and Lithuanian and NERA1 for Romanian were developed. For term extraction, for example, a tool named CollTerm was adapted for Latvian and Lithuanian (initially developed for Croatian) as well as a language specific term extraction tool was developed for Romanian. As CollTerm is language independent, it was used also for English term extraction. For other languages existing tools were used, for instance, OpenNLP[6] for English NER.

The monolingual terms and named entities were cross-lingually mapped using GIZA++ dictionaries and various string-similarity measures (Ştefănescu, 2012).

Identifying corresponding named entities in different languages works reasonably well. However, this is not the case for mapping technical terms. Term mapping is more challenging, because many single word terms cannot be mapped using string-similarity measures (e.g., "*computer*" in English and "*dators*" in Latvian). In the case if the dictionaries do not contain such terms, mapping becomes difficult. This is true for under-resourced languages and new domains. The string similarity measures are also highly affected by language specific compounding rules and the morphological characteristics of both languages.

Due to the lack of terminological and name-entity gold-standards for the project's language pairs, we could not fully evaluate the performance of our tools (the Recall and F-measure). Instead, we manually evaluated the precision of the extracted bilingual named entities from randomly selected 100 pairs of documents. The results are summarized in Table 4 below.

Lang. Pair	Correct	Partially Correct	Incor-rect	Total	Cor-rect in %
EN-LV	49	1	4	54	90%
EN-LT	80	20	5	105	76%
EN-RO	113	4	4	121	93%
EN-DE	141	11	7	159	88%
EN-EL	60	6	0	66	90%
EN-HR	59	51	4	114	51%

Table 4: NE mapping evaluation.

The partially correct mappings (column 3) refer to cases where some parts of a named entity were missing or falsely added in the mapping (for instance, a person's first name mapped to the first name and surname).

For term mapping a similar experiment to evaluate precision of the developed methods was applied on slightly larger comparable disaster news corpora. For instance, for English-Latvian the corpus consists of 2911 document pairs and the mapper achieves a precision of 85.89% with 489 term pairs extracted.

[6] http://incubator.apache.org/opennlp/index.html

4.2 Extraction of parallel phrases and sentences

The extraction of chunks of parallel phrases and sentences from comparable corpora is a more difficult task than monolingually extracting named entities and terms and bilingually mapping them afterwards.

The usual sentence alignment techniques applicable for parallel corpora rely on a fundamental property: the translation equivalent paragraphs (and to a large extent, sentences) have the same order in the two parts of the bitext. This property, which significantly reduces the alignment search space, is not valid anymore in comparable corpora. Given this limitation of a comparable corpus in general and the sizes of the comparable corpora that we will have to deal with in particular, we have devised a variant of an Expectation Maximization (EM) algorithm (Dempster et al., 1977) that generates document alignment from comparable corpus using only pre-existing translation lexicons. The EMACC (Expectation Maximization Alignment for Comparable Corpora) tool (Ion et al, 2011) allows alignment of different types of textual units: documents, paragraphs, and sentences. The document alignment evaluation experiments performed on the previously mentioned ICC corpora showed that for the parallel documents EMACC provided almost perfect results for the languages of the project paired to English. While for EN-RO, EN-EL and EN-LV the alignment was perfect (P=100%, R=100%), the lowest results were obtained for EN-SL (P=89.1%, R=89.1) and the rest was in the range of 91-97%.

For strongly comparable corpora, EMACC achieved also high precision and recall (between 72-83%.): the highest result was obtained for RO-EN pair (P=85.7%, R=85.7%) and the lowest performance for EN-ET pair (P=55.2, R=55.2%).

For weakly comparable corpora the results were much more dispersed among languages with the highest scores for EN-RO (P=66.2%, R=66.2%) and the lowest scores for EN-EL (P=7.7%, R=7.7%).

The phrase extraction algorithm from comparable corpora (PEXACC) is the next module in the processing chain. Similarly to EMACC, this program has been designed and implemented with the main emphasis on weakly comparable documents. The workflow of PEXACC is described in details in Ion (2012).

In order to evaluate the performance of the PEXACC algorithm we needed a gold standard comparable corpus with the parallel sentences marked-up. As such a corpus does not exist, we constructed various artificial comparable corpora for which we knew the result of perfect extraction: starting with sentence aligned parallel corpora, we inserted in each part of the bitext, at random positions, arbitrary sentences in the respective languages. The experiments conducted involved a number of added sentences that was first equal to and then double the number of initial parallel sentences. Thus we created for three language pairs (EN-RO, EN-LV, and EN-ET) two controlled test comparable corpora with noise ratio (non-parallel sentences / parallel sentences) of 1 and 2.

We ran the PEXACC extraction algorithm with three relevance feedback loops on the artificially created noisy comparable corpora. The major observation in these experiments was that the recall of the PEXACC algorithm is relatively insensitive to the level of noise. For the bilingual corpora with noise ratio of 2 the extraction precision/recall varied between 95.8% / 80.5% (EN-RO) and 97.4% / 21.4%, depending on the threshold value for the parallelism score. The downside of this algorithm is its high computing time. We use a cluster with a total of 56 CPU cores (4 nodes) with 6-8 GB of RAM per node and, with this configuration, the total running time is between 8 h and 30 h per language pair (about 2000 documents per language), depending on the setting of the various parameters.

