LREC 2016

11th Workshop on Building and Using Comparable Corpora

PROCEEDINGS

Edited by

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

ISBN: 979-10-95546-07-8
EAN: 9791095546078

8 May 2018

Proceedings of the 11th Workshop on
Building and Using Comparable Corpora, 8 May 2018 – LREC 2018, Miyazaki, Japan

Edited by Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

https://comparable.limsi.fr/bucc2018/

Acknowledgments: Part of this work was supported by a Marie Curie Career Integration Grant (MULTILEX) within the 7th European Community Framework Programme.

Organising Committee

- Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany), Chair
- Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France), Shared task organizer
- Serge Sharoff (University of Leeds, UK)

Programme Committee

- Ahmet Aker, University of Sheffield, UK
- Hervé Déjean, Xerox Research Centre Europe, Grenoble, France
- Éric Gaussier, Université Joseph Fourier, Grenoble, France
- Gregory Grefenstette, INRIA, Saclay, France
- Silvia Hansen-Schirra, University of Mainz, Germany
- Kyo Kageura, University of Tokyo, Japan
- Philippe Langlais, Université de Montrèal, Canada
- Shervin Malmasi, Harvard Medical School, Boston, MA, USA
- Michael Mohler, Language Computer Corp., US
- Emmanuel Morin, Université de Nantes, France
- Dragos Stefan Munteanu, Language Weaver, Inc., US
- Lene Offersgaard, University of Copenhagen, Denmark
- Ted Pedersen, University of Minnesota, Duluth, US
- Reinhard Rapp, Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany
- Serge Sharoff, University of Leeds, UK item Michel Simard, National Research Council Canada
- Pierre Zweigenbaum, LIMSI, CNRS, Université Paris-Saclay, Orsay, France

Preface – 11th BUCC at 11th LREC

In the language engineering and the linguistics communities, research on comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is primarily motivated by the need to use comparable corpora as training data for statistical Natural Language Processing applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

Comparable corpora are collections of documents that are comparable in content and form in various degrees and dimensions. This definition includes many types of parallel and non-parallel multilingual corpora, but also sets of monolingual corpora that are used for comparative purposes. Research on comparable corpora is active but used to be scattered among many workshops and conferences. The workshop series on "Building and Using Comparable Corpora" (BUCC) aims at promoting progress in this exciting emerging field by bundling its research, thereby making it more visible and giving it a better platform.

Following the ten previous editions of the workshop which took place in Africa (LREC'08 in Marrakech), America (ACL'11 in Portland and ACL'17 in Vancouver), Asia (ACL-IJCNLP'09 in Singapore and ACL-IJCNLP'15 in Beijing), Europe (LREC'10 in Malta, ACL'13 in Sofia, LREC'14 in Reykjavik and LREC'16 in Portoroz) and also on the border between Asia and Europe (LREC'12 in Istanbul), this year the 11th edition of the BUCC workshop is co-located with the 11th edition of the LREC conference in Miyazaki, Japan.

Given the hosting country and the impressive growth of Asian research in our field, this year the workshop's special theme is "Comparable Corpora for Asian Languages", and last year's shared task is continued and extended under the title "Identifying Parallel Sentences in Comparable Corpora". A major paradigm change in the field concerns the prevalence of Artificial Neural Networks, also appearing under the more catchy title of *Deep Learning*. Within the last five years, the Deep Learning methods shifted the balance in multilingual NLP processing towards less parallel and more comparable resources, e.g., by providing multilingual embedding spaces from monolingual corpora and by enabling Neural MT with minimal or no reliance on parallel data. Neural Networks finally make it possible to take long distance dependencies (e.g. between the words within a sentence) into account, thus overcoming a fundamental limitation of traditional n-gram-based approaches. The proceedings of this workshop present the new horizons for multilingual research with limited resources.

We would like to thank all people who in one way or another helped in making this workshop once again a success. Our special thanks go to Kyo Kageura and Yves Lepage for accepting to give invited presentations, to the members of the program committee who did an excellent job in reviewing the

submitted papers under strict time constraints, and to the LREC'18 workshop chairs and organizers for hosting the workshop. Last but not least we would like to thank our authors and the participants of the workshop.

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff May 2018

Programme

08:55–9:00 ***Opening Remarks***

Session 1: Invited Presentation

09:00–10:00 Kyo Kageura
Cross-lingual Correspondences of Terms in Texts and Terminologies: Theoretical Issues and Practical Implications

Session 2a: Applications of Comparable Corpora

10:00–10:30 Laurent Prévot, Matthieu Stali and Shu-Chuan Tseng
Grouping conversational markers across languages by exploiting large comparable corpora and unsupervised segmentation

10:30–11:00 ***Coffee Break***

Session 2b: Applications of Comparable Corpora

11:00–11:30 Bartholomäus Wloka
Identifying Bilingual Topics in Wikipedia for Efficient Parallel Corpus Extraction and Building Domain-Specific Glossaries for the Japanese-English Language Pair

11:30–12:00 Firas Sabbah and Ahmet Aker
Creating Comparable Corpora through Topic Mappings

12:00–12:30 Pierre Lison and A. Seza Doğruöz
Detecting Machine-translated Subtitles in Large Parallel Corpora

12:30–14:00 ***Lunch Break***

Session 3: Invited Presentation

14:00–15:00 Yves Lepage
Quasi-Parallel Corpora: Hallucinating Translations for the Chinese-Japanese Language Pair

Session 4a: Shared Task: Parallel Sentence Extraction

15:00–15:30 Pierre Zweigenbaum, Serge Sharoff and Reinhard Rapp
Overview of the Third BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora

15:30–16:00 Houda Bouamor and Hassan Sajjad
H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings

16:00–16:30 ***Coffee Break***

Session 4b: Shared Task: Parallel Sentence Extraction

16:30–17:00 Andoni Azpeitia, Thierry Etchegoyhen and Eva Martinez Garcia
Extracting Parallel Sentences from Comparable Corpora with STACC Variants

17:00–17:30 Chongman Leong, Derek F. Wong and Lidia S. Chao
UM-pAligner: Neural Network-Based Parallel Sentence Identification Model

17:30–17:35 ***Closing***

Table of Contents

Cross-lingual Correspondences of Terms in Texts and Terminologies: Theoretical Issues and Practical Implications

Kyo Kageura
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
kyo@p.u-tokyo.ac.jp

Abstract

Terms are items in language that represent concepts. This relation of representation does not change through use. As such, terms have a unique status in language, second only to proper names. Due to this, clarifying the identity of concepts represented by terms becomes an important issue at the level of what is represented, and control of terms representing the same concept also becomes an important issue at the level of representation. These problems with which terminologists are concerned, though not clear at first glance, are in fact relevant to general words and vocabulary to a lesser extent. In this paper I first clarify theoretical issues of terms and terminologies and what they imply for terminology processing in particular and lexical and lexicological processing in general. I then pick up some terminological applications, examine their status and suggest a few issues that can be addressed in terminology processing.

Keywords: Terminology, Concept, Comparability

1. Background

1.1. Concepts, Knowledge and Terminology

Let me start this paper with a rather theoretical discussions.

Forgeries do not destroy science. Science is destroyed when people, including "scientists," start regarding claims and "arguments" based on forged or fake data as part of science. That we can safely assume that the concept and act of science, in its proper sense, exists and is shared enables us, not only practically but also *logically*, to identify what are to be identified as forgeries as forgeries.

An argument homomorphic to this holds for the changes in the meaning of words in general. When we say a word changes its meaning in accordance with its use, we *logically* presuppose the existence of the *identity* of meaning of the word. Otherwise we cannot talk about the meaning or a meaning or meanings of a word in the first place. This logical identity indeed restricts the *practical* range of changes in the meaning of a word: whenever I have responded "oh, yes, the meaning of a word is sweet and tasty, but it's too expensive" to a person who has asserted that "the meaning of a word changes in use," they have always been puzzled. In other words, the meaning of a word does not change beyond a certain limit, which reflects, at least within a certain range of duration, the identity of the meaning. One can say with confidence that the meaning of a word changes as long as – and precisely because – the underlying identity of the meaning of a word remains intact.

While this *identity* of the meaning tends to work implicitly in the background in the case of general words, it is one of the main and explicit concerns for technical terms. Crudely speaking, it is this *identity* represented by a term that is referred to as a *concept*. Though it is not easy to recognise the essential difference between the relationship between concept and term on the one hand and the relationship between meaning and word on the other (Kageura, 1995), especially when terms and words are handled in practical setups as in compiling dictionaries or terminologies, there is a logical necessity for terminologists to talk about concepts represented by terms rather than meanings of terms.

What is more, this concept-term relation as distinct from meaning-word relation constitutes a part of the essential language infrastructure that supports social construction and organisation, and issues related to this relation can cause practical – and sometimes serious – problems in our social life. A while ago, when US-based insurance companies started operating in Japan, the difference in the definition of "cancer" caused trouble in the application of insurance policies[1]. As this is a cross-lingual case, it is easily noticed that the issue is not to do with the change in the meaning in the process of use, but with the concept referred to by corresponding terms.

Now let us consider the following example:

> Responsibility accompanies freedom.

This clause is written in the draft revision to the Japanese constitution proposed by Liberal Democratic Party, which is the governing party of Japan as of this writing. How should we behave in the face of this statement? If one adopts the stronger version of the Firthian view of meaning, one must accept that freedom should be accompanied by responsibility, although to what degree one must accept that depends on how widespread this discourse is. From the point of view of terminology, this statement is just *false* from start to finish, simply in terms of the concept represented by the term "freedom". Freedom includes such passive forms of freedom as freedom from torture (Berlin, 1969). If we apply the LDP statement to the concept of freedom from torture, we end up with the following:

> If you do not take due responsibility, you may not be free from torture.

This reveals the following essential fact about the concept of "freedom":

[1] Personal communication with Professor Kazuhiko Ohe, Graduate School of Medicine, The University of Tokyo.

Proceedings of the 11th Workshop on Building and Using Comparable Corpora at LREC 2018, Miyazaki, Japan.
Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff (eds.)

> That responsibility does not accompany freedom is the *sine qua non* trait of the very concept of "freedom," without which this word is nullified and we cannot talk about "freedom" at all.

So the statement "responsibility accompanies freedom" should not change the concept of "freedom." If such abuse of language spreads, however, it may become impossible to talk about freedom. In such a situation, we are not talking about the changes in the meaning of "freedom" as it becomes nothing to do with freedom if responsibility accompanies it. This is tantamount to killing the concept of freedom, and this is tantamount to killing the conditions which enable us to maintain the concept of freedom. Incidentally, *learning* for human being is not related to accepting the statement "responsibility accompanies freedom" as part of the determining feature of the concept of freedom, but to gain a system of judgement that enables one to properly identify this statement as false. The former is relevantly called *disciplinisation* or *indoctrination*, which is not – and indeed is the complete opposite of – learning.

It is often the case that the concepts represented by terms are not constitutively accessible and can only be *presumed* as a regulatory ideal (Kant, 1781). In other words, the identity of the concept represented by a term may not be described fully. But this does not mean that the identity of the concept does not exist and everything depends on usage. Reflecting this theoretical status of concepts and terms, practical study of terminology is also concerned with the identity of concepts.

1.2. Machine Learning/Disciplinisation

One of the standard ways of handling the "meaning" of words is word embedding or distributed representation of words. That representations obtained by `word2vec` enabled such operations as follows showed the power of distributed representation of words (Mikolov et al., 2013):

$$\text{Madrid} - \text{Spain} + \text{France} = \text{Paris}.$$

In the same manner, it is pointed out that the following also becomes possible:

$$\text{Doctor} - \text{Male} + \text{Female} = \text{Nurse}^2$$

We can immediately see the qualitative difference between these two cases, i.e. the former reflects the relationships among the meanings of these words, while the latter has nothing to do with the meanings of "doctor," "nurse," "female," or "male." and just reflects gender biases that exist in society and in social discourse. We can also recall what happened to Microsoft Tay, soon started tweeting about its admiration for Hitler and using racist slurs against Jewish and black people. Using the term we introduced above, we have to say that machines did not learn, but rather were disciplinised or indoctrinated[3].

Can corpus-based or data-oriented terminology processing get around these or similar issues? We have been (mostly unconsciously) assuming yes, for the following reasons:

- Specialised knowledge is created and expressed in the proper manner, and biases are filtered out through peer review in each specialised domain of knowledge;
- Popularisation and wider dissemination of specialised knowledge is also carried out in a due manner, reducing the granularity of discourse but essentially keeping the wholesomeness of the specialised knowledge.

Assuming these hold, we can safely use domain corpora for a narrower or wider range for different domains in different languages, even if machines can only be disciplinised and cannot learn in the proper sense of this word.

Unfortunately, however, a range of recent events indicate that relying on these assumptions is becoming more and more dangerous:

- Forgeries have repeatedly come to light and a number of papers have been retracted;
- Some authors have tried to cheat journal editors by supplying fake e-mail addresses for real scientists as potential reviewers;
- Unfounded historical revisionism and views based on such revisionism has appeared in descriptions of history in some school textbooks in Japan (and perhaps in other countries as well);
- Funding bodies require more and more short-term social "impact";
- Mass media pick up more and more sensational aspects of research with improper use of terms.

Together, these blur the distinction between scientific activities which are carried out in accordance with established norms of science and those activities that are not. Recall that science is destroyed when people, including "scientists," start regarding claims and "arguments" based on forged or fake data as part of science.

In such a situation, automatic terminology processing may contribute to the destruction of science through unconsciously extracting the abuse of concepts as normal and spreading them. Daille once argued for the necessity of detailed text profiling (Daille, 2008). If we start from corpora or textual data, text profiling becomes more and more important. Theoretically, however, the relation between concepts (and terms) and texts is the other way round. Texts are constructed in such a way that they make proper sense and concepts and terms are assumed beforehand. Text profiling is concerned with providing machines with appropriate information while assuming that machines are disciplinised rather than that they learn. Can we add the ingredient of learning rather than only avoid inappropriate disciplinisation? What does this mean?

This is the situation which terminology processing currently is facing. Having this in mind, I introduce some practical terminological tasks and some trials. In fact, since the mid-1990s, at the background of terminology processing, I have kept thinking of these issues. Words are grandiose, deeds are miserably tiny. Worse still, the practical tasks introduced below are only remotely related to what we have discussed so far. But let us move on anyway.

[2] An example cited in the Q&A session for Steedman, M., "On distributional semantics," invited talk at the Australian Language Technology Association 2016 Workshop.

[3] I owe this recognition to Dr. Hideto Kazawa of Google.

2. Issues in Terminology Processing

2.1. Terminology and Textual Corpora

Research in and the practice of terminology as an independent area of activity was first consolidated in Wüster's seminal work (Wüster, 1959), in which he put emphasis on the rigidity of concepts and terms. Felber states that terminology starts with concepts rather than terms, is concerned with the system of concepts in its synchronic state, and is not concerned with the linguistic features of terms that are unrelated to concepts (Felber, 1984).

In terminology, terms and concepts are defined as follows (de Besse et al., 1997):

term: A lexical unit consisting of one or more than one word which represents a concept inside a domain.

concept: An abstract unit which consists of the characteristics of a number of concrete or abstract objects which are selected according to specific scientific or conventional criteria appropriate for a domain.

Kageura showed that, theoretically, terminology as a coherent set of terms conceptually precedes individual terms; terms are items within a terminology which in its totality reflects the conceptual system of a domain (Kageura, 2015).

Two features of concepts and terms can be pointed out here:

1. A concept represented by a term may be updated, but does *not change* through casual use. This update of the concept is understood as a step towards the ideal state of that concept, which exists as a regulatory ideal.
2. Terms and the concepts they represent are attributed to the system of knowledge of the domain.

Since the 1990s, more descriptive approaches have appeared (Budin and Oeser, 1995; Temmerman, 2000). While these approaches have advanced *how* concepts and terms can be described, understanding of *what* concepts are seems to have remained intact behind the scenes[4]. The Wüstarian view of terms has always been there as the regulatory ideal for terminology. This also holds for corpus-based automatic terminology processing. After all, without this regulatory ideal, we do not need to and we cannot talk about terms and terminologies as something different from ordinary words, compounds or collocations anymore.

In corpus-based terminology processing such as monolingual and bilingual automatic term extraction, this regulatory ideal that links the work to terminology is implicitly taken into account when domain corpora are defined. Domain corpora are the discoursal part of the linguistic representation of the system of knowledge of the domain. This discoursal part, to be relevant, makes use of the terminological part, which is the other part of the linguistic representation of the system of concepts and knowledge of the domain. Though every now and then concepts are updated through discourse, specialised discourse at the same time critically depends on the system of concepts and the corresponding terminology.

[4]This perception may bring us back to Frege but we do not elaborate on this further here.

Thus term extraction thus should *not* be the task of extracting linguistic elements that are relevant to a given set of texts or domain corpora; it is the task of extracting terminology that represent a system of concepts and thus the system of knowledge of the domain *through* domain corpora. This contrasts with keyword extraction, which is defined as the task of extracting linguistic elements that are relevant to texts. One can extract keywords from a document which consists only of fake information and the extracted keywords can be valid, but one cannot extract terms from such a document.

This is the theoretical reason why text profiling becomes critical in corpus-based terminology processing (Daille, 2008). The practical result that text profiling can improve the performance of such tasks as bilingual term extraction (Morin et al., 2010) can be a reflection of this theoretical point. For text profiling, we can resort to external information at a variety of levels, such as the reliability of authors, of institutions authors are affiliated with, of journals, thus of publishers, or of the format of documents, etc. Unfortunately, it is not sufficient. We can see this from the example we observed above, i.e. the planned insertion of the statement "responsibility accompanies freedom" into the Japanese constitution. The agent trying to do this is the governing party and once inserted the statement will constitute a part of the Japanese constitution. In view of the external criteria, this statement is to be regarded as "reliable," even if it is nonsense. Ultimately, therefore, we need to resort to knowledge *itself* to avoid this sort of misjudgement. But how? Note that here the problem has gone beyond text profiling.

2.2. Conceptual Systems and Normativity

Two clues exist that guide us when dealing with this issue, though neither of them provides us with direct solutions to the problem we have discussed so far.

First, at a certain stage in the process of learning, human beings start judging information or a chunk of knowledge that is given to them and start refusing to accept it. This is because they have nurtured their *system* of belief, which is supported by the *system* of knowledge. One of the core parts of this system of knowledge is a vocabulary, which is not just a set of words but "a coherent, integrated system of concepts" (Miller, 1986). In the arena of sciences, the most basic part of this system of knowledge is reflected in terminology, which represents a coherent, integrated system of the concepts of the domain. A system of concepts is not just a set of concepts randomly collected. It embodies normativity, to the extent that we can talk about degree of systematicity and whether something is relevant to the system or not. Explicitly dealing with the terminology as a reflection of the system of concepts rather than dealing with individual terms or a set of given terms, therefore, can be a step towards properly handling terms, terminology and concepts, i.e. dealing with terms consistently and systematically in such a way that they collectively reflects the meaningful part of the system of concepts of the domain.

Let me cite an example here, though it is not terminological. Suppose we are interested in extracting words from textual corpora to construct a dictionary. Suppose that we

extracted a set of words from a corpus of 10,000 word tokens, and obtained two words that indicate types of fruit, i.e. "orange" and "apple". From the point of view of constructing a dictionary, given the range of words referring to fruit that are used in daily life in many English-speaking areas in the world, it is most natural that a dictionary which includes "apple" and "orange" as entries would also have "banana" as an entry.

To obtain the word "banana" from the corpus, we may have to extend the corpus to 100,000 word tokens. We would then obtain "banana", but would also obtain "mango" and "kiwi fruit". We would most probably think that a dictionary that contains "mango" and "kiwi fruit" as entry words should also have "papaya" as an entry. Otherwise, the set of entries lacks systematicity and coherency. To obtain "papaya", we may have to extend the size of the corpus to, say, 1,000,000 word tokens. In addition to "papaya," then, we would obtain "kiwano" and "star fruit," in which case we would need "dragon fruit" to make the set of entries in the dictionary coherent and systematic. This is the so-called "orange, apple, banana problem" [5].

Although this description is imaginary, a situation equivalent to it can happen in real-world dictionary-making situations. Kilgarriff et al. (2014) found that, in a project that aimed at developing monolingual and bilingual word lists for language learning using corpora, for nine languages and thirty-six language pairs, it was preferable to define a set of common key domains and populate the domains with words independently for each language. As domains they defined calendar, i.e. days of week, months, time, celebrations, colours, clothes, numbers, etc (Kilgarriff et al., 2014). This is partly because there is no guarantee that all the names for the days of week exist in a given corpus. This also indicates that the system of vocabulary or terminology is not a secondary, artificial derivation while discourse and texts are the first-hand manifestation of languages (Wilks et al., 1996).

The concept of normativity is also relevant to the textual or discoursal sphere. For instance, we do not refer to the New York Times, let alone AmericanNews.com, to make a legally *learned* argument about the Indonesian invasion of East Timor in 1975[6]. We refer to binding international law and authentic political records[7]. Indeed, researchers, irrespective of their research area, should be fully aware of what is called normativity here; they refer mostly to peer-reviewed and other academically reliable papers. They do not regard these papers and anonymous blog posts as equivalent. This observation is closely related to the proposal of text profiling mentioned above.

Documents thus have a degree of normativity. This concept is also frequently valid within cross-lingual setups. We often observe that a bilateral contract made between institutions in different countries in two languages adds such a statement as "in case of discrepancies between the versions in two languages, a preference is given for the interpretation according to a version in which the contract was originally drawn up." In bilingual or multilingual situations, it is usually the case that the document is written in one language, which is the source language. The corresponding documents in other languages are created through translation. This implies that, not infrequently, when context vectors for corresponding terms in different languages have some gaps, they are not relative to each other. We sometimes need talk about *deviations* of the usage of a term or the concept represented by a term, as in the case of the definition of cancer in insurance policies.

Linguists may say that normativity of terms and documents is not inherent in languages. May be true, but terms and terminologies are the functional class of languages and the determining factor is social and/or conceptual, which are not linguistic in the first place. What we see is that linguistics in its narrower sense falls short of addressing the issue we have observed so far. I see no merit to sticking to the purity of linguistics or whatever that cannot counter the destruction of the very conditions which enable us to sensibly communicate with each other, without resorting to physical violence. Freedom, in its essence, is never accompanied by responsibility, even if 99 percent of people claim that this is so. The rights of individuals, in their essence, are never accompanied by duties, even if the governing party of a nation declares this to be so. The concepts of freedom and rights should be properly maintained, logically, even when oppressive and discriminative discourse becomes prevalent.

3. Directions in Terminological Research

We can define a range of terminological studies and terminology-processing tasks that take into account the concepts of normativity and/or systematicity, both in monolingual and in bilingual or multilingual situations. To do so, we can conveniently distinguish two phases of terms and terminology: individual terms and concepts they represent as they are and in their use in texts, and the system of terminology and conceptual system.

3.1. Terms, Concepts and Use of Terms

If we focus on individual terms, their relation with concepts is the point of central importance, as has been pointed out by theoretical terminologists. Terminologists pay great attention to how to define concepts properly. Note that terminological lexicons with proper entries and reliable definitions are used as a resource that people commonly refer to and attain normative status. Normativity inevitablly accompanies tasks dealing with concepts represented by terms. Referring to terminologists and other specialists activities, we can define, for instance, the following tasks as taking into account the issues we raised in the previous section:

[5] Personal e-mail communication with the late Dr. Adam Kilgarriff on April 1, 2014. Though the conversation took place on April Fool's Day, the content was academic.

[6] The Indonesian invasion of East Timor began on 7 December 1975, one day after then U.S. Secretary of State Henry Kissinger left Jakarta.

[7] The National Security Archive of George Washington University revealed the conversation between then U.S. President Gerard R. Ford and Kissinger and then Indonesian president Suharto, responsible for the invasion and the massacre that followed. The record showed that U.S. had given "greenlight" to Suharuto's planned invasion. See https://nsarchive2.gwu.edu/NSAEBB/NSAEBB62/.

Creating/extracting definitions: We can define a task of creating or extracting a normative definition(s) for a given term using corpora or other resources. In its ordinary sense, definition extraction is a well-established NLP task (Sierra et al., 2009). We can also regard word embeddings as the task of defining word meanings.

Detecting deviations: In the context of what we have discussed so far, what matters about definitions is their normativity. So one possible application – or evaluation scheme of definition extraction through application – can be the task of detecting deviated use of terms in terms of their definitions. Automatically judging that the statement "responsibility accompanies freedom" is misusing the concept of *freedom* gives a concrete image of the objective of this task. It is somewhat similar to word sense disambiguation and also outlier detections used for evaluating word embeddings (Camacho-Collados and Navigli, 2016), though these tasks regard meanings as relative. At a different level, this task is related to detecting logically inappropriate statement. We started a research for detecting deviations of usage of technical terms in Japanese mass media, currently focussing on the domain of law and politics. A very embryonic observation was reported in (Tang and Kageura, 2017). When term variations exist, i.e. different representational forms are regarded as representing the same concept (Daille, 2017), controlling the surface form of terms also becomes an issue accompanying deviation detection.

Detecting cross-lingual gaps: As in the case of "cancer" and its Japanese "equivalent," terms in different languages that are regarded as representing the same concept can be different in some critical details. Not only judging the degree of correspondence but also evaluating the critical difference will be an important task as an extension of bilingual term extraction from parallel or comparable corpora. If we take into account the fact that very frequently corresponding documents in two or more languages do *not* have the same status (the goodness of TL texts should be evaluated by using original SL texts as the norm), the task is defined as directed, using the normative concepts represented by SL terms to judge the concepts (also normative in a monolingual setup) represented by TL terms[8]. At a different level, dealing with representational variations also becomes an issue.

Text profiling: Social profiling of texts can provide corpus-based processing with external criteria of normativity. To what extent texts themselves can be used to evaluate their normativity is also an important issue. This is technically related to text clustering or classification.

As technical problems involved in these tasks are similar to related tasks that have been well established, methods proposed for these related tasks may be adopted to tackle these problems. The difference resides in definitions of problems. Note that issues related to above topics are recently being dealt with in NLP. For instance, fact checking and analysing and detecting biased language are listed as topics relevant to the Workshop "Natural Language Processing meets Journalism." In the field of terminology, most of these have been dealt with manually.

3.2. Terminologies and System of Concepts

While we have witnessed a great advance in methods of both monolingual and bilingual automatic term extraction (ATE), the systematicity or coherency of extracted terms have not been taken into account when these methods were evaluated. Indeed, it is not stated as one of the aims of ATE in most cases. It is understandable, as we do not really know how to measure systematicity or coherence of terminologies. In lexicography, selecting a coherent set of headwords for a dictionary is left to the expertise of experienced lexicographers and remains one of the last frontiers of lexicography yet to be systematised[9]. Nevertheless, as we discussed above, systematicity is one of the essential features in knowledge and thus to address this issue is critical to the study of terminology.

Taking into account that terminologies represent conceptual systems, we can define, among others, the following tasks:

Evaluating systematicity of terminologies in terms of conceptual systems they represent: Terms are relatively motivated. Complex terms, which constitute 70 to 85 percent of all the terms in most domains in many languages, represent concepts by showing their main conceptual characteristics and their relations through constituent elements (Sager, 1990). A terminological representation thus reflects conceptual system to a certain extent. How systematic a terminology represents the corresponding conceptual system depends on domains and languages. Here we can define the issue of systematicity of terminological representations vis-à-vis the conceptual system. Once we can establish a method that can evaluate the systematicity of terminological-conceptual system, we may be able to judge to what extent a newly obtained term is relevant to the conceptual system and thus to the domain. The dynamic modelling of terminological and conceptual growth can be considered as an extension of this task. Ontology building shares a similar concern, though it focuses on conceptual system rather than representations.

Evaluating cross-lingual differences in the systematicity of terminologies: Terminologies of different languages represent the same conceptual system[10] differently. Evaluating the difference in the systematicity of terminological representations in different languages not only is important for the theoretical terminology but also contributes to cross-lingual terminological applications.

These studies are important not only from the theoretical point of view but also for real-world applications. Recall

[8] One may argue that MT deals with this issue indirectly when TL expressions are selected. For professional translators, being able to explain the difference among possible choices of TL expressions and the reason why a particular expression was chosen is not only part of their competence but also part of the *end*-product; the definitions of end-to-end in MT and in human translation are different.

[9] Personal communication with Dr. Judy Pearsall of the Oxford University Press on 8 July 2010 at the occasion of Euralex 2010 held in Leeuwarden, The Netherlands.

[10] Though the case of "cancer" indicates that it is not necessary safe to assume the identity of conceptual systems represented in corresponding terminologies in different languages, we can assume that they are approximately the same.

that in the keynote presentation in BUCC 2017, Professor Philippe Langlais stated "Despite numerous studies devoted to mining parallel material from bilingual data, we have yet to see the resulting technologies wholeheartedly adopted by professional translators and terminologists alike (Langlais, 2017). One of the reasons that not many advanced term extraction methods are not used in the real-world terminology management tasks or in translation is that it is difficult to know what are extracted and what are missed. If one could judge the status of extracted terms in relation to the existing set of terms, terminologists would be able to take advantage of the results of advanced methods more comfortably. The problem of dealing with the systematicity of terminologies within data-oriented or corpus-based language processing framework is that the size of terminologies are small. This may be one of the reasons why not much work has been carried out that deals with terminologies *per se*[11].

4. Two Concrete Studies

We introduce here two concrete studies we have been and are carrying out, which (remotely) take into account the issues of systematicity and normativity of terminologies and terms. The first is augmentation of bilingual terminologies, and the other is controlling term translations.

4.1. Augmentation of Bilingual Terminologies

In some languages pairs, as in the case of Japanese and English, manually constructed high-quality bilingual terminologies exist, and there is a strong demand for updating these terminologies. Standard corpus-based bilingual term extraction, unfortunately, cannot satisfy this demand, because new terms mostly occur with low frequency in the corpus and often hard to extract, and the relationship between extracted terms and entries in the existing terminologies is not transparent. Against this backdrop and taking into account issues we have discussed so far, we are developing a terminology-driven method for augmenting existing bilingual terminologies (Iwai et al., 2016a; Iwai et al., 2016b). The framework is simple:

1. Assuming that terminologies systematically reflects conceptual systems, we define terminological network which represents termino-conceptual structure of the domain with terms as vertices and edges as common constituent elements among terms. Figure 1 shows a terminology network of a small putative terminology.
2. Apply partitive clustering to the terminological network to obtain subclusters of terms which represent conceptual subsystem (Figure 2). Corresponding terminologies in different languages show similar tendencies, though differences are not small.
3. Complex terms are formed in accordance with the dynamics of these subclusters. Head-modifier bipartite graphs are created for terms in these subclusters, and new term candidates are generated by interpolating the missing links.
4. Bilingual candidates are generated by compositional matching, assuming that terms are motivated roughly in the same manner.
5. Candidates are validated by Web search.

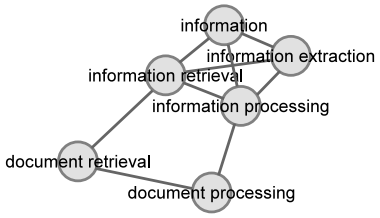

Figure 1: An exemplar terminological network.

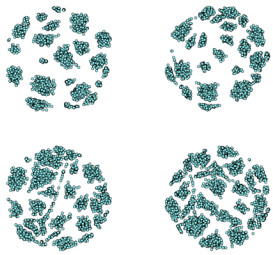

Figure 2: Subclusters in computer science (top) and economy (bottom) in Englih (left) and Japanese (right).

As of now, the method has several shortcomings:

1. The degrees of systematicity at the representational level vis-à-vis the conceptual systems have not been taken into account. The method just assumed that terms are reasonablly well motivated and thus terminologies systematically reflect the underlying conceptual systems to a degree that we can simply use terminological representations as a key to approximate conceptual systems. This assumption holds in monolingual situation, but as shown in Figures 1 and 2, different languages represent different parts of conceptual systems.
2. The gap between the systematicity of English and Japanese terminologies as reflected in the terminological networks can be explored to further capture the terminological structures that reflect conceptual systems. We have not elaborated on this.
3. Terminology networks are defined in a very rudimentary way. As edges are made when two terms (vertices) have common constituents, the hierarchical relations encoded in the forms of terms are not reflected in the network. Also, the dependency relations between constituent elements within terms are not encoded in the networks. This is the other side of the fact that the method currently does not take into account the

[11] Some other reasons are: terminologies and dictionaries are generally regarded as secondary creations compared to documents, which is based on the misunderstanding of languages; and terminologies are not too large and stable so it has been held that manual handling suits better than automatic processing.

conceptual system (see 1 above) and carries out candidate term generation purely at the level of terminological representations. It would be more theoretically proper to define the conceptual system separately from terminological networks, make correspondences between these two layers, and resort to the information at these two levels. To define the conceptual system that corresponds to a given set of terms, we are currently examining the use of distributional representation of constituent elements of terms in terminologies.

4. Currently all the candidates are treated equally. Using the information contained in termino-conceptual structure, we can give weight to candidates in terms of their status within the termino-conceptual system.

We are currently working to overcome these issues.

4.2. Controlling Term Translations

In translating terms, one TL term for an SL term is the basic principle for properly controlled documents (Sharoff and Hartley, 2012). In practice, it is frequently the case that several different TL terms are used for a single SL term. Terminology control should be made at the early stage of translation projects, i.e. controlled bilingual terminologies should be provided with translators involved in the project before they start translating documents. While language service providers generally adopt this procedure, it is still difficult to control terms properly. For instance, across Japanese municipalities, Japanese terms for administrative procedures are the same, but their translations vary because each municipalities translate their documents independently to each other. In these cases, "posterior" terminology control is essential; it is posterior in relation to already translated documents, but prior in relation to future documents to be written and translated.

One of the theoretically essential and practically important issues is to estimate the coverage of collected terms[12]. This issue is related to several other questions, i.e. whether or not the size of the corpus should be extended to collect sufficient number of terms, how many more texts should be checked, and how controlling terms affect these tasks.

We carried out TL term control for Japanese municipality documents manually (Miyata and Kageura, 2018). We collected three Japanese-English parallel documents that describes municipal procedures and extracted bilingual terms from them. Table 1 shows the number of terms ($V(N)$ indicates the number of term types, N the number of term tokens). Although we collected corresponding terms from parallel documents, the number of terms both in types and in tokens differ between two languages. We identified 374 Japanese term variations (12.4%) and 1258 English term variations (36.3%); TL terms have more variations than SL terms (Warburton, 2015).

Variations were groped and a preferred term for each group was assigned, based on three types of evidence, i.e. frequency evidence, topological evidence (expressions of terms) and dictionary evidence. After the terminology control, the numbers of term types in Japanese and English became closer.

[12]That extracted terms be evaluated in terms of coverage is a prerequisite for evaluating systematicity of terminologies.

	$V(N)$	N	$N/V(N)$
Japanese	3012	15313	5.08
English	3465	15708	4.53

Table 1: The number of extracted terms

	$V_c(N)$	$V(N)_c/V(N)$	$N/V_c(N)$
Japanese	2802	93.0%	5.47
English	2740	79.1%	5.73

Table 2: The number of extracted terms

What do they mean for the status of terms we collected? First, to evaluate the status of terms we collected vis-à-vis potential terminology we are dealing with in the domain, we adopted self-referring evaluation of collected terms. The idea is simple: (a) estimate the population number of terms using the distributional information of the terms we collected, and (b) evaluate the coverage of the collected terms against the estimated size of terminology. For the estimation of population number of terms, we used LNRE models (Baayen, 2001; Evert and Baroni, 2007).

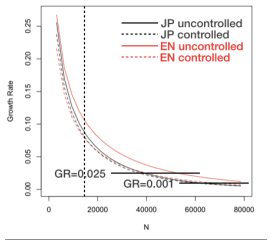

Figure 3: Growth rate of terms to the corpus size.

Figure 3 shows the growth rate of terms, before and after terminology controll was applied. We can observe several points: coverage became higher when terms were controlled and if we extend the corpus size to 40,000 word tokens, only one out of 40 terms is expected to be new. These enable us to evaluate the status of terms and terminology controll and ROI-based evaluation of the usefullness of extending the corpus.

5. Conclusions

We examined theoretical and social issues related to terminology, and clarified the position of terms and terminologies in relation to textual corpora together with issues in corpus-based terminology processing. We argued that the identity of concepts represented by terms is supported by the regulatory ideal, which provides the conditions upon which we can rationally communicate with each other in the first place. The concepts of systematicity and normativity were then introduced as on-the-ground concepts that reflect the regulatory ideal of the identity of concepts. We defined a range of tasks that take into account these issues and introduced two concrete studies as examples.

Much of what we discussed here is yet to be fully pursued, although relevant technologies exist. Indeed, the same technologies that can be used to pursue the tasks defined here can easily be used to promote pseudo-communication, including "fake news" and other forms of communication that promote hatred and discrimination. Unfortunately, current data-driven ML technologies do not internalise the regulatory ideal that human beings have tried to pursue painstakingly, so it is still upon us to decide how these advanced technologies are used. Cross-lingual comparable corpora contain interesting and important gaps, which we can explore to promote mutual understanding, as understanding starts from the recognition and identification of differences.

6. Acknowledgements

I would like to thank the organisers of BUCC Workshop, Professor Reinhard Rapp, Professor Pierre Zweigenbaum and Professor Serge Sharoff for giving me an opportunity to give this talk. Terminology-driven augmentation is a joint work with Dr. Koichi Takeuchi and Mr. Kazuya Ishibashi of Okayama University and Ms. Miki Iwai and Mr. Long-Huei Chen of the University of Tokyo, and was partly supported by JSPS Research Grants 24650122 (2012–2014). Evaluation of controlled terminology is a joint work with Dr. Rei Miyata of Nagoya University, and was partly supported by JSPS Research Grants 25240051 (2013–2017) and KDDI Foundation Research Grant Program (2014–2016).

7. Bibliographical References

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht.

Berlin, I. (1969). *Four Essays on Liberty*. Oxford University Press, Oxford.

Budin, G. and Oeser, E. (1995). Controlled conceptual dynamics: From 'ordinary language' to scientific terminology — and back. *Terminology Science and Research*, 6(2):3–17.

Camacho-Collados, J. and Navigli, R. (2016). Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 43–50.

Daille, B. (2008). Terminologie et traitement automatique des langues. In *TAMA 2008*, Ottawa.

Daille, B. (2017). *Term Variation in Spericalised Corpora*. John Benjamins, Amsterdam.

de Besse, B., Nkwenti-Azeh, B., and Sager, J. C. (1997). Glossary of terms used in terminology. *Terminology*, 4(1):117–156.

Evert, S. and Baroni, M. (2007). zipfr: Word frequency distributions in r. In *45th ACL Poster and Demo Session*, pages 29–32.

Felber, H. (1984). *Terminology Manual*. UNESCO, Paris.

Iwai, M., Takeuchi, K., Ishibashi, K., and Kageura, K. (2016a). A method of augmenting bilingual terminology by taking advantage of the conceptual systematicity of terminologies. In *Computerm 2016*, pages 30–40.

Iwai, M., Takeuchi, K., and Kageura, K. (2016b). Cross-lingual structural correspondence between terminologies: The case of english and japanese. In *TKE 2016*, pages 14–23.

Kageura, K. (1995). Toward the theoretical study of terms: A sketch from the linguistic viewpoint. *Terminology*, 2(2):239–257.

Kageura, K. (2015). Terminology and lexicography. In Hendrik J. Kockaert et al., editors, *Handbook of Terminology*, pages 45–59, Amsterdam. John Benjamins.

Kant, I. (1781). *Critique of Pure Reason*. Cambridge University Press (1999), Cambridge.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., janne Bondi Johannessen, Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Langlais, P. (2017). Users and data: The two neglected children of bilingual natural language processing research. In *BUCC 2017*, pages 1–5.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates.

Miller, G. (1986). Dictionaries in mind. *Language and Cognitive Process*, 1:171–185.

Miyata, R. and Kageura, K. (2018). Building controlled bilingual terminologies for the municipal domain and evaluating them using a coverage estimation approach. *Terminology*, 24(2) (to appear).

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2010). Brains, not brawn: The use of 'smart' comparable corpora in bilingual terminology mining. *ACM Transactions on Speech and Language Processing*, 7(1).

Sager, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam.

Sharoff, S. and Hartley, A. (2012). Lexicography, terminology and ontologies. In Alexander Mehler et al., editors, *Handbook of Technical Communication*, pages 317–346, Boston. Mouton De Gruyter.

Sierra, G., Pozzi, M., and Torres, J.-M. (2009). *Proceedings of the 1st Workshop on Definition Extraction*. ACL, Borovets.

Tang, L. and Kageura, K. (2017). 'Fighting' or 'conflict'? An approach to revealing concepts of terms in political discourse. In *EMNLP Workshop on Natural Language Processing meets Journalism*, pages 90–94.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. John Benjamins, Amsterdam.

Warburton, K. (2015). Terminology management. In Sin-Wai Chan, editor, *Routledge Encyclopedia of Translation Technology*, pages 644–661, New York. Routledge.

Wilks, Y., Slator, B. M., and Guthrie, L. M. (1996). *Electric Words*. MIT Press, Cambridge, Mass.

Wüster, E. (1959). Das Worten der Weld, schaubildlich und terminologisch Dargestellt. *Sprachforum*, 3(3):183–204.

Grouping conversational markers across languages by exploiting large comparable corpora and unsupervised segmentation

Laurent Prévot[1,2], Matthieu Stali[1], Shu-Chuan Tseng[3]
[1] Aix-Marseille Univ, LPL, Aix-en-Provence, France
[2] Institut Universitaire de France, Paris, France
[3] Institute of Linguistics, Academia Sinica, Taipei, Taiwan

Abstract

This work approaches *Conversational and Discourse Markers* (hereafter DM) from a radical data-driven perspective grounded in large comparable corpora of French, English and Taiwan Mandarin conversations. The key features of our approach are (i) to account for lexicalization as a by-product of unsupervised segmentation applied to our corpora, (ii) to exploit simple metrics for clustering DM (both within a language and within multilingual clusters). We explore the benefits and the drawbacks of such a radical approach to DM. In particular we compare the DM clusters obtained from traditional segmentation into tokens (as given by manual transcription of the corpora) vs. unsupervised segmentation. The metrics on which we ground the clustering experiments are based on contrast between (i) short vs. longer utterances distribution and (ii) position within longer utterances.

1. Introduction

Leaving aside some interesting descriptive studies, there are not many attempts to perform systematic and quantitative comparative analysis of social interactions (such as conversations and task-oriented dialogues) from a linguistic perspective. Language resources and natural language processing tools still rely on written canonical data. In the context of studying, comparing and exploiting social interactions; in which speech is fiercely spontaneous and exhibits its own patterns; appears to be a major handicap. Once situated within a multilingual or translational task, it becomes even more difficult to handle by adding the bias towards written canonical data of each language before being able to consider the multilingual or translational aspects themselves. Thus, we propose here to adopt a relatively shallow and data-intensive approach to consider directly the spoken data without passing through resources and tools created for canonical written data.

Comparable corpora are extremely useful for a range of Human Language Technology tasks but also for exploring phenomena across languages. In this paper we are developing a data-driven approach to study discourse and interactional markers (hereafter DM) in a comparative way thanks to large conversational comparable corpora. Our work aims at identifying and grouping discourse markers into homogeneous classes through a purely bottom-up approach carried out on large corpora. Studying discourse markers has a long history in linguistics and corpus linguistics (see Section 2.) but our approach combine some methodological choices that makes it original. This approach relies on rather large comparable conversational corpora across the languages scrutinized (introduced in section 3.). Moreover those corpora have to be transcribed. More precisely the two key ingredients are (i) to explore unsupervised segmentation of our data sets as explained in 4.1. ; (ii) to explore a set of distributional measures of the word-like units for characterizing them (See 4.2.). Finally, in our experiments, standard clustering techniques are used to obtain groups of clusters that we try to label with categories in section 5..

2. Discourse Markers

Discourse markers, such as *like* and *well* in English to quote a few, are key elements in conversations which help speakers build their speech's structure. The main issue when studying DMs lies in the lack of consensus and thus in the various definitions and denominations that can be found among works in the literature related to conversational speech. We can mention the following terms, being the most frequently used: *discourse markers* (Schiffrin, 1988; Fraser, 1999); *pragmatic markers* (Furko, 2009; Garric and Calas, 2007), *discourse particles* (Schourup, 1985; Fischer, 2006), *spoken particles*(Fernandez, 1994; Fernandez-Vest, 2015) and *discourse connectives* (Roze et al., 2010; Lenk, 1998).

Even though we can understand why a categorization task for DMs remain difficult given their poly-functionality and the various stages of functional multi-word expressions' lexicalization, scholars would usually agree on several main aspects. DMs' primary functions are described as being related to a relatively defined set of functions: turn-taking system, discourse relations cuing, discourse structuring, interpersonal relationships marking, speech management or politeness (Fischer, 2006).

Recently, linguists have been interested in automatically identifying DMs for translation purposes. Some results have shown there were discrepancies between bilingual dictionaries translations and the semi-manual annotation ones for a given pair of DMs from two different languages (Roze and Danlos, 2011). Other works include *The TextLink project*[1] which is specifi-

[1] `http://textlinkcost.wixsite.com/textlink`

cally analyzing this aspect, by focusing on discourse-annotated corpora to allow cross-linguistic studies of discourse. The corpus based method seems an adequate tool for caterorizing DMs as it unites a theoretical task consisting in setting parameters of definition variables with an empirical study on spontaneous speech corpora (Crible et al., 2015).

3. Data

The comparable corpora we used for this experiment were : the CoFee collection of corpora (Prévot et al., 2016) (made of CID (Blache et al., 2009), MapTask(Gorisch et al., 2014) and DVD(Prévot et al., 2016)) together with DECODA corpus for French ; Switchoard transcripts for English (Godfrey et al., 1992) ; and Academia Sinica conversational corpora (MCDC, MTCC, MMTC) for Taiwan Mandarin (Tseng, 2013). We experimented with various subcorpora and across languages as illustrated in table 1 and with different potentials *base units*: syllables and letters for French; Characters, Pinyin (with and without tone) for Mandarin and only letters for English.

Corpus	# Tokens	# pseudo-Utterances
CID	125 619	13 134
MTR	42 016	6 425
MTX	36 923	5 830
DVD	64 023	7 989
DECODA (part)	580298	88 982
French	851202	122 360
MCDC	316 422	61 000
MTCC	122 200	26 000
MMTC	34 500	8 300
Mandarin	472 000	95 000
SWBD (English)	2 967 028	391 592

Table 1: Corpora used in the study

Some of those corpora are truly comparable while it is more debatable for others. MTR + MTX on French and MTCC for Mandarin are perfectly comparable since they have been recorded using the same protocol. CID for French and MCDC + MMTC are also very similar by nature. English Switchboard is perhaps a bit different in principle but in practice, it shares most of the features present in the previous corpora. The less similar of the set is French DECODA since it is recorded in a specific context (call center of Paris public transportation enquiries number). However, we add criterion during the extraction to try to avoid too many corpus specificities in a given language. Overall, all those corpora are truly conversational ones exhibiting the usual range of phenomena involved in fiercely spontaneous and interactional speech data. For all these corpora, the transcripts have been force-aligned at the word level.

Concerning the transcription, a standard orthographic transcription had been adopted for thes corpora. The spokenb particles do have standardized written forms in French (*euh, mh,...*) and English (*uh, um, mh...*). In the Taiwan Mandarin corpora, discourse particles, discourse markers, and fillers were transcribed with capital letters to distinguish themselves from foreign words such as English. Fillers are transcribed according to their phonetic forms. For instance, *UHN* is equivalent to *uhn* in English; MHM is something that is frequently observed in Mandarin, but not in English. In particular, multi-syllabic fillers are transcribed in one single unit, separated by H, e.g. *UHNHN*. See (Tseng, 2013) for more details.

4. Methodology

4.1. Segmentation

We use non-supervised machine learning algorithms (based on Branching Entropy, already applied to written Mandarin (Magistry and Sagot, 2012; Magistry, 2013)) for segmenting our sequence of characters coming from the conversational transcripts into our *base units* (spoken tokens). There are currently better methods for segmentation, especially for Chinese Word Segementation, but they require extremely large corpus that are not available for spoken language. Moreover, we were interesting in using the very same methodology on Mandarin, French and English with the idea in mind that the data set segmented in this same way across the languages could exhibit less divergence than being biased by the written form tradition of each language.

More precisely we use Eleve[2] *(Extraction de LExique par Variation d'Entropie - Lexicon extraction based on the variation of entropy)* toolkit. This method is helpful for our study because it allows us to get units grounded on the same principles and therefore not being biased by written processing techniques or conventions employed in different languages. Such an approach results in a new starting point for the type of lexical experiments we will perform later. An illustration of new units for French and English created by our approach are illustrated by Table 2. A benefit of such an approach is that we do not have to define what an individual word or multi-word expression is. We have done our experiments both with traditional segmentation (space-based) and with the output of unsupervised segmentation (in which, for example, *'you-know'*, turned out to be a unit). For a related work see (Dobrovoljc, 2017) which compare different association measures applied to discourse marker items.

While our unsupervised segmentation is very interesting to gather functional multi-words expressions into one unit as a result of the segmentation, it also presents some issues. For example, in French and English, it tends to split bound morphemes such as plural and gender marks as well as some verbal inflections. However, for our purpose of studying DM this feature should not be an issue.

[2] `https://github.com/kodexlab/eleve`

French	English	Mandarin
tu-vois	you-know	
mh-mh	uh-huh	MHMHM
ah-ouais	oh-yeah	對 A
c-est-vrai	that-s-right,that-s-true	
et-euh , donc-euh	and-uh, and-um	
et-puis, mais bon	and-then	
comme-ça	like-that	
dans-le, sur-le	in-the	
il-y-a	there-is	

Table 2: Examples of word like units created at segmentation stage *('-' in the units correspond to spaces in a traditional transcription)*

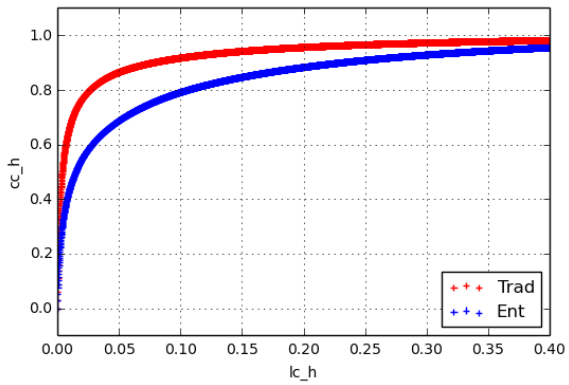

Figure 1: Comparative lexical growth (French) between traditional segmentation and Branching-entropy (x-axis : Coverage of the lexicon ; y-axis: Coverage of the corpus) segmentation

The unsupervised segmentation step provides a segmented corpus and a derived lexicon. In figure 1), we illustrate the lexicon coverage vs. corpus coverage of traditional vs. unsupervised segmentation.

In these corpora, we approximate the notion of utterance by using *Inter-Pausal Units* defined by continuous stretches of speech in between pauses of at least 200 milliseconds. Therefore, both our *lexical units* and our utterances are objective as possible, only relying on speech timoing and on the transcript.

4.2. Quantitative measures

Scores We argue that conversationally speaking, words distribution -Discourse Markers in particular- varies significantly depending on the type of utterances they occur in. A first relevant method being easy to apply in the study of conversation consists in separating the shortest sentences from the longer ones. Besides, it is a known fact that DMs can be found at specific positions in utterances (initial, median, final) with the initial and final ones being the most frequent (Aijmer, 2013; Filippi-Deswelle, 1998; Fraser, 1998; Muller, 2005; Stali, 2015; Stali, 2016). We propose to cross the two parameters mentioned above (type of utterance vs position in the utterance) to chart DMs.