Recently we have developed LEXACC, a new and much more effective extraction algorithm (Ștefănescu et al., 2012), which reuses the similarity measure of PEXACC but replaces the brute force search (analysis of all the sentence pairs in the Cartesian product of sentences contained by a comparable document pair) with a CLIR technique. The idea is relatively simple: for each sentence in the source corpus, the content words are selected and translated into the target language (using available translation tables; (Ștefănescu et al., 2012) used the GIZA++ tables). The translated content words are used to form a Boolean query for a search engine (Lucene) for which the target corpus has been indexed at the sentence level. The speed of the extraction process has been increased more than 1200 times with performance slightly improved over PEXACC.

Besides full sentences, LEXACC may extract sub-sentential fragments as well. In this case the size of extracted data is significantly larger. Because manual validation is a very time consuming task, we restricted ourselves only to parallel sentence pair evaluation. The quantity of (quasi-)parallel sentences extracted from the news corpora is shown in the Table 5 for the language pairs we manually checked the data.

Lang. Pair	Size of comparable corpora (MB)	# Extracted sentence pairs/size (MB)	Confidence threshold	Precision
EN-LV	78.46	4781 /1.24	0.35	85%
EN-LT	76.34	1794 /0.55	0.35	85%
EN-ET	35.77	542 /0.17	0.35	85%
EN-RO	71	2019 / 0.6	0.45	93%
EN-SL	22.3	930 /1.3	0.25	84%

Table 5: Parallel sentences extracted by LEXACC from the ACCURAT News Comparable corpora.

The parallel sentence extractor is a parameterized tool, being tuned for each language pair in the project. The user may decide to re-estimate these parameters and change the extraction confidence thresholds. Because of different comparability degrees of the collected comparable corpora for the considered language pairs, the confidence thresholds, the outcome and its precision are different from language pair to language pair.

The harvesting of comparable corpora is an on-going process as is the extraction of parallel data. Therefore data in Table 5 reflects the status at the writing time of this article. The project has 4 months more to go and we estimate that the final figures will be substantially higher.

5. Comparable corpora in MT applications

In order to evaluate the impact of data extracted from comparable corpora on machine translation performance, several experiments have been carried out for narrow domains. These experiments aim to evaluate translation quality of MT elaborated with data from comparable corpora and to assess usability of MT in real life scenarios.

To test the quality and effect of the data extracted with ACCURAT tools, we ran an experiment with EN-DE domain-adapted SMT for the automotive industry domain.

The baseline system for this experiment was trained on the Europarl (Koehn, 2005) and news-commentary corpora[9]. These corpora were used for both translation and language models. For the adapted system, an additional language model was trained on the data extracted from automotive domain comparable corpus with LEXACC tool described in Section 4.2. In total 45,952 sentence pairs were obtained from a comparable corpus of about 3.5 million lines.

All corpora were aligned using GIZA++ (Och & Ney, 2000), the language models were trained using SRILM (Stolcke, 2002) and the MT systems were trained using the Moses SMT Toolkit (Koehn et al., 2007). Tuning via MERT was performed on a domain-specific development set. For testing text from the automotive domain was used. The translations were evaluated using BLEU (Papineni et al., 2002) and presented in Table 6.

System	BLEU
Baseline	18.81
Automotive extracted	25.44

Table 6: Evaluation of narrow domain SMT system enriched with data from comparable corpus.

As Table 6 shows, with the extracted data, it is possible to gain about 6.5 BLEU points over the baseline system. This means that the data LEXACC extracts is of high enough quality to be useful for SMT purposes, as the noise is filtered out during the training phase.

The second task is to assess the usability and translation quality of MT in real life scenarios. We evaluated influence of MT on precision of recommendations provided by Zemanta's Authoring assistant tool for bloggers [10] using ACCURAT baseline SMT systems trained on publicly available parallel corpora. Human

[9] http://www.statmt.org/wmt11/translation-task.html#download

[10] http://www.zemanta.com

evaluation results for recommendations obtained for 100 Slovenian and German documents and corresponding machine translated English documents are summarized in the Table 7. Looking at the difference between the precision of related articles for texts in the original language and for the translations (11% for Slovenian and 20% for German) we can conclude that the MT solution improved the results obtained from recommendation engine.

Related article set	No. of related articles	Average precision (%)	NDCG score[12]
SL original	990	15.34	0.243
SL translated	988	**26.52**	0.422
DE original	1000	14.13	0.214
DE translated	1000	**34.35**	0.550

Table 7: Comparison of precision and NDCG score for sets of related articles.

As the next step we plan to evaluate Zemanta's Authoring assistant tool using ACCURAT MT systems enriched with the data extracted from the comparable corpora.