Based on those two principles, we define a series of values aiming at characterizing quantitatively any form of the corpus (N: corpus size, S: number of tokens in short utterances, L: number of tokens in longer utterances, F_{all}: frequency of the token, F_{short}: frequency of the form in short utterances;F_{long}:frequency of the form in non-short utterances, F_{ini}:frequency of the form in initial position of longer utterances, F_{fin}:frequency of the form in final position of longer utterances

- $\frac{F_{all}}{N}$: relative frequency
- $\frac{F_{short}}{S}$: relative frequency of the form within short utterance forms
- $\frac{F_{long}}{L}$: relative frequency of the form within longer utterance forms
- $\frac{F_{short}}{F_{all}}$: tendency to occur in short utterances
- $\frac{F_{short}+F_{ini}+F_{fin}}{F_{all}}$: a sort of *"dm-hood"* of the form (tendency to occur in all canonical DM and interactional markers positions)
- $\frac{F_{ini}}{F_{long}}$: tendency to occur in initial position within longer utterances
- $\frac{F_{fin}}{F_{long}}$: tendency to occur in final position within longer utterances

We also use some of those scores to filter the set of items under consideration. More precisely we tested different thresholds for *relative frequency* and *dm-hood* scores. For French and Mandarin, we made sure that the relative frequency threshold was met for at least two-subcorpora in order to avoid domain-based items that could come from Maptask or DECODA corpora. This was both impossible and unnecessary to do on Switchboard corpus which is a lot larger and already more diverse thematically.

5. Experiments

In the context of this work, we were interested in comparing the clustering (and its implicit discourse marker characterization) in two approaches: traditional tokenisation and unsupervised segmentation. After segmenting the data sets and computing the scores as described in the previous sections we processed as follows. We filtered for relative frequency (threshold= 0.0005) and dm-hood (threshold= 0.3). Since we are at an exploratory stage of our work, those thresholds were chosen after inspection of results for a range of values for the both of them. We normalized all the resulting values, then applied PCA to the output and checked the *explained variance ratio* for deciding a number of principal components. The way DM are spread into the dimensions is illustrated for English DM in Figures 2 and 3 for traditional and unsupervised segmentation

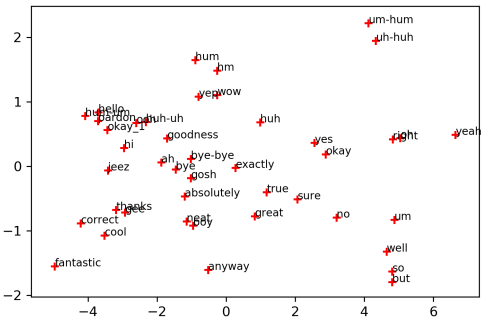

Figure 2: English DM plotted on the 2 principal components, based on traditional segmentation

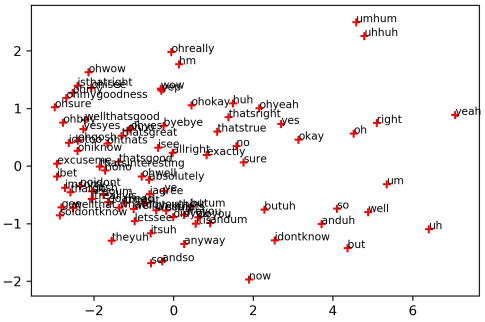

Figure 3: English DM plotted on the 2 principal components, based on unsupervised segmentation

respectively. Finally, after using the elbow techniques to determine an optimum number of clusters, we computed the clusters presented in Figures 4 and 5 for traditional and unsupervised segmentation respectively.

It is not straightforward to label the resulting clusters. However, it is possible to identify some known groups of markers in the clusters. We attempted to use the same color for similar clusters in both tradition and unsupervised results. Concerning traditional segmentation, the green and the yellow clusters host typical feedback items and a relative good match across languages. The division into two clusters is probably due to the fact that the items in the green cluster, in addition to be used frequently isolated as feedback items, may also occur in initial position (which is less the case for the 'yellow' items). Similar structure is observed for the unsupervised segmentation.

The 'blue' cluster corresponds to more evaluative and attitudinal items, at least for French and English. It is interesting to note that our very rough distributional measures are able to discriminate those items from the previous yellow and green clusters. We can see there is an adequate match between French and English but

ah,oh	oh	
ouais,oui	yeah,yes	
voilà	okay,right	
		EIN, EN, HON
mh	uh-huh,um-hum	MHM,MHMM,UHM...
d'accord,ok		
		HEIN,HEN,ON
alors,donc	so	然後, 所以
ben,bon	well	
et		
euh	um	NA
mais	but	可是
non	no	沒有
	sure	
		EI
		其實,就是,因為
cestca	correct	
exactement	absolutely,exactly	
hein		
hum		
putain	boy,goodness,gosh,jeez	
	anyway	
	bye,byebye,thanks	
	cool,fantastic,great,neat,gee	
	ah	HO
	true	
		就是說

Figure 4: Cluster (one per color) grounded on traditional segmentation

not so much for Mandarin.

Finally the red cluster includes at least two kinds of items: discourse connectives but also filled pauses and even interactional management items (French 'hein' in the unsupervised case). This is probably due to a lack of discrimination capacity for forms occurring within longer utterances at different places. For example, we know that *'hein'* tends to be more final but it is not enough to generate a specific cluster. Another explanation can be found in abandoned utterances. Those abandoned utterances typically end with a filled pause marker (French *'euh'*, English *'um,uh'*, Mandarin *'NEGE', 'NA'*. This (frequent) phenomenon therefore tends to make those items more distributionally similar to final particles like *hein*. Similarly, it may be rather surprising to see contrast connectives (*mais* / *but* / 可是) in those clusters. As mentioned above, this is probably due to unfinished utterances or utterance segmentation (based on pauses). However, in this cluster, there is a very satisfying match across the three languages.

6. Conclusion and Future Work

The exploration of DM spaces based on comparable corpora allowed us to show it remains possible to identify DM clusters, even through a cross-linguistic approach. The benefits of the unsupervised segmentation are not clear at this stage, specially for Mandarin

ah,oh		
ahoui,nonnon	no,ohno	
voilà	thatsright,thatstrue	對YA
ben		
	ohokay,okay,yes	
		EIN,EN,MHMM,ON,UNH
d'accord	right	對A
ouais,oui	yeah	
mh	uhhuh,unhum	MH
oh		
		HEIN,HEN
cestbon,cestvrai,cestça	sure	
eteuh,cesteuh,maiseuh,donceuh	andum,butuh,butum	
exactement	exactly	
tuvois	isee	
hum, maisbon,putain		
		HO,對不對
ahbon	ohreally,ohwow,wow	
	isthatright	
hm,hmhm	hm,huh	MHMHM, NHNHN, UHM
ok,toutàfait		
	yep	
alors,donc	so	
bon	well	
et	anduh	
euh	uh,um	NEGE,NA
hein		
mais	but	可是
non		
		EI,HON,其實,因為,就是,然後

Figure 5: Cluster (one per color) grounded on unsupervised segmentation

data. However, the method and approach adopted tend to demonstrate that the traditional segmentation already benefits from adapted transcription convention which includes rules for grouping specific words together. However, we believe it might be interesting to dig further in how much can be achieved without too many supervisions and bias from written resources. In the future, our first objective is to deeper scrutinize the elements in the structure of the Mandarin utterances which prevents DMs to be better clustered correctly with French and Mandarin items.

7. Bibliographical References

Aijmer, K. (2013). *Understanding pragmatic markers: a variational pragmatic approach*. Edinburgh: Edinburgh University Press.

Blache, P., Bertrand, R., and Ferré, G. (2009). Creating and exploiting multimodal annotated corpora: the toma project. *Multimodal corpora*, pages 38–53.

Crible, L., Bolly, C. T., Degand, L., and Uygur-Distexhe, D. (2015). Mdma: un modèle pour l'identification et l'annotation des marqueurs discursifs "potentiels" en contexte. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (16).

Dobrovoljc, K. (2017). Multi-word discourse markers and their corpus-driven identification. *International Journal of Corpus Linguistics*, 22(4):551–582.

Fernandez-Vest, J. (2015). *Detachments for cohesion: toward an information grammar of oral languages*, volume 56. Walter de Gruyter.

Fernandez, M. J. (1994). Les particules énonciatives dans la construction du discours. *Linguistique nouvelle*.

Filippi-Deswelle, C. (1998). *La relation dite de concession - Etude de Though, Although, Even Though et Even If antéposés en anglais contemporain*. Paris: Université Paris 7.

Fischer, K. (2006). *Approaches to Discourse Particles*. Amsterdam: Elsevier.

Fraser, B. (1998). Contrastive discourse markers in english? *Discourse markers: Descriptions and theory*, pages 305–312.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, (31):931–952.

Furko, P. B. (2009). *The pragmatic marker - discourse marker dichotomy reconsidered - the case of 'well' and 'of course'*. dea.lib.unideb.hu.

Garric, N. and Calas, F. (2007). *Introduction à la pragmatique*. Paris: Hachette Education.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Gorisch, J., Astésano, C., Bard, E. G., Bigi, B., and Prévot, L. (2014). Aix map task corpus: The french multimodal corpus of task-oriented dialogue. In *LREC*, pages 2648–2652.

Lenk, U. (1998). *Marking discourse coherence - Functions of discourse markers in Spoken English*. Gunter Narr Verlag Tübingen.

Magistry, P. and Sagot, B. (2012). Unsupervized word segmentation: the case for Mandarin Chinese. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 383–387.

Magistry, P. (2013). *Unsupervised Word Segmentation and Wordhood Assessment*. Ph.D. thesis, Paris Diderot; Inria.

Muller, S. (2005). *Discourse markers in native and non-native English discourse*. John Benjamins Publishing.

Prévot, L., Gorisch, J., and Bertrand, R. (2016). A cup of cofee: A large collection of feedback utterances provided with communicative function annotations. In *Proceedings of 10th Language Resources and Evaluation Conference*, Portoroz.

Roze, C. and Danlos, L. (2011). Traduction (automatique) des connecteurs de discours. *TALN 2011, Montpellier*, (18).

Roze, C., Danlos, L., and Muller, P. (2010). Lexconn: A french lexicon of discourse connectives. *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010), Moissac, France*.

Schiffrin, D. (1988). Discourse markers. *Language*, (64):633–637.

Schourup, L. (1985). *Common discourse particles in English conversation*. New York: Garland.

Stali, M. (2015). *A corpus driven study: the use of dis-*

course markers during Twitch's streams. Master's Degree at Université d'Avignon.

Stali, M. (2016). *Les marqueurs discursifs genre et du coup: une étude comparative de corpus*. Master's Degree at Laboratoire Parole et Langage, Aix-Marseille Université.

Tseng, S.-C. (2013). Lexical coverage in Taiwan Mandarin conversation. *International Journal of Computational Linguistics and Chinese Language Processing*, 1(18):1–18.

Identifying Bilingual Topics in Wikipedia for Efficient Parallel Corpus Extraction and Building Domain-Specific Glossaries for the Japanese-English Language Pair

Bartholomäus Wloka
University of Vienna
Centre for Translation Studies
bartholomaeus.wloka@univie.at.at

Abstract

This paper presents an approach and its implementation as a software toolset for examining what portion of the multilingual content of Wikipedia is viable for harvesting bilingual data in order to build parallel corpora and domain-specific glossaries. An algorithm is presented which analyzes the link topology of topics and subtopics and the co-occurance in another language. This algorithm is implemented in the Python language and can be used to examine an arbitrary number of topics for Japanese-English as well as other language pairs with minor adjustements. The goal of the toolchain is ease of use and transparency as well as flexibility towards language combinations. The findings of a test with several thousands topics is presented as a showcase. The toolchain is open source under the Creative Commons license.

Keywords: automatic language resource harvesting, parallel corpora, data retrieval, wikipedia harvesting, multilingual comparison

1. Introduction

Wikipedia, as commonly known, is a conglomeration of articles written independently by people from all over the world. It is an extensive collection of knowledge represented in many languages. It is obvious that the content of these articles across these languages would vary in structure and semantics depending on the point of view of the particular country or region. However, there are also pages which are directly translated by groups of people who dedicate their time to make the articles consistent and to close gaps which might occur with certain topics across languages. Furthermore, some articles are written independently, but the information is often derived from articles in languages, where the content is more comprehensive. Often the English Wikipedia serves as a pivot for this purpose, given the status of English as lingua franca, and the fact that on Wikipedia English is represented more than any other language, as seen in Fig. 1. A probably even more representative quantifier for the overall activity in a given language, and perhaps an indicator for the quality of the content is the count of active contributors, i.e. contributors that have edited at least one article in the past month (Fig. 2). Here we see even more clearly how dominant the English Wikipedia is in terms of community contribution. Due to the activity of English, the English-Japanese language pair was chosen for this showcase.

The following chapters elaborate on related works (Chapter 2.), the approach of comparing multilingual content on Wikipedia articles in Chapter 3., the algorithm of comparison in Chapter 4., and its implementation in Chapter 5.. Chapter 6. presents a showcase of a comparison. Finally, a conclusion is presented in Chapter 7. as well as a description of future work plans in Chapter 8..

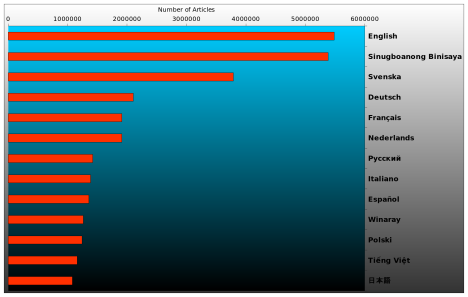

Figure 1: Count of Wikipedia articles by language

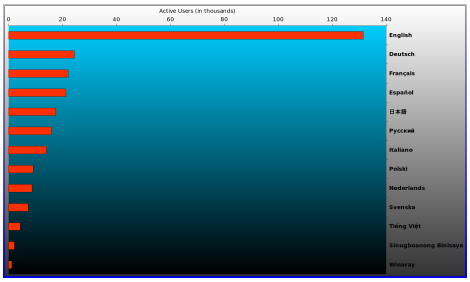

Figure 2: Count of active contributors for the largest Wikipedias

2. Related Work

Multilingual language resource extraction from Wikipedia is becoming increasingly popular, due to a consistent structure of Wikipedia articles. Comparable corpora have been built for many language pairs (Otero and López, 2010; Reese et al., 2010), especially for resource-scarce languages. However, the automatic generation of parallel, i.e.

sentence aligned corpora, across different languages is a very difficult task and requires evaluation and assessment of multilingual correspondence between articles (Paramita et al., 2012). (Ljubesic et al., 2016) shows that crawling text from Wikipedia results in acceptable results, measured by BLEU (Papineni et al., 2002). An approach to crawl depending on a domain is shown in (Labaka et al., 2016), by a breadth first search starting from a certain root-article and advancing via comparison with a domain specific dictionary and the assumption that occurrence of words from this dictionary signify that it belongs in this domain. These and other numerous works show promising results, however, it seems that their approach focuses mainly on few language combinations.This paper attempts to complement these results by a software toolchain which computes straight forward co-occurrence between articles, while not relying on machine translation or dictionary lookup, but on the structured architecture of topic IDs (Vrandecic and Krötsch, 2014), hence providing ease of use and flexibility regarding the language combination.

3. Comparing Wikipedia Articles

The goal of multilingual Wikipedia topic comparison is to find out whether the corresponding articles content can be regarded as a good translation. However, it is not feasible to compare each sentence of every article, especially if many language combinations are to be examined. Due to an exponential computational complexity, dictionary lookups, and the fact that multiple millions of tokens have to be processed, these methods have limits regarding multilingual flexibility. Furthermore, the task of automatic translation quality assessment of entire texts is in itself a computationally intensive process, since it often includes *machine translation* and/or *word sense disambiguation* as well as the use of external language resources such as lexical resources, terminology-bases, etc. The method described in this paper suggests a comparison, which selects articles by identifying the subtopics, hence omitting the problem of huge amounts of data. It uses the Wikipedia *pageID* property to identify articles of the same topic in different languages, which helps to avoid the usage of dictionary lookups, thus eliminating the problem of ambiguity and eliminates the need for additional language resources.

The process begins with choosing a topic and selecting its article page. This Wikipedia page is analyzed for all of the topic links mentioned in this article. Next, all the topics which were found in the first step are analyzed in the same way resulting in a collection of tuples of topics and subtopics. This process is repeated for the second language. Finally, the two lists of tuples are compared to each other and the co-occurring topics/subtopics are counted. The articles with a high percentage of co-occurrences indicate potentially similar content and indicate candidates for bilingual language harvesting and parallel corpora creation.

In addition to identifying parallel corpora candidate pages the algorithm finds the equivalents terms for each topic, which results in a term glossary in a certain domain, depending on the starting point, i.e. the initial topic.

4. Comparison Algorithm

The algorithm presented in this paper starts with harvesting topic links within one article page. Wikipedia article pages are well structured and consistent, so the topic acquisition is easily achieved via extraction of all *href links* with the *title* tag. These tags are stored in a *tuple* (data structure which stores two objects of data), and the tuples themselves are stored in a list. Once all topics of the current page are collected, the first subtopic of the first article, i.e. the second element of the first tuple in the list is used as the main topic and all its subtopics are extracted in the same manner. This process is repeated until the list of tuples of the initial list is exhausted. A formal representation of the algorithm is noted below.

```
function extract(topic)
   for all subtopics in topic
      extract subtopic
      add topic, subtopic to tuple
      store tuple in list
   return list of tuples
call function extract(topic)
for all subtopics in tuple
   call function extract(subtopic)
```

This process is done for the starting topic in English and its equivalent topic in Japanese. These two lists are then compared for co-occurance. In order to do so, the Japanese topics are translated by finding the equivalent topic via the Wikipedia API using the pageID property. Each Wikipedia page has a json file associated with it, which contains the set of all available language representations of this topic. This is slightly different and in some cases more precise than a direct translation of the word, describing this concept, since this translation is focused on the concept, rather than a dictionary entry, which may present several options, or be very generic.
After getting the English topic equivalents for the Japanese list, the two lists are compared for co-occurance of topics. At the same time this results in a glossary for this list of topics.

5. Algorithm Implementation

The implementation of the algorithm described in Chapter 4. is done with the Python language. The modules used in the toolchain are: *BeautifulSoup*, for extraction of data from HTML, *requests* for HTTP access, *json* for reading pageID's from Wikipedia's API, and *re* for string comparison with regular expressions. The source code is open source under the Creative Commons License, and is available from the author upon request.

6. Showcase

The algorithm described in Chapter 4. is used to examine three topics and all of their subtopics up to the second level. The starting articles are *cat*, *language*, and *airplane*. The corresponding Japanese articles are ネコ, 言語, and 飛行機. The output of the results and intermediate results for this starting topic *airplane* are described in this chapter. The article *airplane* yielded 406 topic entries, while their

subtopics resulted in a combined total of 62,634 topics. A sample of the output for the first 40 topics found in English and Japanese are shown in Fig. 3.

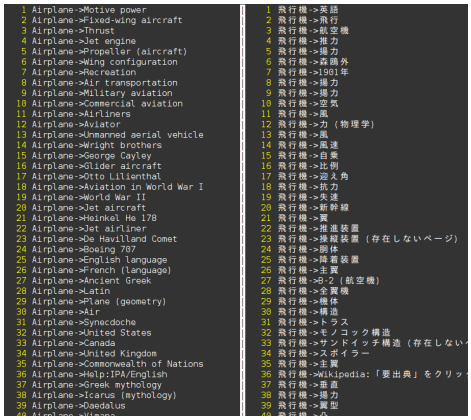

Figure 3: Output from collecting topics

In the next step, the equivalent English topics for each Japanese topic is found, as described in Chapter 4.. Furthermore, the number of common topics is found in these lists. A high occurance of co-occuring topics indicates a high overlap of information and indicates a potential topic collection for parallel corpus harvesting.

Figure 4 shows the first 50 entries of the Japanese topics with their English Wikipedia concept counterparts. Apart from being the basis of comparison of topic overlap, this collection is a glossary in a domain that stems from the initially chosen topic. This way, a glossary of any spesific topic domain can be compiled quickly and dynamically.

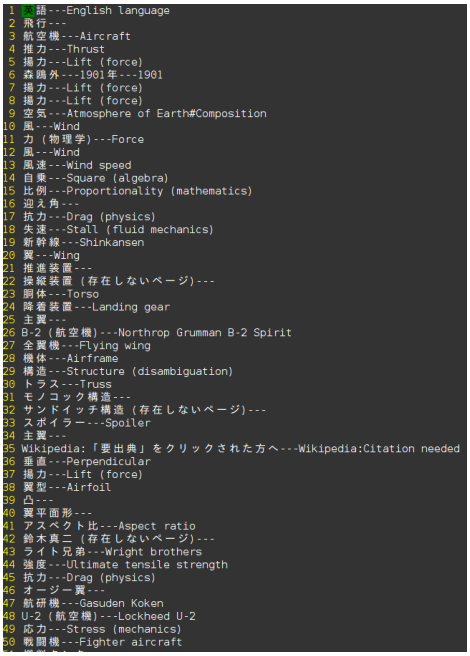

Figure 4: Output from collecting topics

7. Summary

This paper presents a method for extracting Wikipedia articles and all its subtopics up to the second link level for the English-Japanese language pair and is extensible to other language pairs. A showcase of a topic search is presented as an example. Since this is a work in progress, there are no exact numbers yet on the precise topic overlap, although first samples indicate promising results. In the process of analyzing the topic co-occurance several domain specific terminology glossaries have been produced.

8. Future Work

It is planned to analyze large amounts of topic collections to identify parallel corpus harvesting candidates across Wikipedia. Further it is planned to identify and process article sections for parallel corpus extraction. Additionally, the toolchain will be expanded with a graphical user interface, which will make it easy and intuitive to use. A graphical implementation of the step to build glossaries will be tested at the Centre for Translation studies in a class room setting.

9. Bibliographical References

Labaka, G., Alegria, I., and Sarasola, K. (2016). Domain adaptation in mt using titles in wikipedia as a parallel corpus: Resources and evaluation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Ljubesic, N., Espla-Gomis, M., Toral, A., Rojas, S. O., and Klubicka, F. (2016). Producing monolingual and parallel web corpora at the same time - spiderling and bitextor's love affair. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Otero, P. G. and López, I. G. (2010). Wikipedia as multilingual source of comparable corpora. In *In Proceedings of the LREC Workshop on BUCC*, pages 30–37.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Paramita, M. L., Clough, P., Aker, A., and Gaizauskas, R. (2012). Correlation between similarity measures for inter-language linked wikipedia articles. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Reese, S., Boleda, G., Cuadros, M., Padró, L., and Rigau, G. (2010). Wikicorpus: A word-sense disam-

biguated multilingual wikipedia corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 19–21, Valletta, Malta, may. European Language Resources Association (ELRA).

Vrandecic, D. and Krötsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, pages 78–85.

Creating Comparable Corpora through Topic Mappings

Firas Sabbah, Ahmet Aker
Department of Information Engineering, University of Duisburg-Essen
{sabbah, aker}@is.inf.uni-due.de

Abstract

Aligning multilingual documents is considered one of the most important steps in building comparable and parallel corpora. Bilingual lexicons have commonly used to detect the similarity level between the bilingual documents. However, high quality bilingual lexicons are not free and not readily available for many language pairs. In this work, we present a new approach to detect the similarity level between documents written in two different languages. The basic idea is to analyze the topical structure of texts and use it for detecting the similarity level between the documents. The results show that enhancing the lexicon-based methods by the topical structures improves the alignment process. Besides the model, this work introduces a tool for automatic comparable document search for English-Arabic languages.

Keywords: Comparable Corpora, Document Alignment, LDA, topic mapping

1 Introduction

In many cases an event is captured by many new agencies and reported in diverse languages. Being able to track all news about the same event opens many doors for different kind of analyses such as understanding how different countries observe the event, what are their agreement and disagreements in terms of argumentations, what are the reactions of respective readers[1], where are the topical focuses, etc. to name just few.

In our broader research agenda we have the vision to perform multi-lingual argument mining and perform analyses about the differences and commonalities between the arguments found in two different articles reported in two different languages. Our current focus is in English and Arabic. To perform this there are several steps: (1) determining comparable documents, (2) annotating both articles for arguments, (3) aligning arguments and finally (4) making sense of the aligned arguments. The focus of this paper, however, is at step 1 which is also the backbone of the later tasks.

Two documents written in two different languages are comparable if they talk about the same topic or event. Related work (see Section 2 for details) have investigated different ways for obtaining comparable corpora – data collection containing large sets of comparable documents.

In our work we focus on topic mappings. For this we use the Latent Dirichlet Allocation (LDA) to extract the topics of both source and target documents. Each topic is represented by a set of key words. We do this for each language separately. Then we map topics which result in a topic dictionary allowing us to query with source language topics and obtain topics in the target language. With this our approach becomes independent of translation sources which would be needed for translating source topics to target key words. However, we also extract traditional translation based features to boost the alignment performance. Both topic mappings and translation based features are combined to determine the similarity level between two documents written in two different languages. We integrated this solution into a tool enabling users to search for documents in the source language and also automatically retrieve documents in the target language which are comparable to the source documents.

In Section 2 we discuss related work. Next, in Section 4 we introduce our method of aligning the documents. We provide the evaluation results in Section 5. Next in Section 6 we present the tool for automatic comparable document search. Finally, we conclude the paper in Section 7.

2 Related work

Indeed, many approaches for creating comparable corpora were proposed. A common paradigm for obtaining a comparable corpus involves collecting monolingual data for each language and matching documents by comparing document contents (Talvensaari et al., 2007; Hashemi et al., 2010; Aker et al., 2012). These methods have one common aspect; they extract the top keywords of an English text, perform automatic translation of these to the target language and perform the pairing based on the source and translated key words.

Other studies(LU et al., 2013; Kraaij et al., 2003) use the page structure and URLs to detect the similarity level between the documents. The idea of similarity in these studies is that the HTML structure and the document path URL of the source and the target documents have to share an acceptable level of symmetry.

Since Wikipedia supports the inter-language links for its articles, we notice the intensive usage of such resource to produce such corpora (Adafre and De Rijke, 2006; Saad et al., 2013). These studies focus on how to measure the quality of similarity between the Wikipedia pages, and to select the similar articles for building comparable corpora.

Topical structures have been also used for building comparable corpora. (Zhang et al., 2013) propose a model to mine bilingual topics from Wikipedia in order to tackle the problem of cross-lingual linking. In this study the similarity is a score computed by the inner product of topic distributions of the documents. (Zhu et al., 2013) uses also the topics of documents to measure the similarity. The similarity value is calculated using three different measures: Kullback-Leibler (KL), Cosine Similarity and Conditional Probability. For these measures, the similarity is defined by the closeness of

[1] This assumes that each article has available reader comments.

a document to a specific topic.
In our work, however, we focus on topic mappings. The topic mappings do not rely on translation sources and are a way of bridging two articles written in two different languages. With this if a user determines topic of a source document she can easily query from the mappings how to express that topic in the target language and use the expression to look for documents expressing the mapped topic. We use this idea to align two documents written in two different languages. However, to boost the performance of the alignment we also make use of simple and light translation features and combine those together within an SVM classifier.

3 Data

For our targeted languages (English and Arabic), we extracted document collections from HuffingtonPost website [2]. However, HuffingtonPost is not the only news website that offers news in many languages. Tens of news agencies also offer multilingual news like BBC and Reuters. What makes HuffingtonPost different than other news agencies is that some HuffingtonPost news contains a specific phrase or link that leads to another version of HuffingtonPost that contains a near translation of the first article.

3.1 Collection method

For crawling the articles from HuffingtonPost we proceed the following steps: 1) We automatically track the news articles from the twitter page of the target language (Arabic-HuffingtonPost), 2) we check whether the news article has a parallel English version by searching the article text for specific phrases that indicate that news page is originally published by another source (this include phrases like "This article is translated from ..." or "this topic is originally published ..."), finally 3) we extract the texts of both documents.

3.2 The collected data

We crawled the articles over the period from July 2015 to July 2017. Over two years, Arabic-HuffingtonPost had published about 3543 Arabic Articles that have nearly parallel English versions. Table 1 presents a detailed information about the crawled data. The crawled collections cover political, sport, science, technology as well as life style domains. The data[3] is publicly available on GitHub.[4]

English articles	3543
Arabic articles	3543
Total number of English words	2320583
Total number of Arabic words	2153295
Total number of unique English words	74255
Total number of unique Arabic words	154957

Table 1: Collected data specifications

4 Methodology

In order to detect comparable documents we make use of topic mappings between source and target languages. Given a pair of documents (English-Arabic), we extract LDA topics from both documents.
Next, we measure how strongly the topics correlate and decide based on this how strongly comparable the pair of documents is. However, since the LDA topic extraction is performed independently on each document and the topic-describing words are written in two different languages it is not straightforward to compute the topic similarities. One way of doing this is through using dictionaries for translating from one language to other and compute a similarity metric over them. Another way is to generate topic mappings and use them instead of translation dictionaries. In this work we adopt the latter approach.
In the next sections we describe how we create our topics and the topic mappings. We also describe how the mapping information is transferred to features to perform the alignment process. In addition to mapping information, we also make use of traditional features which are also outlined in this section. Figure 1 presents an overview of our methodology phases.

4.1 Training LDA models

LDA (Blei et al., 2003) is a statistical unsupervised learning algorithm. It generates a distribution of how objects constitute hidden themes and how different objects constitute observable entities. LDA regards the hidden topics as a group of tightly co-occurring words.
We learn LDA models for both English and Arabic documents described in Section 3, extract topics from both document collections and align the topics. To align the topics there is an assumption that the pair of documents have an acceptable level of comparability which is the case for the HuffingtonPost data.
Before training, we pre-processed our collected dataset by removing the stop words from both languages, and applying further text processing on the Arabic dataset. To train an LDA model over a dataset, we have to know the following variables: First, we need to decide the number of topics which should be used to produce the best words divisions. Of course, the number of the topics is pertinently related to the dataset size, more documents in the dataset means more vocabulary which implies more topics. Therefore, the number of topics is determined experimentally. From the experiments, we find that the topics number around 70-75 is giving us the best results within HuffingtonPost dataset.
Secondly, we need to decide on the values of LDA parameters alpha and beta. The document-topic density is represented with alpha, the higher alpha, the more topics documents contain. The topic-word density is represented with beta. The higher beta, the more words from the corpus a topic contains. After experiments, we find that using 1.0 for alpha and 0.1 for beta is providing the best division of

[2] https://www.huffingtonpost.com

[3] Due to copyright issues we only publish the URLs to the English and Arabic articles.

[4] https://github.com/fsabbah/lda-comparable-corpora/blob/master/en_ar_URLs.csv

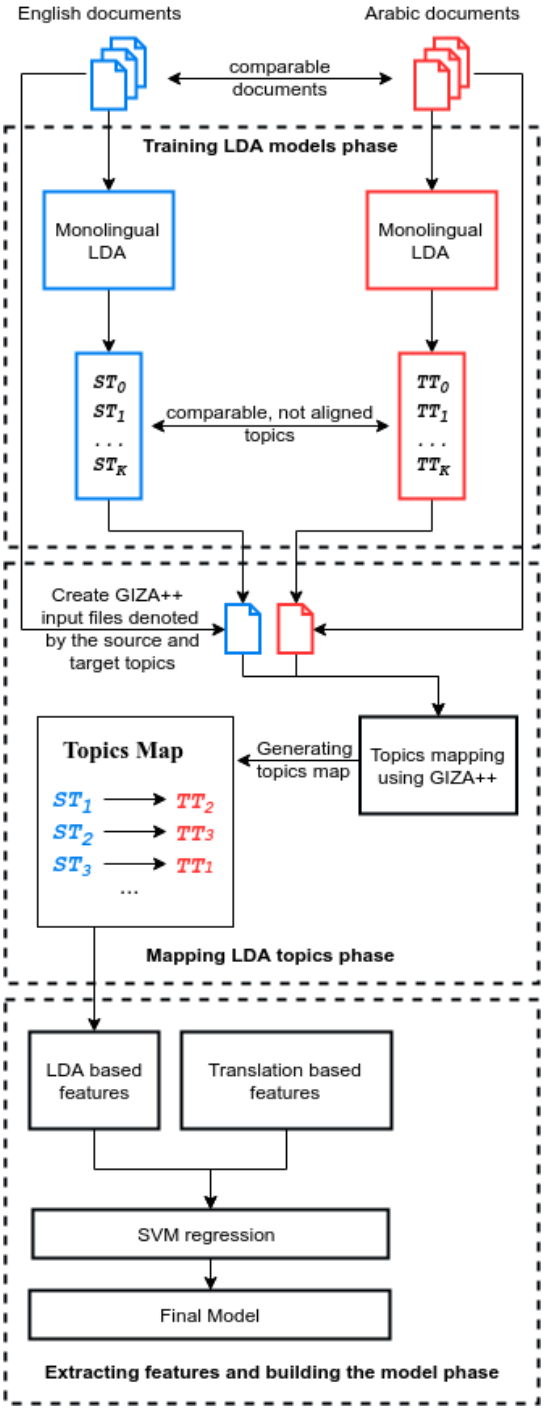

Figure 1: Methodology phases

topics. To perform topic modeling, we used the LDA implementation within the mallet project[5].

[5] http://mallet.cs.umass.edu/

4.2 Mapping LDA topics

The training phase of LDA produces two sets of topics, one for each language. The topic mapping process aims to match the LDA topics of the source and the target languages. For matching topics, we use GIZA++ tool. Since each pair of English and Arabic documents are translation of each other or strongly comparable, we can assume that they share exactly the same or similar topics, just expressed in different languages. Our aim is to find the topic mappings and use this knowledge for finding comparable corpora. For this purpose we create analog to parallel sentence files parallel topic files where the English file contains the ids of the English topics and the Arabic file contains the ids of the Arabic topics. The files are line-aligned where a pair of English-Arabic lines represent a pair of English-Arabic documents. In our approach, we use topics which have at least 5% probability according to LDA. To also express the frequency of a topic or its coverage within a document in each line, we repeat the topic id according to its probability in the original document. For example, if we have probability of 80% of a topic within a document, then we repeat for the document line the topic id eight times in case LDA topics number is selected as K=10.

In its original setting Giza++ produces words alignment. In our case the words are topics. Using this GIZA++ output, we are able to build a mapping matrix between the source and the target topics. Table 2 presents an example map between topics.

	ST_0	ST_1	...	ST_k
TT_0	0.12	0.29	...	0.02
TT_1	0.81	0.05	...	0.01
...	...	...	...	...
TT_k	0.49	0.28	...	0.03

Table 2: Alignment of source and target topics. TT stands for target topic and ST for source topic.

As we see in Table 2 each source/target topic is aligned with every target/source topic. Each alignment is associated with a probability score which is computed by GIZA++. With this matrix it is possible to obtain for a given source topic all target topics which are above a specific probability, determine target documents entailing those topics and based on results make statements about the similarity between the source document and the determined target documents.

Table 3 presents examples of the aligned pairs of topics. These topics contain only the top 20 terms per topic.

4.3 Extracting features and building the model

We use Support Vector Machines (SVMs) with a linear kernel and the trade-off between training error and margin parameter C = 1 for the alignment purposes. Within the classifier, the used features are extracted from the trained LDA models and their topics mapping for the 3366 near parallel articles. Furthermore, we also make use of features extracted using a home-trained GIZA++ dictionary.

English topic	Best aligned Arabic topic	Translation of the Arabic topic
sugar, diet, fat, weight, foods, eat, eating, healthy, health, food, calories, high, body, drinks, risk, energy, low, per, blood, protein	تناول، سكر، غذائية، وزن، طعام، اطعمة، غذائي، جسم، صحية، دهون، تحتوي، نظام، حرارية، كمية، نسبة، سعرات، قلب، تغذية، اصابة، صحي	eating, sugar, food, weight, food, foods, body, healthy, fat, contain, system, calories, quantity, ratio, calories, heart, nutrition, injury, healthy
trump, president, trumps, donald, house, white, obama, washington, campaign, former, election, elect, administration, york, national, presidential, bush, presidency, office, america	ترامب، رئيس، اوباما، اميركية، اميركي، ولايات، دونالد، ابيض، اميركا، بيت، اميركيين، واشنطن، ادارة، جمهوري، لترامب، منتخب، بوش، رئاسة، انتخابية، حملة	trump, president, obama, american, american, states, donald, white, america, house, americans, washington, administration, republican, trump, team, bush, presidency, electoral, campaign
iraq, isis, iraqi, mosul, city, forces, islamic, baghdad, state, sunni, shia, war, saddam, battle, fighting, falluja, government, kurdish, people, iraqs	تنظيم، داعش، عراق، دولة، قوات، اسلامية، مدينة، قاعدة، موصل، معركة، عراقية، ابو، عراقي، ميليشيات، بغداد، عمليات، حسين، حرب، عبد، سوريا	organization, isis, iraq, state, forces, islamic, city, base, mosul, battle, iraqi, abu, iraqi, militias, baghdad, operations, hussein, war, abdul, syria

Table 3: English and Arabic topics represented by top 20 LDA words.

4.3.1 LDA-based features

The procedure of extracting the LDA based features proceeds the following steps: 1) for each document in the training dataset, we fetch the top LDA topics from the trained LDA model, 2) we connect each document in the source dataset to two documents in the target dataset (correct and incorrect target documents), 3) from each connection, we extract four features related to each top LDA topic.

To fetch the top LDA topics of a document we infer the probabilities of topics from a document. We sort the topics according to their probabilities. After that, we define two relationships between the source document and the target documents. The first one represents a positive case; it represents a connection between the source document and its correct aligned target document. The second represents the negative case; it is a link between the source document and a random document from the target dataset. We make sure the random document must not be the aligned target document of that source document. That means we create one correct connection and another wrong connection. For each connection, we extract the following features:

1. The probability of the top LDA source topic.
2. The probability of the top aligned target LDA topic (we find the top aligned target topic using the GIZA++ topics mapping).
3. The probability of the top LDA target topic.
4. The probability of the top aligned source LDA topics to that target topic.

Figure 2 shows the process of extracting the features of the top LDA source and target topics.

However, using only the features of the top topic is not enough to capture all topics within a document. To solve this, we used the top ten LDA topics. As described in the procedure above, we extract the same four features for each of these top 10 topics leading to 40 features in total.

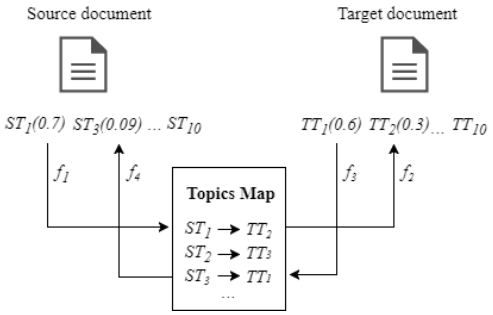

Figure 2: LDA topics features

4.3.2 Translation-based features

In order to improve the accuracy results, we added more features to the LDA-based features described above. This time we extracted the features from the texts. Since we need to determine the similarity between two different texts written in two different languages, we need to convert these texts or parts of them into one language. A translation system is the perfect tool used in these cases.

However, translation systems are not readily available. To overcome this problem, we used a home-trained GIZA++ dictionary. The parallel resources used for training and building this dictionary are brought from the OPUS project[6]. The main idea of using translation is to find how many similar words are shared between different parts of the source and the target texts. Such parts include titles, first and second sentences of both documents. In addition, we extracted also the most important 20 words of each document by calculating tf*idf values of the documents words. As we need numerical values for the classifier, we use cosine similarity to define a numerical value of the similarity between the original texts and the translated texts. As a result, we created the following features:

[6] http://opus.lingfil.uu.se/

1. The cosine similarity between the source document's title and the translated title of the target document,
2. The cosine similarity between the target document's title and the translated title of the source document,
3. Repeating these also for the first and second sentences in the source and target documents,
4. The cosine similarity between the top 20 tf*idf words of the source text and the translated top 20 words of the target text and
5. As in feature 4 with changing the direction of translation, i.e., from target text to source text.

In total, we collected 48 features, 40 from LDA topics and eight based on GIZA++ dictionaries. We set the similarity value 1.0 for each correctly aligned pair of documents, 0 for the connections that are not correctly paired.

5 Evaluation

For evaluation purposes we again use the huffingtonPost data described in Section 3. We split this data into a training (3366 articles) and a testing (177 articles) set. The training data is used to extract topic models and later to create the topic mappings (see previous section).
To evaluate our approach, we perform an automatic evaluation on the testing data. We compare LDA based features against the translation-based ones. In our evaluation we pair each English document with every Arabic document resulting in 177 pairs for each English document. Note among these 177 pairs there is only one pair that is correct. For all pairs features are extracted and SVM used to rank them. The document pair that is ranked top is evaluated whether it is the correct pair. If yes then we have a positive hit otherwise negative. Once we have repeated this for every English document we compute the accuracy scores which is the ratio of positive hits to all hits. Results are shown in Table 4. Note that the table shows only accuracy figures of the translation features. From the table we can see that best results are obtained when all translation-based features are combined. The combined translation-based features lead to close 69% accuracy.

Experiment	SVM classifier
Title	29.94%
Title + First sentence	40.67%
Title + First sentence + Second sentence	44.63%
20-top ranked tf*idf words	50.84%
Title + 20-top ranked tf*idf words	62.71%
Title + 20-top ranked tf*idf words + First sentence + Second sentence	68.92%

Table 4: Accuracy of the translation-based features

Figure 3 presents the results of the LDA-based features. Unfortunately LDA based features are not able to outperform the translation based features and achieve maximum

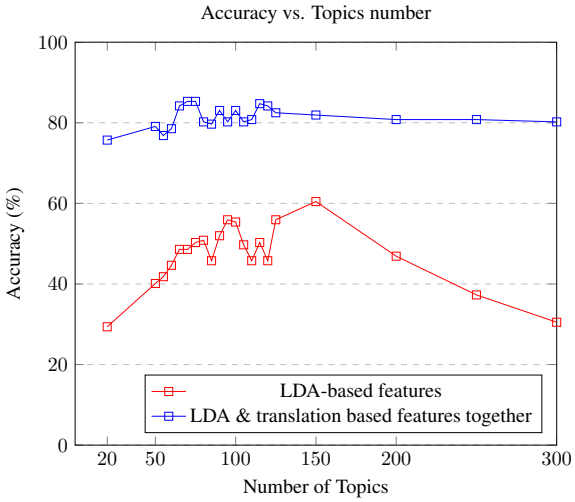

Figure 3: Accuracy of LDA-based features (alone and combined with translaiton based features).

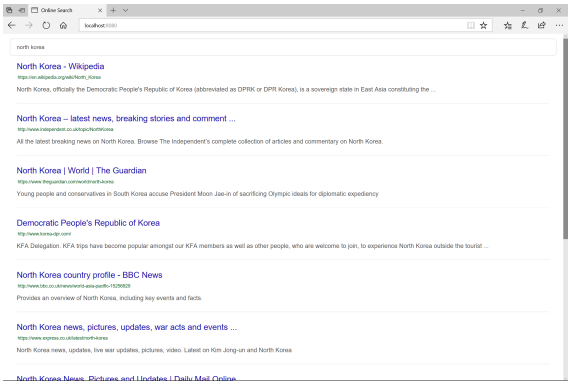

Figure 4: Search results for the English query.

60% accuracy (with K=150). However, we see that the LDA based features boost the results when they are combined with the translation ones. This is again shown in Figure 3 but this time with K=70-75. The combined approach leads to an accuracy of around 85%. This means in 85% the case our alignment is able to capture the correct target document of each source one. In our tool we use this combined approach to align documents.

6 Tool for comparable document search

Our current tool supports the gathering of Arabic documents comparable to an English document. The system allows users to enter English queries to search for English documents. The tool uses the Bing search API to search the web. The retrieved English documents are shown in a list similar to a search engine result list (see Figure 4).
Within the tool the user can select any English document and preview it before asking for comparable documents. To find the comparable Arabic documents, the tool first translates the title of the selected English document using an in

house created GIZA++ dictionary and uses the translated title to query for Arabic documents. Once the Arabic documents are retrieved it applies the alignment method described earlier to rank them. The user can then select the Arabic documents to display – this time the English and the Arabic document are displayed side by side (see Figure 5).

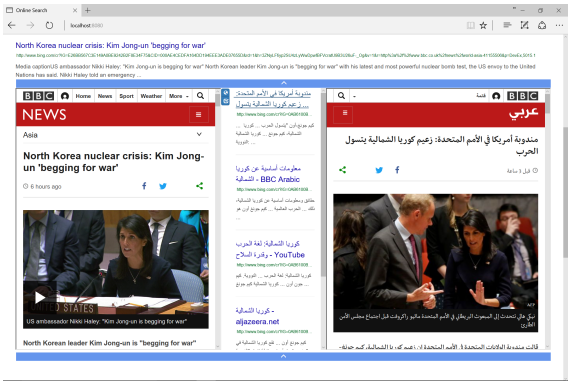

Figure 5: Documents are displayed side by side

7 Conclusion and Outlook

In this work we described a new approach for aligning English and Arabic documents for the purpose of comparable corpora construction. The proposed approach make use of LDA topics to analyze the topical structures of the documents. Based on the LDA topics we created topic mapping dictionary to automatically transfer a set of key-words describing the topics within the source document to the target language and use the transferred knowledge to judge whether two documents written in English and Arabic are comparable. Besides the topical mappings, we also use the traditional translation-based features to boost the alignment performance. Our results show that topic mappings as well as traditional features alone have performance around 60% to 70% accuracy. However, when both are combined the performance increases to the 80% level. We also integrated our alignment approach within a search tool that enables users to search for English documents, select an English document and retrieve Arabic documents comparable to the selected English document. The Arabic documents are ranked according to how comparable they are to the selected English document. In both cases the tools lets the user to read the articles.

In future we plan to further work on our vision to have a complete tool that supports multi-lingual argument mining. We will enhance our current tool with state-of-the-art argument mining approaches to determine arguments in the English and Arabic documents. However, due to the lack of argumentative training data for the Arabic language we will use for now only English argument mining solutions, tag English arguments and investigate mappings of those English arguments to the Arabic document. In terms of argument mapping we will follow the strategy discussed in (Aker and Zhang, 2017). However, in close future we aim to construct Arabic argument mining solutions using the data collection idea described in (Sliwa et al., 2018).

8 Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

9 Bibliographical References

Adafre, S. F. and De Rijke, M. (2006). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Aker, A. and Zhang, H. (2017). Projection of argumentative corpora from source to target languages. In *Proceedings of the 4th Workshop on Argument Mining*, pages 67–72.

Aker, A., Kanoulas, E., and Gaizauskas, R. J. (2012). A light way to collect comparable corpora from the web. In *LREC*, pages 15–20. Citeseer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Hashemi, H. B., Shakery, A., and Faili, H. (2010). Creating a persian-english comparable corpus. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 27–39. Springer.

Kraaij, W., Nie, J.-Y., and Simard, M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29(3):381–419.

LU, Y., ZHANG, X., and ZHENG, D. (2013). Automatic english-chinese parallel corpus acquisition and sentences extraction. *Journal of Computational Information Systems*, 9(6):2365–2372.

Saad, M., Langlois, D., and Smaïli, K. (2013). Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia-Social and Behavioral Sciences*, 95:40–47.

Sliwa, A., Ma, Y., Liu, R., Borad, N., Fatemeh Ziyaei, S., Ghobadi, M., Sabbah, F., and Aker, A. (2018). Multilingual argumentative corpora in english, turkish, greek, albanian, croatian, serbian, macedonian, bulgarian, romanian and arabic. In *Proceedings of LREC 2018*.

Talvensaari, T., , J. L., Jarvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)*, 25:4.

Zhang, T., Liu, K., Zhao, J., et al. (2013). Cross lingual entity linking with bilingual topic model. In *IJCAI*.

Zhu, Z., Li, M., Chen, L., and Yang, Z. (2013). Building comparable corpora based on bilingual lda model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 278–282.

Detecting Machine-translated Subtitles in Large Parallel Corpora

Pierre Lison, A. Seza Doğruöz
Norwegian Computing Center, Independent Researcher
Oslo, Norway, Turkey
plison@nr.no, a.s.dogruoz@gmail.com

Abstract

Parallel corpora extracted from online repositories of movie and TV subtitles are employed in a wide range of NLP applications, from language modelling to machine translation and dialogue systems. However, the subtitles uploaded in such repositories exhibit varying levels of quality. A particularly difficult problem stems from the fact that a substantial number of these subtitles are not written by human subtitlers but are simply generated through the use of online translation engines. This paper investigates whether these machine-generated subtitles can be detected automatically using a combination of linguistic and extra-linguistic features. We show that a feedforward neural network trained on a small dataset of subtitles can detect machine-generated subtitles with a F_1-score of 0.64. Furthermore, applying this detection model on an unlabelled sample of subtitles allows us to provide a statistical estimate for the proportion of subtitles that are machine-translated (or are at least of very low quality) in the full corpus.

Keywords: Parallel corpora, Machine Translation, Quality Estimation

1. Introduction

The availability of movie and TV subtitles for a large number of languages and linguistic genres makes them particularly useful for the construction of parallel corpora. Currently, the largest collection of subtitles is the OpenSubtitles corpus with 3.35 billion sentences covering 60 languages (Lison et al., 2018). In addition to their textual content, subtitles are also associated with precise timestamps indicating when each subtitle block should be displayed. These timestamps allow subtitles to be efficiently aligned across languages based on time overlaps (Tiedemann, 2007; Tiedemann, 2008). These time-based alignments can in turn be used to extract multilingual parallel corpora.[1]. In addition to the OpenSubtitles corpus, other corpora based on subtitles include the SUMAT data collection (Petukhova et al., 2012), the collection of dual subtitles from (Zhang et al., 2014), the Tehran English-Persian parallel corpus (Pilevar et al., 2011) and the Japanese-English subtitle corpus (Pryzant et al., 2017).

From a linguistic perspective, parallel corpora derived from subtitles are appealing due to their coverage of a broad range of conversational genres and speaker styles. Subtitles are also widely used in practical NLP applications, notably for neural and statistical machine translation (Belinkov and Glass, 2016; van der Wees et al., 2016; Wang et al., 2017), but also conversational modelling (Lison and Bibauw, 2017; Krause et al., 2017), semantic role labelling (Akbik et al., 2016) and distributional semantics (Lison and Kutuzov, 2017; Speer and Lowry-Duda, 2017).

Despite their benefits, corpora extracted from subtitle repositories also have some shortcomings. The most important issue is the varying quality of the subtitles in terms of linguistic fluency, faithfulness to the dialogues in the source material (movie or TV episode), and adherence to formatting guidelines. Subtitles made available in online repositories such as OpenSubtitles[2] are typically created by movie and TV fans rather than translation and subtitling professionals. An important portion of subtitles are not even produced by human translators at all (professional or not) but are merely generated using online machine translation engines based on other existing subtitles. The linguistic quality of these machine-generated subtitles is typically quite low, as they are typically left unedited and contain numerous grammatical and translation errors.

This paper presents a machine learning model for detecting such machine-translated subtitles based on a combination of linguistic and extra-linguistic features. Despite the difficulty of the detection task, the model achieves a reasonable performance and can be used to either filter out low-quality subtitles from the corpus or assign them with a document weight that can be passed to downstream applications.

The rest of the paper is organised as follows. The next section reviews related work on detecting machine-translated texts. Section 3 presents the dataset employed in this paper and provides several examples of translation errors observed in machine-translated subtitles. Section 4 defines the linguistic and extra-linguistic features that can be employed for detecting such subtitles. Section 5 details the empirical evaluation and error analysis of the approach. Section 6 shows how the detection model can be used to estimate the number of machine-translated subtitles in the full corpus and ultimately enhance the overall corpus quality. Finally, Section 7 concludes the paper.

2. Background

The comparison between machine-translated, human-translated and "original" (non-translated) texts has been the subject of numerous studies in translation studies and machine translation research. Translated texts can often be distinguished from non-translated texts due to interferences from the source language (where some aspects of the source language "spill" onto the translation output) combined with artifacts of the translation process that are independent of the source language (Koppel and Ordan, 2011). In particular, human-translated texts often make use of a more "standard" language than original texts (Toury, 1995), allowing them to be detected automatically (Kurokawa et al., 2009).