Conclusion

In this paper we presented a model for exploiting comparable corpora to increase MT quality for under resourced languages and narrow domains. We presented tools and resources for the collection, evaluation and alignment of comparable texts for application in machine translation.

MT-related data extracted from comparable corpora (parallel named entities pairs, parallel term pairs, parallel sub-sentential chunks and parallel sentences) can be reliably found even in weakly comparable corpora. Given that comparable corpora can be collected in very large quantities, even a few percentages of extracted MT-related data can provide a significant help in building or adapting an SMT system for which proper training parallel corpora cannot be easily found.

We also presented the initial results for the enhancement of a narrow domain SMT system with data extracted from comparable corpora and the application of SMT into a blog writer's recommendation system. Tools presented in the paper allow parallel data to be extracted from comparable corpora that can be used for SMT adaptation for particular domain.

Results achieved so far are promising in terms of collected data, precision of comparability metrics and alignment algorithms. Tools and resources described in this paper are publicly available from ACCURAT project website: www.accurat-project.eu.

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. Many thanks to colleagues in ACCURAT partner organizations: Sabine Hunsicker from DFKI (Germany), Gregor Thurmair from Linguatec (Germany) and Marko Tadić from University of Zagreb (Croatia).

[12] Normalized Discounted Cumulative Gain (NDCG) ranking score (Järvelin & Kekalainen, 2002)

References

Abdul-Rauf, S.; Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In: *EACL 2009: Proceedings of the 12th conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, pp 16–23.

Abdul-Rauf, S.; Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. In: *Machine Translation*, 25(4), pp.341--375.

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.

Adafre, S. F. and de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In Proceedings of the EACL Workshop on New Text, Trento, Italy.

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.

Bekavac, B. and Tadič, M. (2007). Implementation of croatian nerc system. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, Association for Computational Linguistics, pp. 11--18.

Braschler, M. and Schäuble, P. (1998). Multilingual Information Retrieval Based on Document Alignment Techniques. In *Research and advanced technology for digital libraries: second European conference*, *ECDL'98*, Heraklion, Crete, Cyprus, September 21-23, 1998, Springer Verlag, pp. 1998.

Dempster, A. P.; Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39(B):* pp. 1—38.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16:264–285.

Filatova, E. (2009). Directions for exploiting asymmetries in multilingual Wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3 '09).*

Hewavitharana, S. and Vogel, S. (2008). Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. In *Proceedings of the Workshop on Comparable Corpora*, LREC'08, pp. 7-10.

Huang, D.; Zhao, L.; Li, L. and Yu, H. (2010). Mining large-scale comparable corpora from Chinese-English

news collections. In *Proceedings of the 23rd International Conference on Computational Linguistics (Posters)*, pp. 472–480.

Ion, R.;Ceauşu A. and Irimia, E. (2011). An Expectation Maximization Algorithm for Textual Unit Alignment. In *Proceedings of 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pp. 128--135.

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

Järvelin, K. and Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems 20(4)*, pp. 422--446.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit* Phuket, Thailand, AAMT, pp. 79--86.

Koehn, P.; Hoang, H.; Birch, A., Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A. and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 177--180.

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

Li, B.; Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceeding of COLING 2010*, Beijing, China, pp. 644-652.

Lopez, C.; Prince, V. and Roche, M. (2011). Automatic titling of Articles Using Position and Statistical Information. In *RANLP, 2011*, pp. 727-732.

Lu, B.; Jiang, T.; Chow, K.; Tsou, B.K. (2010). Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. In: *Proceedings of the 3rd workshop on building and using comparable corpora: from parallel to non-parallel corpora*, Valletta, Malta, pp. 42–48.

Maia, B. (2003). What are comparable corpora? In *Proceedings of the Corpus Linguistics workshop on Multilingual Corpora: Linguistic requirements and technical perspectives*, 2003, Lancaster, U.K., pp. 27-34.

McEnery, A., Xiao, Z. (2007). Parallel and comparable corpora? In Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters, Clevedon, UK.

Munteanu, D. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4), pp. 477--504.

Munteanu, D. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from nonparallel corpora. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Morristown, NJ, USA, pp. 81–88.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 440--447.

Papineni, K.; Roukos S.; Ward T.; Zhu W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, pp. 890-899 .

Pinnis M. (2012). Latvian and Lithuanian Named Entity Recognition with TildeNER. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.

Resnik, P.; Smith, N. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. In *Proceedings of 3rd Web as Corpus Workshop*, Louvain-la-Neuve, Belgium

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.; 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.

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

Stolcke, A. (2002). SRILM: an extensible language modeling toolkit. In *Proceedings of the Int. Conf. on SpokenLanguage Processing*, pp. 901–904.

Su, F. and Babych, B. (2012). Development and application of a cross-language document comparability metric. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.

Talvensaari, T.; Pirkola, A.; Järvelin, K.; Juhola, M. and Laurikkala, J. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5), pp. 427–445.

Tufiş, D.; Ion, R.; Ceauşu, A. and Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, Trento, Italy, 3-7 April, 2006, pp. 153—160.