[1] http://opus.nlpl.eu/OpenSubtitles2018.php
[2] http://www.opensubtitles.org

Language	**Number**				
English	669	Swedish	46	Chinese (simplified)	10
Indonesian	580	Danish	45	Tamil	10
Spanish	519	Russian	43	Norwegian	10
Portuguese (Brazilian)	462	Serbian	41	Catalan	7
Romanian	327	Slovenian	40	Thai	6
Hebrew	326	Malay	37	Chinese (traditional)	6
Turkish	269	Albanian	35	Esperanto	5
Bulgarian	220	Dutch	31	Bengali	4
Arabic	193	Vietnamese	29	Basque	4
Polish	127	Ukrainian	26	Finnish	4
Persian	101	Japanese	23	Lithuanian	3
Portuguese	100	Hungarian	22	Korean	3
Italian	98	Estonian	18	Galician	2
Croatian	97	Slovak	17	Macedonian	2
German	92	Sinhalese	15	Malayalam	1
French	82	Hindi	14	Tagalog	1
Czech	79	Bosnian	11	Urdu	1
Greek	76	Telugu	10		
				Total:	**4 999**

Table 1: Number of subtitles explicitly marked with a "machine-generated" flag in the OpenSubtitles corpus, distributed by subtitling language.

This standardisation make them well-suited for language modelling (Lembersky et al., 2012). The term "*translationese*" is often used to refer to these peculiarities of translated documents compared to non-translated ones.

In comparison with human-translated texts, machine-translated documents are of course subject to various type of translation errors (Vilar et al., 2006; Stymne and Ahrenberg, 2012) that degrade the quality of the resulting texts. Arase and Zhou (2013) presented a data-driven approach aimed at detecting low-quality translations in web texts, using monolingual corpora only as input. Their features specifically focused on "phrase salads" in which the phrases of sentences are correct in isolation but become inaccurate when put together as a complete sentence. Aharoni et al. (2014) described a related approach and found a correlation between the performance of the machine learning model and the human evaluation of translation quality.

The two aforementioned approaches focused on specific language pairs for which large quantities of in-domain data is either already available or can be generated. In contrast, the detection model presented in this paper aims to be applicable to any language pair, without relying on the occurrence of translation errors specific to a given source or target language. Indeed, as explained in the next section, machine-generated subtitles can be found in virtually every language present in the corpus. Furthermore, these subtitles do not include any information about the subtitle it was translated from, nor even the source language. The detection model must therefore scale to a broad spectrum of possible language pairs while relying on a relatively small number of parameters (due to the modest amount of machine-generated subtitles available for training).

It should also be noted that machine-generated subtitles have been present in subtitle repositories since the early 2000s. As a consequence, the translations are a result of a broad mixture of translation tools, from early rule-based MT systems to modern APIs for statistical and neural machine translation. This leads to large disparities in the translation quality and typical error patterns observed in these subtitles. This stands in contrast with the aforementioned approaches which only relied on translations generated from specific, well-optimised statistical machine translation systems to train and evaluate their models.

3. Data

3.1. Subtitle corpus

The data employed in this paper comes from the latest version of the OpenSubtitles corpus released as part of the OPUS corpus repository (Tiedemann, 2012; Lison et al., 2018). The latest release comprises 3.73 million subtitles[3] in 60 languages. Each subtitle is converted into Unicode, segmented into sentences and tokenised according to the procedure outlined in (Lison and Tiedemann, 2016). For each language pair, subtitles associated with the same movie or TV episode (identified through their IMDB identifier[4]) are aligned at the sentence level, based on the respective timestamps of the two subtitles (Tiedemann, 2008). This alignment procedure leads to a total of 1 782 bitexts (language pairs must share at least one common movie or TV episode in order to form a bitext).

In addition to the tokenised sentences, each subtitle is also enriched with meta-data information regarding the movie or TV episode (release year, genre, original language) and the subtitle itself (upload date, user ratings, etc.). Unfortunately, we do not have any direct information about who

[3] In this paper, we use the term "subtitle" to refer to the whole file that contains the transcriptions for a given movie or TV episode. Each subtitle is itself composed of many (up to several thousands) subtitle blocks, where each block contains at most two lines of text and is associated with a start time and end time.

[4] http://www.imdb.com

created a given subtitle or for which purpose it was created. Some subtitles are created from scratch by fans, while others are "ripped" from official DVD releases or TV streams (which can sometimes be inferred from the presence of OCR errors in the subtitles). Yet another subset of subtitles are translations from other existing subtitles. For instance, a movie fan might wish to create a Spanish subtitle for a Japanese movie, but, not being fluent in Japanese, might opt for translating from an existing English subtitle instead of creating the subtitle from scratch. The translation quality of these subtitles is uneven at best, especially when translated with the help of online translation engines and left unedited. This is especially the case for subtitles uploaded before 2010, at a period where machine translation engines were of a much lower quality than today.

To address these quality issues, the administrators of the OpenSubtitles website have asked their users to mark machine-generated subtitles with an explicit flag when uploading new subtitles. However, only a small fraction of the machine-generated subtitles have so far been annotated with this flag (4 999 subtitles in total) as users are reluctant to declare that their uploaded subtitles are of lower quality. Table 1 illustrates the distribution of these subtitles by language.

3.2. Translation issues

One reason for this particularly low quality of machine-generated subtitles stems from the fact that, with the possible exception of documentaries, subtitles are conversational in nature and typically contain many short-sentences whose interpretation is tightly coupled with the preceding context. This content is ignored by machine translation engines as they operate at the sentence level.

This leads to problematic translations such as in the example below, extracted from an English subtitle. The subtitle is made for a 1945 Danish movie but the subtitle is apparently translated from an existing French subtitle.

(1) * *And Michael? It must come back, you hear?*
(**French**): Et Michael? Il doit revenir, vous entendez?
'And Michael? He must come back, you understand?'

We observe from Example (1) that the 3^{rd} person pronoun 'il' is mistranslated into 'it', while the preceding utterance makes it clear that the pronoun refers to a person.

Other well-documented translation errors include inaccurate lexical choices, wrong word order or mismatched inflectional endings. Here are two other examples of failed translations from the same subtitle, including both wrong lexical choices and grammatical errors:

(2) * *Come, you will see well.*
(**French**): Venez, vous verrez bien.
'Come, you'll see.'

(3) * *How are you take you?*
(**French**): Comment vas-tu t'y prendre?
'How will you go about it?'

Here is yet another example of failed translation, this time in a Dutch subtitle machine-translated from English:

(4) * *Hij is gonna verkopen ons allen langs de rivier.*
(**English**): 'He's gonna sell us all down the river'

Several translation mistakes are at play in Example (4). First of all, the English expression 'sell X down the river' is translated literally. Second, the word 'gonna' is not translated at all and simply repeated in the Dutch output. Finally, Dutch word order – which is verb-final in subordinate clauses – is not respected.

Another common error when translated into prop-drop languages (Doğruöz, 2014) relates to the use of redundant subject pronouns . The example below illustrates a redundant subject pronoun in Turkish:

(5) * *Ben telefonumu aldı*
I telephone-poss.1sg-acc take-past I
Ben döndü ve bu iki gövde vardı.
turn-past and these two body.
'I took my phone, I turned and there were these two bodies.'

Example (5) illustrates two translation issues. First, the two verbs ('take' and 'turn') lack person agreement markers. In addition, the second subject pronoun is redundant since it was already used in the first sentence and does not deliver new or contrastive information.

4. Approach

The detection of machine-translated subtitles is a challenging task, as we have no direct information about the actual source subtitle (or even the source language) that was used as translation input. Furthermore, the machine-translated subtitles are spread over a wide range of languages, as illustrated in Table 1. The features of the detection model must therefore be as language-independent as possible.

The features employed in the presented approach can be divided in two groups:

- *Target-side features*, extracted from the subtitle itself.
- *Subtitle pair features*, extracted by determining the most likely source subtitle and extracting similarity features between the source and target sentences.

4.1. Target-side features

Target-side features are defined on the sole basis of the subtitle itself. One important observation is that machine-generated subtitles typically contain a slightly larger proportion of rare/unknown tokens than their human-generated counterparts. Indeed, source-side tokens that the MT engine is unable to translate will often be repeated in the target sentence, as in the following example (where the contracted word '*tryin'*' is seemingly not understood by the MT engine and left untranslated in French):

(6) * *Regarde comme il est tryin' pour prendre sa température.*
(**English**): Looks like he's tryin' to take her temperature.

In order to detect such rare or unknown tokens, we relied on statistical language models to (1) determine the

number of tokens unknown to the language model and (2) compute the log-probabilities over the bigrams extracted from a given subtitle. The language models are derived from the Google Web 1T 5-gram corpus (Brants and Franz, 2006) when available and are estimated from the OpenSubtitles corpora otherwise (excluding the subtitles used in the evaluation). The number of unknown tokens (such as "tryin' in French) and the number of bigrams with very low log-probabilities are then integrated as features to the machine learning model. To account for the fact that distinct languages will have distinct distributions for these log-probabilities (due to e.g. differences in the vocabulary size of the various language models), the thresholds are empirically determined as percentiles of these language-specific distributions.

Subtitles are also associated with meta-data such as the release year, movie genre, release type (e.g. DVD) and original language of the movie or TV episode. These variables are also included as features in the machine learning model using one-hot encodings. Finally, a small number of subtitles include explicit clues in the beginning or end of the subtitles indicating that a machine translation engine was used. The occurrence of these cues (notably the presence of the words "Google" or "auto-translated") are also integrated as target-side features.

4.2. Subtitle pair features

The comparison between the source-side and target-side sentences can also yield useful information.

Identification of source subtitle

The first step is to identify the source subtitle that may have served as input to the machine translation engine. To determine this source, we first determine a list of potential candidates, namely subtitles associated with the same movie or TV episode but written in another language.

To find the most likely source subtitle among this list of potential candidates, a good criteria is to look at the timestamps (start and end times of subtitle blocks, in milliseconds) that are used in the subtitle. Indeed, subtitles translated from other subtitles will often have identical or near-identical timestamps, as there is no reason for the user to modify these timings. More precisely, assume a subtitle s_t written in language $l(s_t)$ and associated with the movie or TV episode with IMDB identifier $I(s_t)$. We wish to identify the source subtitle s_s from the same IMDB $I(s_s) = I(s_t)$ but written in language $l(s_s) \neq l(s_t)$ and that stands closest to s_t in terms of timestamps associated with each subtitle block. One way to measure this proximity is to extract the set of all timestamps $T(s_s)$ for subtitle s_s and the set of all timestamps $T(s_t)$ for subtitle s_t, and compute the Jaccard coefficient between the two sets:

$$J\left(T(s_s), T(s_t)\right) = \frac{|T(s_s) \cap T(s_t)|}{|T(s_s) \cup T(s_t)|} \tag{7}$$

We can then rank the list of subtitle candidates s_s for a given target subtitle s_t according to this Jaccard coefficient. To limit the number of candidates to consider, we constrain the possible source languages $l(s_s)$ to be either:

- A large "pivot language", such as English, Spanish, Russian, French, or Arabic ;
- The original language of the movie or TV episode.

The vast majority of machine-translated subtitles are indeed translations from these restricted set (mostly due to the wider availability of subtitles in these languages).

Surface-level features

Once the most likely source subtitle is determined, one can align the sentences from the two subtitles using the time-based method described in (Tiedemann, 2008) and extract features from the aligned sentence pairs.

One simple set of features is defined by the ratio between the number of tokens (and characters) in the source and target sentences. Indeed, machine-generated subtitles will often consist in literal translations of the source-side sentences, and will typically have have an average ratio close to one. On the other hand, subtitles created by human users will often show more variation in their transcription of the original dialogues, with some parts being left out, rephrased or selectively presented. This higher degree of variation will in turn lead to larger differences in the ratios of tokens (and ratios of characters) between the source and target sentences. These length ratios are, however, language-specific, as the average number of tokens may vary from language to language (as modelled in machine translation through word penalties). These differences are taken into account by rescaling the ratios by language.

Syntactic features

We can observe empirically that machine-translated subtitles are also more likely to follow the syntactic structure of the source subtitle than their human-generated counterparts. This is again due to the fact that machine-translated subtitles have more literal alignments than subtitles created by human users.

To capture this similarity, we extract the sequence of POS tags and dependency relations of the source and target subtitles through UDPipe (Straka and Straková, 2017) and extract k-gram precision scores from them:

$$\text{precision}_k = \frac{|k\text{-grams in both source and target}|}{|k\text{-grams in source}|} \tag{8}$$

The precision scores for each pair of (source,target) subtitles are then employed as features.

5. Evaluation

The features described in the previous section can be used to learn a classifier that detects whether a given a subtitle is likely to be machine-translated.

5.1. Experimental design

The dataset used for the experiments consists of a sample of 54 999 subtitles from the OpenSubtitles corpus, divided in two classes. The first class consists of the 4 999 subtitles explicitly marked as machine-generated in their meta-data (see Table 1). The second class comprises 50 000 subtitles that are (presumed to be) human-generated. As there is no absolute guarantee that a subtitle is not machine-generated,

Model	**Hyper-parameters**	**Precision**	**Recall**	F_1 **score**	**Accuracy**
Keyword baseline	"Google" at start/end of subtitle	1.000	0.017	0.030	0.910
Jaccard baseline	Jaccard coefficient ≥ 0.99	0.360	0.248	0.294	0.841
Logistic regression	Regularisation = l_2, C =1	0.266	0.757	0.394	0.787
	Regularisation = l_2, C =10	0.267	0.758	0.395	0.787
	Regularisation = l_1, C =1	0.263	0.756	0.390	0.784
	Regularisation = l_1, C =10	0.262	0.756	0.389	0.783
Support Vector Machines	Kernel = linear, C =1	0.268	0.751	0.395	0.790
	Kernel = linear, C =10	0.244	0.750	0.356	0.744
	Kernel = RBF, C =1	0.372	0.803	0.508	0.858
	Kernel = polynomial, C =1	0.340	0.708	0.460	0.848
K-nearest neighbours	Nb. neighbours = 1	0.610	0.514	0.558	0.925
	Nb. neighbours = 5	0.436	0.684	0.532	0.890
	Nb. neighbours = 10	0.359	0.757	0.486	0.854
Decision tree	Min. samples per leaf = 1	0.436	0.431	0.434	0.897
	Min. samples per leaf = 2	0.428	0.453	0.440	0.895
	Min. samples per leaf = 5	0.399	0.521	0.452	0.884
Random Forest	Nb. estimators = 10	0.718	0.409	0.521	0.931
	Nb. estimators = 50	0.758	0.449	0.564	0.937
	Nb. estimators = 100	0.772	0.448	0.567	**0.937**
Gradient Boosting	Nb. estimators = 10	0.710	0.412	0.521	0.931
	Nb. estimators = 50	0.753	0.449	0.563	0.936
	Nb. estimators = 100	0.762	0.444	0.561	0.936
Neural network (MLP)	1 hidden layer with dim. 10	0.377	0.808	0.513	0.860
	1 hidden layer with dim. 50	0.506	0.697	0.585	0.909
	1 hidden layer with dim. 100	0.580	0.661	0.617	0.925
	1 hidden layer with dim. 200	0.622	0.657	**0.638**	0.932
	2 hidden layers with dim. (10, 10)	0.374	0.812	0.512	0.858
	2 hidden layers with dim. (50, 10)	0.504	0.685	0.580	0.909

Table 2: Experimental results for the task of detecting machine-generated subtitles in a dataset of 54 999 subtitles, of which 9 % are explicitly marked as machine-generated.

we selected the subtitles that had the highest average user ratings (as users are more likely to give a high user rating to a high-quality, human-translated subtitle than a machine-generated one). The 50 000 subtitles were sampled according to the same language distribution than the 4 999 subtitles to avoid statistical biases between the two classes.

All features were scaled by removing the mean and scaling to unit variance. In addition, we found that transforming the feature values to follow a uniform distribution using quantiles information ("quantile transform") improved the performance of most classifiers. Features whose values may depend on language-specific properties (such as the average number of tokens per sentence) were scaled on a language by language basis. Class reweighting was used to compensate for the class imbalance in the dataset.

The performance of these classifiers is evaluated through 10-fold stratified cross-validation on the dataset of 54 999 subtitles, with the precision, recall, F_1-score and accuracy as performance metrics.

5.2. Models

Two simple, rule-based baselines are employed:

1. The first baseline looks at the occurrence of the token "Google" in the first and last sentences of the subtitle (which are typically indicative of a machine-translation, such as in "Tradução by Google"). This baseline has perfect precision, but only covers a small fraction of the machine-translated subtitles.
2. The other baseline looks at whether the Jaccard coefficient from Equation (7) is ≥ 0.99, indicating that the timestamps are identical or near-identical to another subtitle for the same movie or TV episode. This baseline has a higher recall but a lower precision, as many subtitles will share the same timings without being translations from one another (this is notably the case for subtitles extracted from DVD releases).

The following machine-learning models were estimated based on the features in Section 4 :

1. Logistic regression (with L_1 or L_2 regularisation)
2. SVMs (with linear, RBF or polynomial kernels)
3. K-nearest neighbours
4. Decision trees (with Gini as split criterion)
5. Random forests and gradient boosting trees
6. Feed-forward neural networks with one or two hidden layers. The networks use rectified linear units as activation layer and Adam as optimisation algorithm.

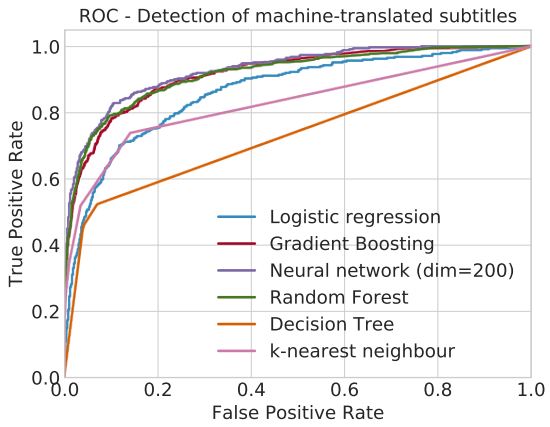

Figure 1: ROC curve for 6 machine learning models on the task of detecting machine-translated subtitles, based on the dataset of 54 999 subtitles (of which 10 % are known to be machine translated).

5.3. Results and error analysis

The results are shown in Table 2. The best performing models are feed-forward neural networks with one hidden layer, with a F_1 score of 0.638. Random forests achieve a slightly higher accuracy on this dataset, but accuracy is a less relevant metric than F_1 given the class imbalance of this task. The performance gain of neural networks seems to indicate the existence of non-linear interactions between the features that cannot be accounted for by "shallow" models such as logistic regression. All feature families described in Section 4 seem to be useful for the task (based on a small-scale feature ablation study). The most discriminative features for the task are the Jaccard coefficient, the occurrence of the "Google" keyword, and the number of unknown tokens according to the language model.

Figure 1 shows the ROC (Receiver Operating Characteristics) curve for each family of machine-learning models with the exception of SVMs which do not directly provide probabilistic estimates. The curve plots the true positive rate (equivalent to the recall) against the false positive rate when the discrimination threshold is varied.

The results demonstrate nevertheless the difficulty of the task. We conducted an error analysis of the classification results, and found most errors to be imputable to two factors. The first factor is that the "machine-translated" flags associated with the 4 999 subtitles are not always accurate. We observed a number of subtitles that were flagged as machine-translated that were surprisingly well written and lacked any obvious translation errors. In other words, their inclusion in the set of machine-translated subtitles is most likely due to a human classification error. Unfortunately, a manual cleanup of this dataset would require finding annotators capable of assessing the fluency of subtitles in most of the languages listed in Table 1, which would constitute a major undertaking.

Furthermore, the set of 50 000 subtitles assumed to be human-generated also has some shortcomings. One important problem, described in (Lison and Tiedemann, 2016) stems from the fact that many subtitles are extracted from video streams through Optical Character Recognition (OCR) and include therefore optical recognition errors, such as the letter 'i' being mistaken for the letter 'l'. These spelling errors are a source of confusion for the language model used to identify unknown tokens and determine bigram log-probabilities. We also observed subtitles including a mixture of machine-generated and human-edited sentences, often combined with numerous spelling and grammatical errors. This leads to a relatively large number of false positives. It should nevertheless be pointed out that these false positives also reflect subtitles of low-quality that one might wish to prune out of the corpus .

6. Discussion

6.1. Estimates on full corpus

The detection models presented in Section 5 can be employed to extrapolate the total number of machine-translated subtitles – or at least on the number of subtitles of suspiciously low quality – in the full corpus. We selected a random sample of 30 000 subtitles from the OpenSubtitles corpus (excluding the subtitles used in the evaluation). We then extracted the features from Section 4 and applied the most accurate detection model (the feedforward neural network with one hidden layer of 200 dimensions) on these features. As the output probabilities of the neural network are not calibrated, we perform probability calibration using Platt's sigmoid model (Guo et al., 2017).

The resulting distribution of probabilities (using Kernel Density Estimation) is illustrated in Figure 2. We can observe from the figure that most of the probability mass lies within the lower half of the distribution, but a small proportion of subtitles has a high probability of being machine-translated according to the detection model.

Base on this empirical distribution, we can proceed to estimate the number of machine-translated subtitles on the full OpenSubtitles corpus through a Poisson Binomial Distribution, which corresponds to the sum of independent Bernoulli trials that are not identically distributed (in this case, the probabilities of being machine-translated). The mean of this distribution is set to 327 K (out of 3.735 million subtitles) with a standard deviation $\sigma = 335.8$. In other words, the proportion of machine-translated subtitles (and other subtitles of similarly low quality) amounts to about 9 % of the total corpus.

6.2. Corpus filtering

The detection models presented in Section 5 can be used to detect at least a substantial portion of the machine-translated subtitles in the OpenSubtitles corpus. As illustrated by the ROC curve in Figure 1, the neural model is notably able to detect 51 % of the machine-subtitles with a false positive rate of just 1 %. Given the sensitivity of the model to the number of unknown tokens and the bigram log-probabilities, the detected subtitles are presumably also the ones with the lowest quality in terms of linguistic flu-

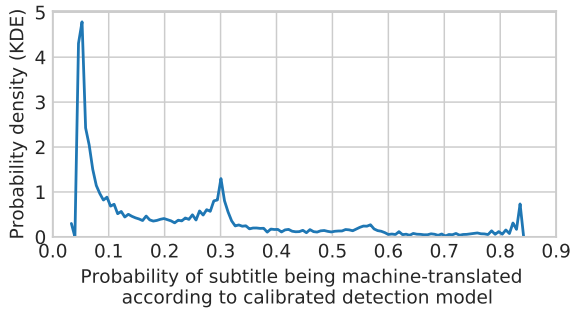

Figure 2: Distribution of probability values given by the calibrated neural network on the set of 30 000 subtitles of unknown class. Kernel Density Estimation is employed for the probability density function.

ency (and thus the ones causing the most important degradation to the quality of the resulting corpus).

The predictions from the detection model can be exploited in several ways. The most straightforward is to directly filter out these (presumed) machine-translated subtitles from the corpus. Alternatively, one can integrate the outputs of the prediction as a distinct feature in the statistical rescoring model of (Lison et al., 2018), which associates each sentence alignment with a numerical score. The latter approach has the advantage of allowing for various filtering levels, from conservative (keeping all subtitles in the corpus) to aggressive (removing all suspicious subtitles), without committing to a specific threshold. Finally, one can also transform the prediction into weights associated with each subtitle. Such weights can be used in various downstream applications, for instance when training machine translation models (Matsoukas et al., 2009)

Although the evaluation presented in this paper focused on subtitles, it should be pointed out that most features employed in the detection models (with the exception of metadata features) are genre-independent and can be extracted on other types of parallel or comparable corpora.

7. Conclusion

Parallel corpus extracted from movie and TV subtitles can be particularly noisy and include a large number of low-quality subtitles. One important cause of this low-quality is the presence of subtitles translated from other subtitles through online machine translation tools. Detecting and pruning out (or downsampling) these subtitles is therefore expected to enhance the overall quality of such corpora.

The present paper described a data-driven approach to the detection of machine-translated documents based on a combination of linguistic and extra-linguistic features. Experimental results show that a detection model based on a feed-forward neural network with one hidden layer is able to achieve reasonable performance on this task. In contrast with previous work, the machine learning models are not optimised for a specific language pair or translation model and can be directly applied to any multilingual corpus. The detection model can be used to filter out machine-translated documents from the corpus or assign them to a lower weight in downstream applications.

Future work will investigate how to further improve our understanding of the relations between subtitles and uncover the "history" behind each subtitle. Subtitles are indeed connected to each other in a myriad of ways:

- Some subtitles are translations of subtitles in other languages, as addressed in this paper. These translations may be done by (professional or amateur) human translators, machine translation tools, or a combination of both (machine-assisted translation).
- A second group consist of subtitles that are part of the same release (for instance, subtitles included in the same DVD). Such subtitles are often created by the same translation/subtitling company and are therefore relatively close at a structural level, although they are typically not translations of one another.
- Finally, many subtitles are corrections of previous subtitles in the same language (for instance to correct spelling, grammatical or formatting errors).

The relations above are important for the construction of parallel corpora from subtitles, as they provide key insights on the relative quality and proximity of each pair of subtitle. In the longer term, we wish to integrate these inferred relations into the ranking model employed for the document-level alignment process (Lison and Tiedemann, 2016).

References

Aharoni, R., Koppel, M., and Goldberg, Y. (2014). Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 289–295. ACL.

Akbik, A., Guan, X., and Li, Y. (2016). Multilingual aliasing for auto-generating proposition banks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 3466–3474.

Arase, Y. and Zhou, M. (2013). Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1597–1607. The Association for Computer Linguistics.

Belinkov, Y. and Glass, J. (2016). Large-scale machine translation between Arabic and Hebrew: Available corpora and initial results. *arXiv preprint arXiv:1609.07701*.

Brants, T. and Franz, A. (2006). Web 1T 5-gram corpus version 1. Technical report, Google Research.

Doğruöz, A. S. (2014). On the borrowability of subject pronoun constructions in Turkish-Dutch contact. *Constructions and Frames*, 6(2):143–169.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting*

of the Association for Computational Linguistics (ACL 2011), pages 1318–1326. ACL.

Krause, B., Damonte, M., Dobre, M., Duma, D., Fainberg, J., Fancellu, F., Kahembwe, E., Cheng, J., and Webber, B. L. (2017). Edina: Building an open domain socialbot with self-dialogues. *CoRR*, abs/1709.09816.

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

Lembersky, G., Ordan, N., and Wintner, S. (2012). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Lison, P. and Bibauw, S. (2017). Not all dialogues are created equal: Instance weighting for neural conversational models. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2017)*, pages 384–394, Saarbrücken, Germany. ACL.

Lison, P. and Kutuzov, A. (2017). Redefining context windows for word embedding models: An experimental study. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (Nodalida 2017)*, pages 284–288, Göteborg, Sweden. Linköping University Electronic Press.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Opensubtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan. (accepted).

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 708–717. Association for Computational Linguistics.

Petukhova, V., Agerri, R., Fishel, M., Penkale, S., del Pozo, A., Maucec, M. S., Way, A., Georgakopoulou, P., and Volk, M. (2012). SUMAT: Data collection and parallel corpus compilation for machine translation of subtitles. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Pilevar, M. T., Faili, H., and Pilevar, A. H. (2011). Tep: Tehran English-Persian parallel corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 68–79. Springer.

Pryzant, R., Chung, Y., Jurafsky, D., and Britz, D. (2017). JESC: japanese-english subtitle corpus. *CoRR*, abs/1710.10639.

Speer, R. and Lowry-Duda, J. (2017). ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89. Association for Computational Linguistics.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. ACL.

Stymne, S. and Ahrenberg, L. (2012). On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Tiedemann, J. (2007). Improved sentence alignment for movie subtitles. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07)*, Borovets, Bulgaria.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 1902–1906, Marrakesh, Marocco.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Benjamins translation library. John Benjamins Publishing Company.

van der Wees, M., Bisazza, A., and Monz, C. (2016). Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 2571–2581.

Vilar, D., Xu, J., D'haro, L., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

Wang, L., Tu, Z., Zhang, X., Liu, S., Li, H., Way, A., and Liu, Q. (2017). A novel and robust approach for pro-drop language translation. *Machine Translation*, pages 1–23.

Zhang, S., Ling, W., and Dyer, C. (2014). Dual subtitles as parallel corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Quasi-Parallel Corpora: Hallucinating Translations for the Chinese–Japanese Language Pair

Yves Lepage
Waseda University
808-0135 Fukuoka-ken, Kitakyûsyû-si, Wakamatu-ku, Hibikino 2-7, Japan
yves.lepage@waseda.jp

Abstract

We show how to address the problem of bilingual data scarcity in machine translation. We propose a method that generates aligned sentences which may be not perfect translations. It consists in 'hallucinating' new sentences which contain small but well-attested variations extracted from unaligned unrelated monolingual data. We conducted various experiments in statistical machine translation between Chinese and Japanese to determine when adding such quasi-parallel data to a basic training corpus leads to increases in translation accuracy as measured by BLEU.

Keywords: Machine translation, Quasi-parallel data, Comparable corpora

1. Introduction

Some languages are well-resourced. This means that tools like segmenters, morphological analysers, syntactic or semantic parsers are available for them. It also means that large amounts of monolingual data are available, usually freely available. Some language pairs are also well-resourced. This means that large amounts of parallel, i.e., well-aligned, data exist for the two languages. Indeed a large number of language pairs are not well-resourced, so that directly building translation systems for these languages is problematic. In this respect, one-shot translation (Johnson et al., 2016) in the framework of neural machine translations raises great expectations. Nevertheless, it is still acknowledged that the lack of aligned or parallel data remains a problem for MT in general.

2. Lack of Parallel Data for Chinese–Japanese

2.1. The Situation

Individually, Chinese and Japanese are relatively well-resourced languages with efficient segmenters, morphological analysers, parsers, etc. However, the language pair itself suffers from a lack of freely available bilingual corpora and this is a problem for machine translation between these two languages.

The BTEC corpus (Takezawa et al., 2002) contains short sentences in the tourism domain, but this corpus is not available for free[1]. The original version contains 160,000 sentences, but it has been extended to more than 1 million.

There also exist one large corpus in the scientific and technological domain, used in the MT evaluation campaign WAT, the ASPEC-JC corpus (Nakazawa et al., 2016). Its use requires to sign a license agreement, to participate in the WAT campaign, and to erase data after a one-year term.

2.2. Possible Answers

Different possible solutions to augment the size of parallel corpora have been proposed in the past. They range from the manual creation of data to the automatic extraction of comparable corpora, with attempts at creating bilingual data from monolingual data (Klementiev et al., 2012; Sun et al., 2013; Chu et al., 2013). In statistical machine translation, where the translation table is crucial, directly augmenting the data in the translation table has also been proposed (Luo et al., 2013). All these methods may solve the problem of data scarcity to some extent and lead to increases in BLEU points in different language pairs when used in addition to existing training data.

2.3. The Proposed Answer

The purpose of this paper is to describe a method to create a corpus of aligned sentences, which are translations of one another only up to a certain extent. Because the translation correspondence may not be perfect, we call such a bilingual corpus a quasi-parallel corpus. The similarities and differences between a quasi-parallel corpus and a comparable corpus can be summarised as follows:

Comparable corpus	Quasi-parallel corpus
not exact translations	not exact translations
natural texts	synthetic data
unit: document	unit: sentences
not sentence-aligned	sentence-aligned by design
usually one topic / doc.	any topic

The method consists in 'hallucinating' linguistic data (Irvine and Callison-Burch, 2014), i.e., in creating hopefully parallel, synthetic data from unrelated unaligned monolingual data. However, a certain amount of parallel data as seed data is necessary.

In previous works, we assessed different sets of such 'hallucinated' data by adding them to a training corpus to build an SMT system. This led to variable improvements, as measured by BLEU, ranging from less than half a point on difficult tasks, to several points in other tasks (Wang et al., 2014a), depending on the experimental conditions.

3. Generation of Quasi-Parallel Corpora

3.1. Collecting Variations in Monolingual Data

Figure 1 gives an illustrated overview of the proposed method. The central object in the method is a list of analog-

[1] Approximate cost as of February 2018: 1 yen per sentence.

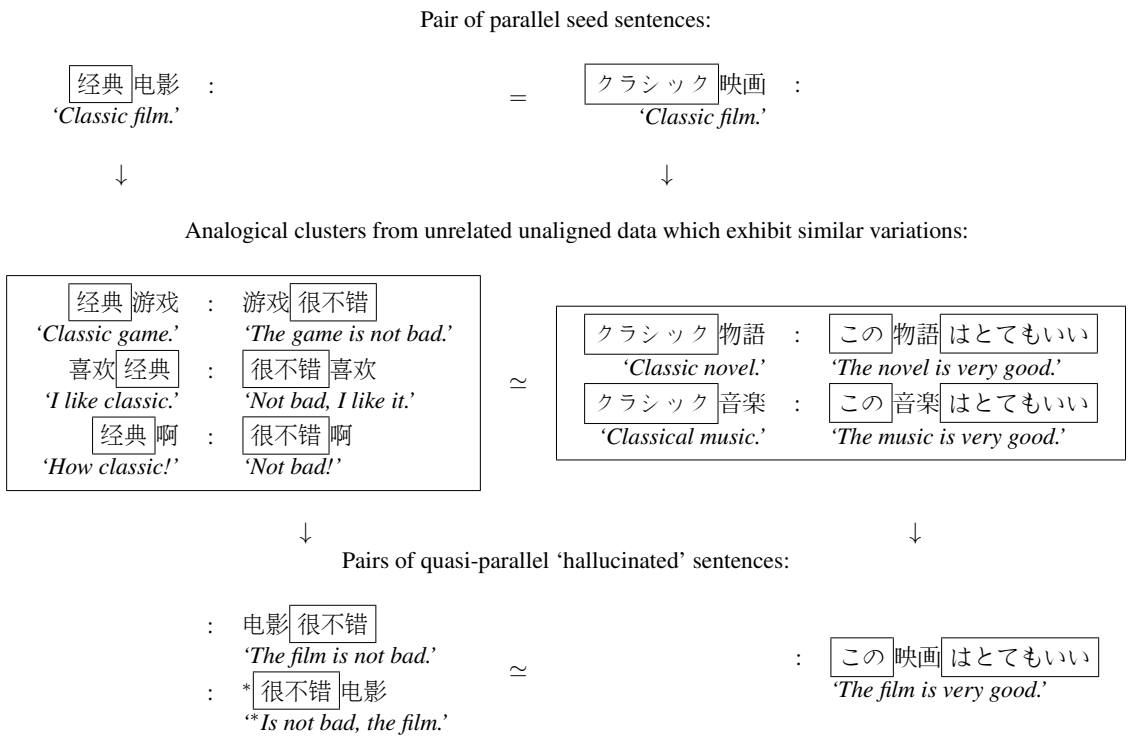

Figure 1: Overview of the generation of hallucinated quasi-parallel translations from parallel seed sentences using analogical clusters produced from unrelated unaligned monolingual data. Chinese on the left, Japanese on the right. The clusters exhibit similar variations so that the sentences obtained from aligned seed sentences can be thought to be almost translations of one another. The variations exhibited by the clusters are framed. The Japanese part shows that the variations may be discontinuous. Notice that the sentences in the analogical clusters are not translations and that the number of sentences in each cluster is different in each language. Notice also that the second hallucinated Chinese sentence is ungrammatical.

ical clusters which exhibit similar variations. In this example, the variation in both languages can be represented as:

经典X:X很不错 ≃ クラシックX:このXはとてもいい
'Classic X.' 'X is not bad.' 'Classic X.' 'This X is very good.'

Basically, this is an illustration of the principle of translation by analogy introduced in (Nagao, 1984). Finding such configurations requires to perform two tasks: firstly, to collect a relatively large number of small variations in each language, which are well-attested; secondly to be able to show that some of the well-attested monolingual variations in one language correspond to some well-attested variations in the other language.

To complete the first task, we deal with the idea of well-attested series of variations (Yang et al., 2013a; Yang et al., 2013b). Such series of variations, extracted from actual monolingual data, are shown in Figures 2 and 3 for Chinese and Japanese respectively. They are instances of what is called analogical clusters. For details concerning the definition of analogical clusters and the fundamental relation this definition relies on, i.e., proportional analogy, see Appendix 7.

As for implementation, (Fam and Lepage, 2018) describes a set of publicly released tools to automatically output analogical clusters from textual data. The example clusters in Figures 2 and 3 have been obtained using these tools.

3.2. Similarity of Variations Across Languages

In order for the method illustrated in Figure 1 to work, it is necessary to complete the second task mentioned in the previous section, i.e., to be able to show that some of the well-attested monolingual variations in one language correspond to some other well-attested variations in the other language. For that, we use classical ways of comparing bags-of-words across languages.

The computation is performed on the variations exhibited in a cluster. Hence, we compute the differences between the left and the right sides of each cluster in each language and compare these differences by use of Dice coefficients. In order to normalise words across languages, in the case of Chinese and Japanese, we make use of hanzi-kanji conversion tables and dictionaries. The use of translation tables is of course possible. See Appendix 8. for formulae used in estimating the similarity between analogical clusters across two languages.

As shown in the appendix, a reasonably high value of 0.833 is obtained for the two clusters shown in Figure 1.

不值得购买 : 很值得购买
'It's not worth buying.' : *'It's worth buying.'*
这个游戏不好玩 : 这个游戏很好玩
'The game is not funny.' : *'The game is very funny.'*
画面不好 : 画面很好
'The frame is bad.' : *'The frame is very good.'*
小朋友不喜欢 : 小朋友很喜欢
'Children don't like it.' : *'Children like it very much.'*
难度不大 : 难度很大
'It's not difficult.' : *'It's very difficult.'*
⋮ : ⋮

太好了 : 效果太好了
'It's very good.' : *'The effect is very good.'*
非常不错 : 效果非常不错
'It's not bad.' : *'The effect is not bad.'*
画面很好 : 画面效果很好
'The frame is very good.' : *'Effect of the frame is very good.'*
很炫 : 效果很炫
'It's very cool.' : *'The effect is very cool.'*
马马虎虎 : 效果马马虎虎
'It's just so-so.' : *'The effect is just so-so.'*
⋮ : ⋮

Figure 2: Two analogical clusters in Chinese. The first one *(top)* illustrates the opposition between negative and affirmative sentences (不 *'not'* replaced by copula 很 *'is'* (etymologically adverb *'very'*)). The second one *(bottom)* illustrates the replacement of unexpressed subjects (expressed in English by the pronoun *'it'*) by the noun 效果 *'effect'*. The framed sentence shows that the same sentence may be found in different analogical clusters.

3.3. Generating Hallucinated Synthetic Data by Application of the Variations

As Figure 1 illustrates, it is possible to apply the variations exhibited in an analogical cluster to any sentence for which it makes sense. The very application of the variations on a sentence is performed by solving equations. E.g., for Figure 1, the equation

$$\boxed{\text{经典}}\text{游戏} : \text{游戏}\boxed{\text{很不错}} :: \boxed{\text{经典}}\text{电影} : x$$

is formed by taking the first line in the Chinese cluster and the Chinese sentence in the pair of aligned sentences at the top of the figure. The solution of this equation is the first Chinese 'hallucinated' sentence: $x =$ 电影很不错.
As all the lines in a cluster are used in turn, it is understandable that the same hallucinated sentence may be generated several times.

3.4. Filtering Out Ill-Formed Sentences

However, as mentioned in the caption of Figure 1 and as is well known with analogy, there is a danger of overgeneration, i.e., a risk of creating sentences which are ill-formed, either because they make no sense (ill combinations of characters) or because they are ungrammatical.

早急に対応して下さい。 : 早急に対応して欲しい。
'Please respond as soon as possible.' : *'I want you to respond as soon as possible.'*
正式版に戻して下さい。 : 正式版に戻して欲しい。
'Please return to the official version.' : *'I want you to return to the official version.'*
元に戻して下さい。 : 元に戻して欲しい。
'Please return to the beginning.' : *'I want you to return to the beginning.'*
やめて下さい。 : やめて欲しい。
'Please stop.' : *'I want you to stop.'*
⋮ : ⋮

本当に良かった : 良かったですね
'It was really good.' : *'It was good.'*
本当に酷い : 酷いですね
'It was so cool.' : *'It was cool.'*
本当に嬉しい : 嬉しいですね
'I am really very happy.' : *'I am very happy.'*
⋮ : ⋮

Figure 3: Two analogical clusters in Japanese. The first one *(top)* illustrates the opposition between a request or a wish expressed by 下さい *'Please'* and 欲しい *'I want'* respectively at the end of the sentence. The second one *(bottom)* illustrates the opposition between informal speech on the left and standard speech on the right (suffixation by a copula です and a sentence marker ね). In addition, the sentences on the left include 本当に *'in fact, really, in reality'* at their beginning.

This is the case with the second Chinese hallucinated sentence in Figure 1.
To remedy this problem, based on extensive experiments and comparison of different methods (SVM, language models), we rely on counts of N-sequences to check for naturalness (Doddington, 2002; Lin and Hovy, 2003). The results of our experiments suggest to take a rigid stance and to reject any sentence which contains an N-sequence not attested in a given reference dataset. In other terms, for a sentence to be retained, all of its N-sequences should be attested in the reference dataset (N = 6 for Chinese and 7 for Japanese in our experiments). The method favours precision to the detriment of recall. Indeed manual inspection suggests that a very large amount of valid sentences are actually discarded. However, in experiments where we assessed the quality of the retained sentences, it was judged that 99 % of the sentences are correct in Chinese and Japanese. As for reference dataset, the monolingual data used to collect analogical clusters or the training data to be used in an MT experiment can be used.

4. Assessment with Statistical Machine Translation

In various experiments in SMT conducted over several years in different settings (Wang et al., 2014a; Wang et al., 2014b; Yang et al., 2014; Yang and Lepage, 2014b; Yang and Lepage, 2014a; Yang et al., 2015; Yang et al., 2017), it was shown that the introduction of the small variations

	Training data (# of lines)	Additional data (# of lines)	(percent)	Baseline (rounded BLEU points)	Increase
PolTAL 2014 / IPSJ 2017	subtitles$_1$	seeds = subtitles$_1$ clusters = Web news$_1$			
	110,000	75,000	(+68%)	11 $\sim$ 13	+6
PIC 2014	subtitles$_2$	seeds = 10% of subtitles$_2$ clusters = Web news$_2$			
	120,000	10,000	(+8.3%)	17 $\sim$ 20	+2 $\sim$ 4
WAT 2014	ASPEC corpus	seeds = 1/6 of ASPEC (length $\leq$ 30 chars) clusters = Web news$_1$		(Moses 1.0)	
	670,000	35,000	(+5%)	23 $\sim$ 29	+2 $\sim$ 3
Additional exp. 2017	ASPEC corpus	seeds = 1/6 of ASPEC (length $\leq$ 30 chars) clusters = ASPEC		(Moses 2.1)	
	670,000	2,800	(+0.3%)	30 $\sim$ 37	+0

Table 1: Synthesis in numbers of several experiments in using quasi-parallel corpora for SMT. Subtitles$_1$ and Subtitles$_2$ are different excerpts from the OpenSubtitles corpus (Tiedemann, 2009). Web news$_1$ and Web news$_2$ are two in-house datasets browsed from various news sites in Chinese and Japanese. Larger improvements are obtained when the training corpus and the quasi-parallel corpus are from different domains and when the quasi-parallel corpus is large relatively to the training corpus (compare framed values on first and last lines).

created by the proposed method of adding a quasi-parallel corpus to the training data explained above, increases the size of the translation tables and that the new phrases are actually used and may contribute to translation accuracy. A synthesis of the results obtained over the years is given in Table 1.

The overall results are mitigated. The improvements in translation accuracy as measured by BLEU vary from large positive values to smaller and less encouraging values. Also, in experiments reduplicated with different versions of the Moses engine, versions 1.0 and 2.1, it was observed that the upgrade of the Moses engine made up for the increases brought by the method on the older version.

Notwithstanding the various improvements in BLEU scores, two main lessons can be drawn from the SMT experiments.

Firstly, several experiments tend to show that the quality of the alignment of the produced sentences is not so crucial. What seems to be crucial is the grammaticality of the sentences produced. For that, different configurations and various methods have been tested so as to automatically ensure a very high level of grammaticality or semantic correctness. The N-sequence filtering method was found to be the most effective technique to filter out ill-formed sentences, despite a very low recall.

Secondly, the relationship or rather the absence of relation between the basic training data and the monolingual data seems to be important. Monolingual data from the same domain or the same collection of texts do not seem to conduct to significant improvements. Thorough experiments still need to be conducted to confirm this impression, but it seems that variations from the general language, are necessary to bring improvements in translation accuracy. Relatively to this, the larger the quasi-parallel corpus added to the training corpus, the better.

5. Conclusions

The method described in this paper to produce a quasi-parallel corpus relies on the application of a large number of small well-attested variations on a relatively small number of parallel seed sentences. As SMT is concerned, these small variations are captured in the translation table and, if such small variations appear in the test set, the test set may be better translated. This is shown by the fact that a larger number of the phrases generated from the quasi-parallel corpus are indeed used to translate the test set, in comparison to a baseline system trained without the quasi-parallel corpus.

What seems important for the method to work is the grammatical quality of the generated sentences, while, relatively, the quality of the correspondence between the clusters may not be so important. It seems that the best configuration is a configuration where the monolingual data for the extraction of analogical clusters is varied enough so as to offer useful variations and where these monolingual data are different from those found in the training data, i.e., new variations can be found. Consequently, the positive effect of the quasi-parallel corpus may be thought as the effect of providing variations found in the general usage of the languages to be translated.

6. Acknowledgements and Thanks

A large part of the work reported here was supported by a JSPS Grant, Number 15K00317 (Kakenhi C), entitled "Language productivity: efficient extraction of productive analogical clusters and their evaluation using statistical machine translation."

Special thanks to YANG Wei (former PhD student) who, as a research assistant, was instrumental and contributed largely to the implementation of the method and to the obtention of the results described and reported here.

Thanks to WANG Hao (PhD student), SHEN Hanfei and GAO Mengru (former master students) for collecting data, writing programs and running experiments.

7. Appendix: Definition of Analogical Clusters

An analogical cluster is defined in the following way, where the s's stand for sentences, i.e., strings of characters (computation in strings of words is also possible):

$$\begin{array}{c} s_1^1 : s_1^2 \\ s_2^1 : s_2^2 \\ \vdots \ \ \vdots \\ s_n^1 : s_n^2 \end{array} \quad \stackrel{\Delta}{\Longleftrightarrow} \quad \begin{array}{r} \forall (i,j) \in \{1,\ldots,n\}^2, \\ s_i^1 : s_i^2 :: s_j^1 : s_j^2 \end{array} \tag{1}$$

In this definition, it is understandable that the underlying relation between four sentences, noted by semi-colons and double semi-colons as $s_i^1 : s_i^2 :: s_j^1 : s_j^2$, is the most important notion. This notion is that of proportional analogy, for which we adopt the characterisation introduced in (Lepage, 1998; Lepage, 2003):

$$A : B :: C : D \ \Rightarrow \left\{ \begin{array}{c} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{array} \right. \tag{2}$$

where $d(A,B)$ is the distance between two strings A and B and $|A|_a$ stands for the number of occurrences of character a in string A.

In order to make Characterisation (2) operational, we read it in the other direction, i.e., we assume that an analogy holds when the constraints on distance and character counts are met.

8. Appendix: Computation of Analogical Cluster Similarity Across Two Languages

For simplicity, we compare analogical clusters across languages by first extracting the differences in words on their left and right sides and then compare two analogical clusters in two different languages by taking the mean of the Dice coefficients for the differences on each of their sides. This is expressed by Formula (3).

$$\begin{array}{l} \text{Sim}((L_{zh} : R_{zh}), (L_{ja} : R_{ja})) = \\ \qquad \frac{1}{2}(\text{Dice}(L_{zh}, L_{ja}) + \text{Dice}(R_{zh}, R_{ja})) \end{array} \tag{3}$$

We repeat the formula for the Dice coefficient ($|S|$ stands for the cardinality of a set S):

$$\text{Dice}(S_{zh}, S_{ja}) = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \tag{4}$$

To be able to compute the intersection between two sets of words in two different languages, Chinese and Japanese, we normalise the words in one language in the other language by making use of kanji-hanzi conversion, dictionaries, translation tables, etc. Nowadays we should consider bilingual word vector representations.

As an illustration, for the clusters in Figure 1, knowing from some dictionary or translation table that 经典 = クラシック, 很 = とても and 不错 = いい, we perform the following computation:

$$\begin{array}{rl} & \text{Sim}((L_{zh} : R_{zh}), (L_{ja} : R_{ja})) \\ = & \frac{1}{2}\left(\frac{2 \times |\{\text{经典} = \text{クラシック}\}|}{|\{\text{经典}\}| + |\{\text{クラシック}\}|}\right. \\ & \left. \quad + \frac{2 \times |\{\text{很} = \text{とても}, \text{不错} = \text{いい}\}|}{|\{\text{很}, \text{不错}\}| + |\{\text{この}, \text{は}, \text{とても}, \text{いい}\}|}\right) \\ = & \frac{1}{2}\left(\frac{2 \times 1}{1+1} + \frac{2 \times 2}{2+4}\right) \\ = & \frac{1}{2}\left(1 + \frac{2}{3}\right) \\ = & 0.833 \end{array}$$

because the left and right parts of the variations in each of the Chinese and Japanese clusters are

$$(L_{zh} : R_{zh}) = (\{\text{经典}\} : \{\text{很}, \text{不错}\})$$

and

$$\begin{array}{l} (L_{ja} : R_{ja})) = \\ \qquad (\{\text{クラシック}\} : \{\text{この}, \text{は}, \text{とても}, \text{いい}\}) \end{array}$$

respectively.

As the values range from 0 to 1, with higher values showing greater similarity, a value of 0.833 can be interpreted as a high similarity for the variations exhibited by the two clusters.

9. Bibliographical References

Chu, C., Nakazawa, T., and Kurohashi, S. (2013). Chinese–Japanese parallel sentence extraction from quasi–comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 34–42, Aug.

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, March. Morgan Kaufmann Publishers Inc.

Fam, R. and Lepage, Y. (2018). Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, May.

Irvine, A. and Callison-Burch, C. (2014). Hallucinating phrase translation for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170, Ann Arbor, Michigan, June. Association for computational linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. In *Proceedings of the*

13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, pages 130–140, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lepage, Y. (1998). Solving analogies on words: an algorithm. In *Proceedings of COLING-ACL'98*, volume I, pages 728–735, Montréal, August.

Lepage, Y. (2003). *De l'analogie rendant compte de la commutation en linguistique (Of that kind of analogies capturing linguistic commutation)*. Habilitation thesis, Joseph Fourier Grenoble University, May.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 150–157, Edmonton, May.

Luo, J., Max, A., and Lepage, Y. (2013). Using the productivity of language is rewarding for small data: Populating SMT phrase table by analogy. In Zygmunt Vetulani, editor, *Proceedings of the 6th Language & Technology Conference (LTC'13)*, pages 147–151, Poznań, December. Fundacja uniwersytetu im. Adama Mickiewicza.

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial & Human Intelligence*, pages 173–180.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia, may. European Language Resources Association (ELRA).

Sun, J., Qing, Y., and Lepage, Y. (2013). An iterative method to identify parallel sentences from non-parallel corpora. In Zygmunt Vetulani, editor, *Proceedings of the 6th Language & Technology Conference (LTC'13)*, pages 238–242, Poznań, December. Fundacja uniwersytetu im. Adama Mickiewicza.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). Toward a broad coverage bilingual corpus for speech translation of travel conversation in the real world. In *Proceedings of LREC 2002*, pages 147–152, Las Palmas, May.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, et al., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Wang, H., Yang, W., and Lepage, Y. (2014a). Improved Chinese-Japanese phrase-based MT quality using extended quasi-parallel corpus. In Yinglin Wang, et al., editors, *Proceedings of the 2014 IEEE International Conference on Progress in Informatics and Computing (PIC 2014)*, pages 6–10. IEEE Computer Society Press, May.

Wang, H., Yang, W., and Lepage, Y. (2014b). Sentence generation by analogy: towards the construction of a quasi-parallel corpus for Chinese-Japanese. In *Proceedings of the 20th Annual Meeting of the Japanese Association for Natural Language Processing*, pages 900–903, Sapporo, March.

Yang, W. and Lepage, Y. (2014a). Consistent improvement in translation quality of Chinese–Japanese technical texts by adding additional quasi-parallel training data. In *Proceedings of the First Workshop on Asian Translation (WAT)*, pages 69–76, October.

Yang, W. and Lepage, Y. (2014b). Inflating a training corpus for SMT by using unrelated unaligned monolingual data. In Adam Przepiórkowski et al., editors, *Advances in Natural Language Processing: Proceedings of the 9th conference on language processing (PolTAL 2014)*, volume LNAI 8686, pages 236–248, Warsaw, Poland, September. Springer.

Yang, W., Wang, H., and Lepage, Y. (2013a). Automatic acquisition of rewriting models for the generation of quasi-parallel corpus. In Zygmunt Vetulani, editor, *Proceedings of the 6th Language & Technology Conference (LTC'13)*, pages 409–413, Poznań, December. Fundacja uniwersytetu im. Adama Mickiewicza.

Yang, W., Wang, H., and Lepage, Y. (2013b). Using analogical associations to acquire Chinese-Japanese quasi-parallel sentences. In *Proceedings of the tenth symposium on natural language processing (SNLP2013)*, pages 86–93, Phuket, Thailand, October.

Yang, W., Wang, H., and Lepage, Y. (2014). Deduction of translation relations between new short sentences in Chinese and Japanese using analogical associations. *International Journal of Advanced Intelligence*, 6(1):13–34.

Yang, W., Zhao, Z., and Lepage, Y. (2015). Inflating training data for statistical machine translation using unaligned monolingual data. In *Proceedings of the 21th Annual Meeting of the Japanese Association for Natural Language Processing*, pages 1016–1019, Kyoto, March.

Yang, W., Shen, H., and Lepage, Y. (2017). Inflating a small parallel corpus into a large quasi-parallel corpus using monolingual data for Chinese–Japanese machine translation. *Journal of Information Processing*, 25:88–99.

Overview of the Third BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora

Pierre Zweigenbaum
LIMSI, CNRS,
Université Paris-Saclay,
F-91405 Orsay, France
pz@limsi.fr

Serge Sharoff
University of Leeds
Leeds, United Kingdom
s.sharoff@leeds.ac.uk

Reinhard Rapp
Magdeburg-Stendal University
of Applied Sciences and
University of Mainz, Germany
reinhardrapp@gmx.de

Abstract

This paper presents the BUCC 2018 shared task on parallel sentence extraction from comparable corpora. This task used the same data as the BUCC 2017 shared task. 17 runs were submitted by 3 teams, covering all four proposed language pairs: German-English (3 runs), French-English (6 runs), Russian-English (3 runs), and Chinese-English (5 runs). The best F-scores as measured against the gold standard were 0.86 (German-English), 0.81 (French-English and Russian-English), and 0.77 (Chinese-English). All top scores improved over those of 2017.

Keywords: Comparable corpora, parallel sentences, parallel sentence extraction, cross-language similarity, annotated corpus

1. Introduction

Comparable corpora are gaining momentum as a supplement to parallel corpora for multilingual natural language processing (Sharoff et al., 2013; Rapp et al., 2016). After the extraction of word translations (Rapp, 1995; Fung, 1995), the detection of parallel sentences (Utiyama and Isahara, 2003; Munteanu et al., 2004; Abdul Rauf and Schwenk, 2009a) and parallel segments (Munteanu and Marcu, 2006; Hewavitharana and Vogel, 2011) in comparable corpora was addressed and found to improve statistical machine translation (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009b).

This strong interest in comparable corpora created a need for shared tasks that provide common task definitions, datasets and evaluation methods to assess the state of the art. Such shared tasks were created in the context of the BUCC workshop series on Building and Using Comparable Corpora and in other venues: the first one was run at BUCC 2015 and addressed the detection of comparable documents in two languages (Sharoff et al., 2015). It was followed on the same topic by the bilingual document alignment task of WMT 2016 (Buck and Koehn, 2016). A task on parallel sentence extraction from comparable corpora was prepared in 2016 (Zweigenbaum et al., 2016) and organized at BUCC 2017 (Zweigenbaum et al., 2017). It bears relations with but differs in several respects from the cross-language plagiarism detection tasks of PAN (Potthast et al., 2012) and the cross-language semantic text similarity task of SemEval (Agirre et al., 2016).

To let more participants take part in this task, we decided to run it for a second year in 2018 as the Third BUCC Shared Task.[1]. In this paper we describe the task and its datasets (Section 2.), the participants' systems (Section 3.), the results they obtained (Section 4.), and conclude (Section 5.).

[1] https://comparable.limsi.fr/bucc2018/bucc2018-task.html

2. Task and Datasets

As in the Second BUCC Shared Task, the Third BUCC Shared Task aims to examine the ability of algorithms to detect parallel sentence pairs in a pair of monolingual corpora. Its design principles are the following.

Observing that past work took advantage of much existing meta-information, such as links between two matching Wikipedia articles in two languages or article dates in synchronous comparable news corpora (Munteanu and Marcu, 2005), we decided to create a dataset in which algorithms should focus on sentence contents instead of trying to rely on external, contextual clues. This should remove a large part of the heuristic aspects of these algorithms that are not directly linked to detecting cross-language sentence parallelism. Therefore this BUCC dataset has no meta-information attached to documents or sentences. To prevent participants from obtaining such meta-information indirectly, the instructions asked them not to use the original datasets from which the BUCC dataset was built.

The main difficulty in preparing a dataset to evaluate parallel sentence extraction from a pair of comparable corpora is the preparation of gold standard annotations: these annotations must identify the true positive parallel sentence pairs among the much larger set of true negatives, i.e., non-parallel sentence pairs, among the cross-product of sentences of the two corpora. Because the cross-product grows with the product of the sizes of the two corpora, as soon as these sizes exceed a few hundred sentences, it becomes difficult, not to say impossible, to manually spot the few parallel sentence pairs that happen to occur in these comparable corpora.

We therefore designed a dataset in which (i) parallel sentence pairs have been artificially inserted, in a way to make their presence as inconspicuous as possible; and (ii) action has been taken to make naturally occurring parallel sentence pairs less likely to occur. More detail is provided in (Zweigenbaum et al., 2016; Zweigenbaum et al., 2017).

The dataset for the BUCC'18 shared task consits of two parts. The non-parallel part is made of Wikipedia sen-

Pair	Sample (2%)			Training (49%)			Test (49%)		
	fr	en	gold	*fr*	en	gold	*fr*	en	gold
de-en	32593	40354	1038	413869	399337	9580	413884	396534	9550
fr-en	21497	38069	929	271874	369810	9086	276833	373459	9043
ru-en	45459	72766	2374	460853	558401	14435	457327	566356	14330
zh-en	8624	13589	257	94637	88860	1899	91824	90037	1896

Table 1: Corpus statistics (reproduced from (Zweigenbaum et al., 2017)): number of monolingual sentences (*fr*, en) and of parallel pairs (gold) for each split and each language pair. The *fr* column stands for the non-English language in each pair.

Name	Affiliation (reference)	Language pairs (*-en)
H2@BUCC2018	Carnegie Mellon University in Qatar, Qatar & QCRI, Qatar (Bouamor and Sajjad, 2018)	fr (3)
NLP2CT	NLP2CT Lab, Dept. of Computer and Information Science, University of Macau (Leong et al., 2018)	zh (2)
VIC	Vicomtech-IK4, Donostia / San Sebastian, Gipuzkoa, Spain (Azpeitia et al., 2018)	de (3), fr (3), ru (3), zh (3)

Table 2: Shared task systems: system label, team affiliation, publication reference, number of runs for each language pair

tences (dumps as of 20161201[2]) in two chosen languages. The parallel part is made of News Commentary sentences (v11[3]). As mentioned above, the instructions required task participants not to use any of these two corpora in their methods and systems. Datasets were prepared for four language pairs, each of which included English and another language among German (de), French (fr), Russian (ru), and Chinese (zh). Each dataset contained sample, training, and test splits (see Table 1).

Given a dataset containing two monolingual corpora en and fr, systems were expected to produce a set of sentence pairs (s^i_{en}, s^i_{fr}). Evaluation was performed by comparing system pairs to the set of gold standard pairs, and computing precision, recall, and F1-score in the usual way.

Note that the gold standard was defined by artificially inserted sentences. There is however a non-zero chance that some other pairs of sentences naturally happen to be translations too. If a system finds such correct sentence pairs that are not part of the gold standard annotations, these pairs are counted as false positives. As a result, the precision of system runs can be underestimated. By reviewing a small sample of false positive sentence pairs in the most precise en-fr run of one of the Second BUCC Shared Task participants (Zweigenbaum et al., 2017), we computed a very rough estimate of the number of such sentence pairs. We considered as correct translations sentence pairs such that (i) "the two sentences are completely equivalent, as they mean the same thing," possibly also considering cases in which (ii) "the two sentences are mostly equivalent, but some unimportant details differ." These correspond to the top two grades (5 and 4) in the guidelines of cross-language sentence similarity in SemEval 2016 (Agirre et al., 2016). Lower grades, e.g. (3) in which "the two sentences are roughly equivalent, but some important information differs or is missing" were not considered correct translations. Table 3 lists examples of sentence pairs considered false positives according to the gold standard, together with the human judgments (s) they received. Two sentence pairs in Table 3 received different scores from the two judges.

fr	en	s
Le renforcement de la gendarmerie locale par des troupes européennes est vite envisagé.	The reinforcement of the local gendarmerie with European troops was quickly planned.	5
Avant la *Première Guerre mondiale*, l'Allemagne importait annuellement pour 1,5 milliard de Reichsmarks de matières premières en provenance de Russie.	Germany imported 1.5 billion Rechsmarks of raw materials *and other goods* annually from Russia before the war.	4, 5
Le Mozambique est l'un des pays les plus pauvres du monde.	Mozambique is one of the poorest *and most underdeveloped* countries in the world.	4
Le jeu comporte aussi *plusieurs* modes de jeu, qui peuvent être joué en solo ou en multijoueur local:	Competitive multiplayer modes have also *been added*, and can be played locally or over a network.	3, 4
Dans le deuxième, le type cystovarien, les ovocytes sont transmis à l'extérieur, par le biais de l'oviducte.	In the third type, the oocytes are conveyed to the exterior through the oviduct.	3

Table 3: Example sentence pairs found in false positive system output, with associated human cross-language similarity scores s. Italics emphasize extra material

We found that the resulting understimate of precision for that participant was between 0.6 and 4 points depending on whether only grade 5 pairs were considered correct, whether grade 4 pairs were also deemed acceptable, and on

[2] http://ftp.acc.umu.se/mirror/wikimedia.org/dumps/

[3] http://www.casmacat.eu/corpus/news-commentary.html

how discordances across annotators were reconciled. Participants with less precise results were less subject to this phenomenon, therefore this did not change rankings.

3. Participants and systems

16 teams downloaded datasets, among which three teams submitted runs. Table 2 gives more detail about teams and runs.

Systems addressed the bilingual dimension of the task with machine translation systems (H2@BUCC2018, nlp2ct2), or used parallel corpora to obtain word translations (VIC) or to train bilingual word embeddings (H2@BUCC2018) or an autoencoder (nlp2ct2).

Cross-language sentence similarity was handled by the Jaccard coefficient (VIC) or the BLEU score (H2@BUCC2018), possibly with weighting (a function of frequency: VIC) and with a trained classifier (H2@BUCC2018, nlp2ct2).

One team used an Information Retrieval engine for faster search of similar sentences (VIC), where as the others took advantage of the fast computation of the Cosine of word embeddings (H2@BUCC2018) or of the orthogonal denoising encoder output (nlp2ct2).

4. Results and discussion

We present evaluation results for the runs submitted for each language. In each table we show the precision, recall and F1-score of each run in percentages. In addition, we show the best run of 2017 when available for that language pair. Because the evaluation performed through this synthetic dataset, with artificially inserted translation pairs, only approximates what a human evaluation of system results would return, it would not be relevant to compute scores with many digits: therefore we round the computed figures to the nearest integer.

Table 4 shows results for the three runs submitted on the German-English (de-en) language pair (one team). As in 2017, this language pair obtains the best results. Table 5 presents the six runs submitted on the French-English (fr-en) language pair by two teams. Table 6 presents the three runs submitted on the Russian-English (ru-en) language pair by one team. This language pair did not receive any submissions in 2017. Table 7 presents the five runs submitted on the Chinese-English (zh-en) language pair by two teams. They all improve upon the previous year's zh-en results.

5. Conclusion

The third BUCC 2018 Shared Task addressed spotting parallel sentences in comparable corpora. The best results of the participants are high, with precisions of 89–91%, recalls of 75–84%, and F1-scores of 77–86%. The Russian-English language pair was attempted for the first time, and the Chinese-English language pair was again the most challenging. F1-scores improved over 2017 for all language pairs. The BUCC 2018 Shared Task dataset and evaluation program can be downloaded from the shared task's Web page.[4]

run_name	sys_n	P	R	F1
VIC1.de-en	9271	87	**84**	**86**
VIC3.de-en	8265	**91**	79	85
VIC2.de-en	8769	88	81	84
VIC1.de-en in 2017	8640	88	80	**84**

Table 4: Evaluation (%) of de-en runs (n_gold=9,550)

run_name	sys_n	P	R	F1
VIC1.fr-en	8136	86	77	**81**
VIC2.fr-en	7173	**91**	72	80
VIC3.fr-en	8887	80	**79**	80
H2@BUCC18_1_fr-en	7947	82	72	76
H2@BUCC18_2_fr-en	9607	71	75	73
H2@BUCC18_3_fr-en	8300	70	64	67
VIC1.fr-en in 2017	8831	80	79	**79**

Table 5: Evaluation (%) of fr-en runs (n_gold=9,043).

run_name	sys_n	P	R	F1
VIC1.ru-en	11010	86	77	**81**
VIC2.ru-en	10127	**90**	71	79
VIC3.ru-en	11370	79	**79**	79

Table 6: Evaluation (%) of ru-en runs (n_gold=14,330)

Acknowledgments

We thank the participants for their interest in this task. This work was partially funded by the European Union's Horizon 2020 Marie Sklodowska Curie Innovative Training Networks—European Joint doctorate (ITN-EJD) Under grant agreement No:676207 (MiRoR). Part of this work was supported by a Marie Curie Career Integration Grant within the 7th European Community Framework Programme.

6. References

Abdul Rauf, S. and Schwenk, H. (2009a). Exploiting comparable corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 46–54, Singapore, August. Association for Computational Linguistics.

Abdul-Rauf, S. and Schwenk, H. (2009b). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece, March. Association for Computational Linguistics.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego,

[4] https://comparable.limsi.fr/bucc2018/bucc2018-task.html.

run_name	sys_n	P	R	F1
VIC1.zh-en	1680	80	71	75
VIC2.zh-en	1373	**89**	64	74
VIC3.zh-en	1763	80	**75**	**77**
nlp2ct1.zh-en	1169	73	45	55
nlp2ct2.zh-en	1209	72	46	56
zNLP1 in 2017	1985	42	**44**	**43**

Table 7: Evaluation (%) of zh-en runs (n_gold=1,896)

California, June. Association for Computational Linguistics.

Azpeitia, A., Etchegoyhen, T., and Martínez Garcia, E. (2018). Extracting parallel sentences from comparable corpora with STACC variants. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May. ELRA.

Bouamor, H. and Sajjad, H. (2018). H2@BUCC18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May. ELRA.

Buck, C. and Koehn, P. (2016). Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany, August. Association for Computational Linguistics.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183.

Hewavitharana, S. and Vogel, S. (2011). Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 61–68, Portland, Oregon, June. Association for Computational Linguistics.

Leong, C., Wong, D. F., and Chao, L. S. (2018). UMpAligner: Neural network-based parallel sentence identification model. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May. ELRA.

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

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia, July. Association for Computational Linguistics.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In Daniel Marcu Susan Dumais et al., editors, *HLT-NAACL 2004: Main Proceedings*, pages 265–272, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In Pamela Forner, et al., editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rapp, R., Sharoff, S., and Zweigenbaum, P. (2016). Recent advances in machine translation using comparable corpora. *Natural Language Engineering*, 22(4):501–516, July.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, student session*, volume 1, pages 320–322, Boston, Mass.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In Serge Sharoff, et al., editors, *Building and Using Comparable Corpora*, pages 1–20. Springer, Berlin Heidelberg, December.

Sharoff, S., Zweigenbaum, P., and Rapp, R. (2015). BUCC shared task: Cross-language document similarity. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 74–78, Beijing, China, July. Association for Computational Linguistics.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2016). Towards preparation of the second BUCC shared task: Detecting parallel sentences in comparable corpora. In *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*, pages 38–43, PortoroÅŸ, Slovenia, May. European Language Resources Association (ELRA).

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada, August. Association for Computational Linguistics.

H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings

Houda Bouamor[1] **and Hassan Sajjad**[2]
[1]Carnegie Mellon University in Qatar, [2]Qatar Computing Research Institute, HBKU, Qatar
hbouamor@cmu.edu, hsajjad@qf.org.qa

Abstract

This paper presents our solution for the BUCC 2018 Shared Task on parallel sentence extraction from comparable corpora. Our system identifies parallel sentence pairs in French-English corpora by following a hybrid approach pairing multilingual sentence-level embeddings, neural machine translation, and supervised classification. Our system consists of a two-step process. In the first step, to reduce the size and the noise of the candidate sentence pairs, we filter the target translation candidates using the continuous vector representation of each source-target sentence pair learned using a bilingual distributed representation model. Then we select the best translation using a neural machine translation system or a binary classification model. We achieve an F_1-score of up to 75.2 and 76.0 on the BUCC18 train and test sets respectively.

Keywords: Comparable Corpora, Parallel Sentences, Multilingual Embeddings, Neural Machine Translation, Supervised Classification

1. Introduction

Building standard machine translation (MT) systems require a large amount of sentence-aligned parallel corpora. While these resources are available for mainstream languages (i.e., English, French, German, and Arabic) and domains, unfortunately, many low resourced languages and specialized domains suffer from the scarcity of such corpora. The manual generation of parallel data for several language pairs needs human expertise, a costly and time-consuming task. Although this problem can be alleviated by exploiting a pivot language to bridge the source and target languages (Cohn and Lapata, 2007; El Kholy et al., 2013; Sajjad et al., 2013; Cheng et al., 2017), the performance of such systems is never comparable to the ones built using parallel corpora. The scarcity of these resources pushed researchers to investigate the use of comparable corpora (Bouamor et al., 2013a; Rapp et al., 2016).

Comparable corpora include non-aligned sentences, phrases or documents that are not an exact translation of each other but share common features such as domain, genre, sampling period, etc. (Wu and Fung, 2005). Wikipedia articles describing the same topic, but written in two different languages (Barrón-Cedeño et al., 2015) and news topics covered in different newspapers appearing the same day, reporting about the same event or describing the same subject, are both good examples of comparable corpora. These resources could be leveraged to automatically extract parallel sentence pairs and build a parallel corpus between two languages. In recent years, there has been a body of work related to MT based on non-parallel comparable corpora. Rapp et al. (2016) gives a detailed survey of the use of comparable corpora in MT and several other NLP tasks.

In this work, we present our solution for the BUCC 2018 Shared Task on parallel sentence extraction from comparable corpora. Our system identifies parallel sentence pairs in French-English corpora by defining a hybrid approach pairing multilingual sentence-level embeddings, neural machine translation, and supervised classification.

The two monolingual corpora provided in the shared task are of approximately 370K and 270K sentences. Here, every target sentence is a candidate translation of every source sentence. The search space for the number of comparisons is is very large. To tackle this, we propose a two-step process. In the first step, in order to reduce the size of the candidate sentences, we filter the English translation candidates using the continuous vector representation of each French-English sentence pair learned using a bilingual distributed representation model. Then we select the best translation by leveraging the output of a neural machine translation system or a supervised classification model.

The remainder of this paper is organized as follows: We give a detailed description of our approach in Section 2.. Then, we present our experimental setup in Section 3.. We finally report and discuss our system results in Section 4..

2. Approach

When dealing with comparable corpora, every sentence in the target corpus can be a potential translation of every source sentence. Given a source corpus of S sentences and a target corpus of T sentences, the number of comparisons required to find translation pairs are $S \times T$. Given the large size of S and T, the search space becomes very large to find translation pairs from the corpus efficiently. In this work, we split the process of parallel sentence extraction into two steps: The first step reduces the search space from millions of comparisons to a few hundreds of top candidate pairs. In the second step, we select the best translation from the list of candidate pairs.

In the first step, we use multilingual sentence embeddings to identify top N closest target sentences to a source sentence. In the second step, we use machine translation, a machine translation evaluation metric, and binary classifier to select the best translation from the list of N candidate pairs.

2.1. Bilingual Distributed Representations

Monolingual distributed word representations have shown great potential in boosting the performance in several NLP tasks (Iacobacci et al., 2015; Guzmán et al., 2016; Santos et al., 2017). The use of word embeddings was further extended to include multilingual tasks (Zou et al., 2013;

Adams et al., 2017; Ammar et al., 2016), where distributed representations are induced over different language-pairs and thus serve as an effective way of capturing linguistic regularities in words that share same semantic and syntactic space, across languages (Gouws et al., 2015). However, there is a major problem with using monolingual word embeddings in a multilingual scenario. The models are usually trained independently for each of the languages using vector spaces. Thus, measuring the similarity between words is a challenging task, even for similar words.

Much research work has been conducted to address this problem, following several approaches (Luong et al., 2015): (i) *Bilingual mapping*, where word representations are trained for each language independently, and a linear mapping is then learned to transform representations from one language to another (Mikolov et al., 2013; Grégoire and Langlais, 2017); (ii) *Monolingual adaptation* that relies on pre-trained embeddings of the source language when learning target representations (Zou et al., 2013); and (iii) *Bilingual training* aiming at jointly learning representations for both languages using a parallel corpus, benefiting from word alignments (Luong et al., 2015) or without word alignments (Gouws et al., 2015).

In our model, we exploit the power of bilingual distributed representations to identify highly similar sentences in a $fr-en$ comparable corpus. For this, we use multivec (Bérard et al., 2016), an implementation of (Luong et al., 2015)'s bivec model for bilingual distributed representations. This toolkit is used for computing continuous representations for text at different granularity levels (word-level or sequences of words).[1]

Similarly to word2vec (Mikolov et al., 2013), for each pair of sentences in a parallel corpus, bivec tries to predict words in the same sentence, but also uses words in the source sentence to predict words in the target sentence (and conversely).

Following this approach, we first train multivec on a large $fr-en$ parallel corpus, to build a bilingual sentence level embedding model in the same vector space. Then, we use the model to learn a continuous representation for each source and target sentences from the train and test datasets provided in the shared task.

Our system detects a parallel sentence pair by measuring the cosine similarity between a sentence vector $\vec{f_i}$ of each French sentence fr_i (in the source corpus) and each vector $\vec{e_j}$ corresponding to a possible en_j candidate (in the target corpus). We define each sentence embedding defined as an average of the source word embeddings of its constituent words. We create our set of candidate pairs by keeping the top N most similar target sentences for each source sentence fr_i (as per the cosine similarity measure) .[2]

2.2. Candidate Filtering

We follow two approaches to filter further the parallel sentence candidates obtained using the multilingual vector similarity: machine translation and supervised classification.

2.2.1. Machine Translation

To this point, we have a list of translation candidate sentences for every source sentence. We have reduced our search space of comparison from thousands of options to 10 and 100 options by using the bilingual distributed representations. In order to choose the best translation for each source sentence, the ideal scenario would require a reference sentence against which we can compare the candidate translations and keep the closest one. We use machine translation to produce a "reference" translation for each source sentence.

We hypothesize that given a machine translation system of decent quality, translation of a source sentence should be closest to its parallel sentence in the target language. To achieve this, we translate all French sentences in the comparable corpora to English using the French to English machine translation system. Then, for every French sentence, we compare its translation against all the English candidate sentences. The candidate sentence that gives the highest BLEU (Papineni et al., 2002) above certain threshold is selected as a translation of the source sentence. We use a high threshold above 50 BLEU point to discard source sentences that do not have any matching translations among the candidate translations.

2.2.2. Supervised Binary Classification

Machine translation systems are not perfect and can induce translation errors and noise, which impacts the quality of the sentence pairs identified. In order, to experiment with a more straightforward approach that leverages only the source-target sentence pairs without any intermediate step, we explore the use of supervised classification.

After obtaining the top N candidate source-target parallel sentence pairs from the first step (described in Section 2.1.), we build a bi-class classifier to identify parallel sentences among them, without translating the source sentences into the target language. Our system takes as input a French sentence FR and each of its English candidate EN$_n$ (considered here as a possible translation) and outputs a score for each pair FR-EN$_n$ estimating a kind of translation quality. The parallel sentence pair FR-EN$_{best}$ selected is the one that has the highest quality score.

For this, we use a Support Vector Machine (SVM) classifier and exploit a rich set of features to represent a French source language sentence and each of its English translation candidates.

Learning features: We use the following group of features which have been used in work related to translation quality estimation for several languages (Bouamor et al., 2013b; Specia et al., 2015).

- **General features**: For each sentence, we use different features modeling its length in terms of words, the ratio of source-target length, source-target punctuation marks, numerical characters, and source-target content words.[3]

[1] https://github.com/eske/multivec

[2] Since we are working with vector representations, doing the Cartesian product is possible.

[3] As the English candidates are not the output of a machine translation system, there was no need to use language modeling or MT-based features (such as perplexity scores or number of OOVs)

- **Morphosyntactic features:** We use features to model the difference of sequences of POS tags for a pair of source-target sentences. These features measure the POS preservation between a source sentence and its target candidate. We compute the absolute difference between the number of different POS tags. We also indicate the percentage of nouns, verbs, and adjectives in the source and target sentences. The source and target sentences were tagged respectively using the French and English distributions of the Stanford coreNLP pipeline (Manning et al., 2014).
- **Named Entity features** A pair of parallel sentences usually contains the same number and type of Named Entities (NEs) (a translation/transliteration of each other). We use this hypothesis to measure the difference in number of various types of named entities in the source-target candidate parallel sentences. We use the CoreNLP named entity recognizer to extract persons, locations, organizations, and dates.

3. Experimental Setup

We experiment with different configurations and following several approaches. We present in this section our experimental settings and describe the datasets and tools used.

3.1. Dataset

In addition to the $fr-en$ training and testing datasets (BUCC18$_{train}$ and BUCC18$_{test}$) provided in the Shared Task, we use the $fr-en$ Europarl parallel corpus (Koehn, 2005) containing 2 million sentence pairs, as well as a News corpus made available from WMT 2016 with 183,000 aligned $fr-en$ sentence pairs (Bojar et al., 2016)(Europarl+News). All the corpora (French and English) are preprocessed through the following steps: Tokenization, POS tagging, and Name-Entity recognition. These preprocessing steps are completed using The Stanford CoreNLP Toolkit (Manning et al., 2014).

3.2. Model Training Settings

Continuous Vector Modeling: to train our bilingual model, we use the parallel $fr-en$ Europarl+News corpus described above, with the default configuration of the multivec tool. The model was trained using a learning rate α set to 0.05, a sample (a threshold on words' frequency) set to 0.001 and a window size of 5.

Machine Translation: we use the OpenNMT toolkit (Klein et al., 2017) to train a 2-layered LSTM encoder-decoder with attention (Bahdanau et al., 2015). In order to keep the training and test time low, we restrict ourself to uni-directional LSTM model. We use the default settings: embedding layer size: 512, hidden layer size: 1,000.
We limit the vocabulary to 40,000 words using BPE (Sennrich et al., 2016) with 40,000 operations. The sub-word units help us to map various morphological variations of a word to known sub-units. It also fixes the mismatch of vocabulary between our training corpus of machine translation and comparable corpus by splitting the unknowns in the comparable corpus into known sub-word units of the training corpus.

Binary Classification: we use the models described in Section 2.2.2. to build a Support Vector Machine (SVM) binary classifier using the LinearSVC implementation of scikit-learn[4].
To train our classifier we needed a gold standard corpus where a pair of $fr-en$ sentence is labeled as having high or low translation quality.
In order to build this dataset, we use the Europarl-News parallel corpus. Each sentence pair in this corpus is considered as a positive example (high translation quality). We then built a set of negative training examples (low translation quality), by selecting sentences from the French part of the corpus and randomly assigning a sentence from the English part to them. 80% of this corpus is used for training and 20% for testing. None of the sentences provided in the Shared task are used in building this classification model.

3.3. Evaluation Protocol

We evaluate our models, after obtaining the final predicted $fr-en$ parallel sentence pairs, using precision (P), recall (R), and F_1 score, defined in the shared task as follows:

$$P = \frac{TP}{TP+FP};\ R = \frac{TP}{TP+FN};\ F_1 = \frac{2PR}{P+R}$$

Where TP stands for the number of $fr-en$ sentence pairs that are present in the gold standard provided. A false positive (FP) is a pair of sentences that are not present in the gold standard. And a false negative (FN) is a pair of sentences present in the gold standard, but absent from system results.

4. Evaluation and Results

We tested several configurations:

Baseline: Our baseline consists of selecting $fr-en$ sentence pairs predicted only by the cosine similarity between sentence embedding pairs (described in Section 2.1.) with N=10. Since our method looks for a translation for every French sentence, we have a large number of false positives. Later, we use machine translation and classification to filter out these false positive pairs.

Machine translation: We take N=10 best candidates from our baseline system. For every French sentence, we compare its English translation generated automatically using a machine translation system against the ten candidate sentences. We sort the candidates based on BLEU and choose a translation with the best BLEU score above a certain threshold. Table 1 shows the results on the BUCC18$_{train}$ set when tested for different values of BLEU. The multivec-10best shows the highest initial recall of the list before applying BLEU-based filtering. The system achieved best f-score at BLEU value 0.57. It is interesting to see that a small difference in BLEU threshold dropped the recall by more than two points. This could be due to the nature of the BLEU metric that prefers exact ngram matches and penalizes words that are only different

[4]available at:`http://scikitlearn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

by a small morphological change. We suspect that BLEU at the sub-word level or Meteor would be less sensitive to small threshold changes and may result in a better balance between precision and recall.

Method	**P**	**R**	**F_1**
multivec-10best	-	83.4	-
BLEU-0.00	2.8	82.3	5.3
BLEU-0.50	60.6	77.3	67.9
BLEU-0.52	62.2	74.5	72.2
BLEU-0.55	79.0	71.7	75.2
BLEU-0.57	83.9	69.1	75.8
BLEU-0.59	87.5	65.8	75.1

Table 1: Precision, recall and F_1 on BUCC18$_{train}$, when filtered for various BLEU thresholds. **multivec-10best** shows the oracle recall that our system can achieve.

Classification: We measure the accuracy of our classifier on the external dataset (Europarl+News$_{test}$) as well as the train and test sets provided for the French-English task: BUCC18$_{train}$ and BUCC18$_{test}$. The source-target pairs that exist in the training and testing gold standards have been considered as positive examples, and an equivalent number from the rest of the pairs, generated by applying the multilingual word embedding based approach are considered as negative examples. Table 2 reports the accuracy of the classifier on different test sets of different nature and various sizes. The results obtained are encouraging, as we only exploit a group of basic features and do not include any semantic features such as sentence vector similarity or machine translation features.

H2@BUCC-2018 Results: We submitted three runs of our system:

- **Run1**: 10 best candidates with a BLEU filtering threshold of 0.52;
- **Run2**: 10 best candidates with a BLEU filtering threshold of 0.55;
- **Run3**: 10 best candidates with the SVM binary classification model output.

Table 3 summarizes the results for the three runs. Because of the time constraint, we reduced the number of candidate sentences to 10 only. This caused a loss of more than 16% in recall. In future, we would like to increase the candidate list to 100 candidates. This would slow down the filtering process but would result in better F_1 score.
The best machine translation results mentioned in Table 1 dropped the recall by 14 points. In future, we would like to consider other metrics like classification and sentence embedding in combination with MT results to improve the loss in recall.

	#of examples	**Accuracy**
Europarl+News$_{test}$	437,103	**81.05**
BUCC18$_{train}$	18,178	72.60
BUCC18$_{test}$	18,086	72.73

Table 2: Accuracy of the classifier on different test sets. The size of each test set is indicated.

	P	**R**	**F_1**
Run1	71	**75**	73
Run2	**82**	72	**76**
Run3	70	64	67

Table 3: Official results of our system on the BUCC2018 Testset

5. Conclusion

In this work, we presented our system to extract parallel sentences from comparable corpora. Initially, we learned sentence embedding vectors of the source and target languages using a parallel corpus. For every source sentence, we found the closest target sentence embeddings to create a list of candidate sentences. We then chose the best translation from the candidate translations by considering it either as a machine translation evaluation task or a binary classification task of choosing the best translation given a source sentence. Our method achieved an F_1-score of up to 75.2 and 76.0 on the BUCC18 train and test sets respectively.

6. References

Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-Lingual Word Embeddings for Low-Resource Language Modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain.

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively Multilingual Word Embeddings. *CoRR*, abs/1602.01925.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate.

Barrón-Cedeño, A., España Bonet, C., Boldoba, J., and Màrquez, L. (2015). A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13, Beijing, China.

Bérard, A., Servan, C., Pietquin, O., and Besacier, L. (2016). MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož, Slovenia.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Bouamor, D., Popescu, A., Semmar, N., and Zweigenbaum, P. (2013a). Building Specialized Bilingual Lexicons Using Large Scale Background Knowledge. In *Proceedings of the 2013 Conference on Empirical Methods in*

Natural Language Processing, pages 479–489, Seattle, Washington, USA.

Bouamor, H., Mohit, B., and Oflazer, K. (2013b). SuMT: A Framework of Summarization and MT. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 270–278, Nagoya, Japan.

Cheng, Y., Yang, Q., Liu, Y., Sun, M., and Xu, W. (2017). Joint training for pivot-based neural machine translation. In *Proceedings of IJCAI*.

Cohn, T. and Lapata, M. (2007). Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic.

El Kholy, A., Habash, N., Leusch, G., Matusov, E., and Sawaf, H. (2013). Language Independent Connectivity Strength Features for Phrase Pivot Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Sofia, Bulgaria.

Gouws, S., Bengio, Y., and Corrado, G. (2015). BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning*.

Grégoire, F. and Langlais, P. (2017). BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 46–50, Vancouver, Canada.

Guzmán, F., Bouamor, H., Baly, R., and Habash, N. (2016). Machine Translation Evaluation for Arabic using Morphologically-enriched Embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1398–1408, Osaka, Japan.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado, June. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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, Morristown, NJ, USA.

Rapp, R., Sharoff, S., and Zweigenbaum, P. (2016). Recent Advances in Machine Translation using Comparable Corpora. *Natural Language Engineering*, 22(4):501–516.

Sajjad, H., Darwish, K., and Belinkov, Y. (2013). Translating Dialectal Arabic to English. In *Proceedings of the 51st Conference of the Association for Computational Linguistics (ACL)*.

Santos, L., Corrêa Júnior, E. A., Oliveira Jr, O., Amancio, D., Mansur, L., and Aluísio, S. (2017). Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1284–1296, Vancouver, Canada.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Specia, L., Paetzold, G., and Scarton, C. (2015). Multilevel Translation Quality Prediction with QuEst++. In *ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China.

Wu, D. and Fung, P. (2005). Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-comparable Corpora. In *International Conference on Natural Language Processing*, pages 257–268. Springer.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA.

Extracting Parallel Sentences from Comparable Corpora with STACC Variants

Andoni Azpeitia, Thierry Etchegoyhen and Eva Martínez Garcia
Vicomtech, Donostia/San Sebastián, Spain
{aazpeitia, tetchegoyhen, emartinez}@vicomtech.org

Abstract

This article describes our submissions to the BUCC 2018 shared task on parallel sentence extraction from comparable corpora. Our approach is based on variants of the STACC method, which computes similarity on expanded lexical sets via Jaccard similarity. We apply the weighted variant of the method to all four language pairs of the task, demonstrating the efficiency and portability of the approach. Additionally, we introduce a variant which further penalizes mismatches in terms of named entities, improving over the already strong weighted variant baseline in most cases. Our approach reached the highest results in all scenarios, with scores over 80% in terms of f1-measure and 90% in precision.

Keywords: BUCC 2018, Shared Task, Sentence Alignment, Comparable Corpora

1. Introduction

The exploitation of comparable corpora is an important research area (Munteanu and Marcu, 2005; Sharoff et al., 2016), as it contributes to the creation of the parallel corpora that are needed for multilingual natural language processing tasks such as data-driven machine translation (Brown et al., 1990; Bahdanau et al., 2015) or automated bilingual dictionary creation (Rapp, 1995).

Extracting parallel sentences from comparable corpora is a challenging task, which has given rise to the development of a wide range of approaches over the years. Thus, interesting results have been notably obtained with methods based on suffix trees (Munteanu and Marcu, 2002), maximum likelihood (Zhao and Vogel, 2002), binary classification (Munteanu and Marcu, 2005), cosine similarity (Fung and Cheung, 2004), reference metrics over statistical machine translations (Abdul-Rauf and Schwenk, 2009; Sarikaya et al., 2009), feature-based approaches (Stefănescu et al., 2012; Smith et al., 2010) or deep learning with bidirectional recurrent neural networks (Grégoire and Langlais, 2017), among others.

For our participation in the BUCC 2018 shared task on extracting parallel sentences from comparable corpora, we followed the STACC approach of (Etchegoyhen et al., 2016; Etchegoyhen and Azpeitia, 2016), which is based on Jaccard similarity (Jaccard, 1901) over lexical sets, with additional set expansion operations to address named entities and morphological variation.

We selected as our baseline the weighted variant of the approach (Azpeitia et al., 2017), which proved highly successful on the BUCC 2017 shared task (Zweigenbaum et al., 2017), and applied the approach to all four language pairs in the 2018 task. Additionally, we designed a variant of this approach which further penalizes mismatches in terms of named entities, showing that it improves over the strong weighted STACC baseline in most cases.

The results obtained in this shared task confirm the efficiency and portability of our approach, and additionally demonstrate the specific importance of named entities for parallel sentence extraction from comparable corpora.

2. STACC

The STACC approach has been described and explored in detail in (Etchegoyhen and Azpeitia, 2016), and we briefly summarise below how similarity is computed with their method.

Let s_i and s_j be two tokenised and truecased sentences in languages l_1 and l_2, respectively, S_i the set of tokens in s_i, S_j the set of tokens in s_j, T_{ij} the set of lexical translations into l_2 for all tokens in S_i, and T_{ji} the set of lexical translations into l_1 for all tokens in S_j.

Lexical translations are initially computed from sentences s_i and s_j by retaining the k-best translations for each word, if any, as determined by the ranking obtained from the lexical translation probabilities computed with IBM word alignment models (Brown et al., 1990). The sets T_{ij} and T_{ji} that comprise these k-best lexical translations are then expanded by means of two operations:

1. For each element in the set difference $T'_{ij} = T_{ij} - S_j$ (respectively $T'_{ji} = T_{ji} - S_i$), and each element in S_j (respectively S_i), if both elements share a common prefix with minimal length of more than n characters, the prefix is added to both sets. This longest common prefix matching strategy is meant to capture morphological variation via minimal computation.

2. Numbers and capitalised truecased tokens not found in the translation tables are added to the expanded translation sets. This operation addresses named entities, which are strong indicators of potential alignment given their low relative frequency and are likely to be missing from translation tables trained on different domains.

With source and target sets as defined here, the STACC similarity score is then computed as in Equation 1:

$$stacc(s_i, s_j) = \frac{\frac{|T_{ij} \cap S_j|}{|T_{ij} \cup S_j|} + \frac{|T_{ji} \cap S_i|}{|T_{ji} \cup S_i|}}{2} \quad (1)$$

Similarity for the core metric is thus defined as the average of the Jaccard similarity coefficients obtained between sentence token sets and expanded lexical translations in both directions.

2.1. STACC$_w$

In (Azpeitia et al., 2017), the STACC$_w$ variant of the core method is described, where set membership values of 1 in the original approach are replaced with lexical weights. The weights are computed according to Equation 2, where $f(w_i)$ is the relative frequency of word w_i and α is a parameter controlling the smoothness of the curve.

$$W(w_i) = \frac{1}{e^{\sqrt{\alpha \cdot f(w_i)}}} \quad (2)$$

Weighting can be computed on each monolingual corpus to be aligned, as will be the case for all the results reported in this paper, or on separate monolingual corpora. STACC$_w$ similarity is computed according to the weighted Jaccard similarity formula described in Equation 3, for a given lexical translation set T and token set S:

$$WJ(T,S) = \frac{\sum_{w_m \in \{T \cap S\}} W(w_m)}{\sum_{w_n \in \{T \cup S\}} W(w_n)} \quad (3)$$

The complete weighted similarity score is thus computed according to Equation 4.

$$stacc_w(s_i, s_j) = \frac{WJ(T_{ij}, S_j) + WJ(T_{ji}, S_i)}{2} \quad (4)$$

This variant was rather successful on the BUCC 2017 shared task, as it significantly improved over the baseline version of STACC, which would have already obtained the best results on all metrics in the two language pairs alignment scenarios in which the system participated.

2.2. STACC$_{wp}$

For this version of the BUCC shared task, we introduced a new variant, based on STACC$_w$ and on a penalty oriented towards named entity mismatches.

Both STACC and STACC$_w$ include a treatment of named entities, defined in terms of surface forms, by including in the expanded translation sets both capitalised words and numbers. Intuitively though, named entities might be thought of as playing an even stronger role than simply participating in determining similarity: when glancing over sets of comparable sentences, checking mismatches in terms of named entities between a given pair of sentences seems an efficient method to at least quickly discard improbable alignments.

We tested this hypothesis by first defining a penalty as in Equation 5, where N_i and N_j denote the sets of surface-form entities in the source and target sentence, respectively.

$$nep(s_i, s_j) = \frac{|(N_i - N_j) \cup (N_j - N_i)|}{|S_i \cup S_j|} \quad (5)$$

The penalty is thus defined in terms of set differences, taking as numerator the union of entities that are present in one sentence but not in the other. By defining the denominator as the union of all tokens in the source and target sentences, the measure is bound between 0 and 1, and a higher penalty will be assigned to sentence pairs with larger numbers of mismatching entities.

For this STACC$_{wp}$ variant, the penalty is included in the computation of the final score according to Equation 6.

$$stacc_{wp}(s_i, s_j) = stacc_w(s_i, s_j) - nep(s_i, s_j) \quad (6)$$

Thus, this variant preserves the successful core weighted metric for all cases where either no entities are present in the source and target sentences, or when the same entities are present in both sentences. The penalty complements the core metric by gradually reducing the overall score as entity mismatches increase between the source and target sentences.

3. BUCC 2018 Shared Task

The BUCC 2018 shared task on parallel sentence extraction from comparable corpora[1] consists in identifying translation pairs within two sentence-split monolingual corpora. It involves four language pairs and we applied the variants of our approach in all four alignment scenarios. The organisers provided three datasets for each language pair, whose statistics are described in Table 1; gold reference pairs were provided for the training and sample sets.

3.1. Experimental Settings

The volumes of data selected for the task makes it unrealistic to compute the alignments over the Cartesian products of source and target sentences. Thus, we use the STACC system in cross-language information retrieval (CLIR) mode, where target sentences are first indexed using the Apache Lucene toolkit [2] and retrieved by building a query over the expanded sets created from each source sentence.

This strategy drastically reduces the computational load, at the cost of missing some correct alignment pairs. Similarity is computed for each source sentence against all retrieved candidates and a final optimisation is applied to enforce 1-1 alignments, a process which has been shown to improve the quality of alignments (Etchegoyhen and Azpeitia, 2016).

For each language pair, weighting was computed on each monolingual corpus composing the pair to be aligned. Translation tables were generated with the GIZA++ toolkit (Och and Ney, 2003) for all language pairs but Russian-English, for which word alignments were computed with FastAlign (Dyer et al., 2013).

To train the word alignment models, we followed the approach in (Azpeitia et al., 2017) and created generic corpora via bilingual perplexity-based sampling, with an arbitrary upper data selection bound to avoid over-representing individual corpora. Note that, due to time availability to prepare our submissions, this method was not applied to our two new language pairs, Russian-English and Chinese-English, for which we only used the MULTIUN corpus, in totality for the former, and a sample of approximately 2 million for the latter. Table 2 describes the number of sentence pairs selected for each language pair.[3]

[1] https://comparable.limsi.fr/bucc2018/bucc2018-task.html

[2] https://lucene.apache.org.

[3] All original corpora were downloaded from the OPUS repository (Tiedemann, 2012): http://opus.lingfil.uu.se/; the upper selection bound was set to 500,000 sentence pairs after considering the relative weights of the available corpora.

PAIR	LANG	MONOLINGUAL			GOLD		
		SAMPLE	TRAIN	TEST	SAMPLE	TRAIN	TEST
DE-EN	de	32,593	413,869	413,884	1,038	9,580	9,550
	en	40,354	399,337	396,534	1,038	9,580	9,550
FR-EN	fr	21,497	271,874	276,833	929	9,086	9,043
	en	38,069	369,810	373,459	929	9,086	9,043
RU-EN	ru	45,459	460,853	457,327	2,374	14,435	14,330
	en	72,766	558,401	566,356	2,374	14,435	14,330
ZH-EN	zh	8,624	94,637	91,824	257	1,899	1,896
	en	13,589	88,860	90,037	257	1,899	1,896

Table 1: Task data statistics (number of sentences)

PAIR	DATA	CORPUS					
		OPENSUBS	MULTIUN	EUROPARL	JRC	TED	GENERIC
DE-EN	Original	11,473,328	103,490	1,776,292	449,818	138,243	*13,941,171*
	Selected	500,000	103,490	500,000	449,818	139,243	*1,692,551*
FR-EN	Original	28,024,360	9,142,161	1,826,770	708,896	153,167	*39,855,354*
	Selected	500,000	500,000	500,000	316,327	153,167	*1,969,494*
RU-EN	Original	-	9,111,212	-	-	-	*9,111,212*
	Selected	-	9,111,212	-	-	-	*9,111,212*
ZH-EN	Original	-	7,747,328	-	-	-	*7,747,328*
	Selected	-	1,831,016	-	-	-	*1,831,016*

Table 2: Generic data (number of sentences)

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.15	99.04	95.09	**91.51**	**93.27**
SAMPLE	STACC$_{wp}$ (F)	250	0.15	99.04	97.36	89.01	93.00
SAMPLE	STACC$_{wp}$ (P)	250	0.16	99.04	**99.21**	85.54	91.87
TRAIN	STACC$_w$ (F)	250	0.17	98.50	87.00	79.96	83.33
TRAIN	STACC$_{wp}$ (F)	250	0.16	98.50	84.81	**83.74**	**84.27**
TRAIN	STACC$_{wp}$ (P)	250	0.17	98.50	**89.86**	78.28	83.67
TEST	STACC$_w$ (F)	250	0.17	98.65	88.06	80.86	84.31
TEST	STACC$_{wp}$ (F)	250	0.16	98.65	86.81	**84.27**	**85.52**
TEST	STACC$_{wp}$ (P)	250	0.17	98.65	**91.47**	79.16	84.87

Table 3: Results for DE-EN

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.15	99.46	92.44	89.45	90.92
SAMPLE	STACC$_{wp}$ (F)	250	0.14	99.46	92.26	**91.07**	**91.66**
SAMPLE	STACC$_{wp}$ (P)	250	0.15	99.46	**95.33**	87.84	91.43
TRAIN	STACC$_w$ (F)	250	0.16	96.84	78.43	**79.23**	78.83
TRAIN	STACC$_{wp}$ (F)	250	0.16	96.84	83.93	77.58	**80.63**
TRAIN	STACC$_{wp}$ (P)	250	0.17	96.84	**87.81**	71.69	78.93
TEST	STACC$_w$ (F)	250	0.16	96.87	80.27	**78.89**	79.58
TEST	STACC$_{wp}$ (F)	250	0.16	96.87	86.01	77.39	**81.47**
TEST	STACC$_{wp}$ (P)	250	0.17	96.87	**90.62**	71.88	80.17

Table 4: Results for FR-EN

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.12	100.00	91.27	**89.49**	**90.37**
SAMPLE	STACC$_{wp}$ (F)	250	0.12	100.00	95.79	70.82	81.43
SAMPLE	STACC$_{wp}$ (P)	250	0.13	100.00	**98.82**	65.37	78.69
TRAIN	STACC$_w$ (F)	250	0.14	97.05	78.27	**74.72**	**76.45**
TRAIN	STACC$_{wp}$ (F)	250	0.13	97.05	79.26	70.62	74.69
TRAIN	STACC$_{wp}$ (P)	250	0.14	97.05	**86.23**	64.61	73.87
TEST	STACC$_w$ (F)	250	0.14	97.15	80.37	**74.74**	**77.45**
TEST	STACC$_{wp}$ (F)	250	0.13	97.15	79.82	70.73	75.00
TEST	STACC$_{wp}$ (P)	250	0.14	97.15	**88.64**	64.19	74.46

Table 5: Results for ZH-EN

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.15	97.81	95.42	86.98	91.01
SAMPLE	STACC$_{wp}$ (F)	250	0.14	97.81	96.46	**88.37**	**92.24**
SAMPLE	STACC$_{wp}$ (P)	250	0.15	97.81	**97.94**	84.16	90.53
TRAIN	STACC$_w$ (F)	250	0.16	96.64	77.69	**79.77**	78.72
TRAIN	STACC$_{wp}$ (F)	250	0.16	96.64	84.87	77.26	**80.89**
TRAIN	STACC$_{wp}$ (P)	250	0.17	96.64	**88.05**	71.02	78.63
TEST	STACC$_w$ (F)	250	0.16	96.81	79.44	**79.34**	79.39
TEST	STACC$_{wp}$ (F)	250	0.16	96.81	86.31	76.83	**81.30**
TEST	STACC$_{wp}$ (P)	250	0.17	96.81	**89.91**	70.67	79.14

Table 6: Results for RU-EN

Regarding STACC hyper-parameters, k-best lexical translations were limited to a maximum of 4 and the minimal prefix length for longest common prefix matching was set to 4. Lucene indexing was based on words with length of 4 or more characters, and a maximum of 100 candidates were retrieved for each source sentence. For each language pair, English was arbitrarily set to be the target language. For the weighting function, α was set to 250 across the board, as it was established in (Azpeitia et al., 2017) to be an optimal setting overall.

We prepared three variants for the task and applied all three on all four language pairs. The first variant is STACC$_w$, which we take to be our baseline, with an alignment threshold set to maximise the f1-measure on the training set. The second variant is the STACC$_{wp}$ method described in Section 2.2., with an alignment threshold also set to maximise the f1-measure.[4] Finally, we submitted a third variant, based on STACC$_{wp}$ but with a higher alignment threshold meant to maximise precision, as in practical cases it may be optimal to create smaller but more accurate bitexts from comparable corpora.[5]

3.2. Results

Results on all datasets are shown in Tables 3, 4, 5 and 6, along with the hyper-parameters used for each dataset and the percentage of correct candidates retrieved via Lucene indexing and retrieval. Our system competed with other systems in FR-EN and ZH-EN, with our variants reaching the highest scores on all three metrics;[6] for DE-EN and RU-EN, there were no other competing systems.

Since not all gold parallel sentences are known for this task, the results shown here are minimum values, i.e. there may be actually correct alignments identified as false positives.[7] They are nonetheless satisfactory across the board, with f1 scores above 80% on the test sets for French-English, German-English and Russian-English, and precision above 90% for the same three pairs. Although slightly lower, Chinese-English results are close to the 80% mark for the f1 measure and at 89% in terms of precision, improving over the best results obtained for this language pair on the similar BUCC 2017 task by more than 30 f1-measure points and over 40 points in terms of precision.

Our submission this year confirmed the efficiency of the generic STACC approach on Russian and Chinese, two languages that exhibit marked differences with the other two language pairs. Thus, these results further validate the claim of portability for our approach.

As for the STACC$_{wp}$ variant we introduced this year, it provided significant improvements over the already robust STACC$_w$ method, with gains of up to two points in f1-measure. Only for Chinese-English were the results lower than with STACC$_w$, a not completely unexpected result given the peculiarities of Chinese in terms of named entities as well. The results obtained with this variant confirm the specific importance of named entities for the alignment of comparable sentences, and the need to give them special prominence when computing alignment scores.

Overall, we view the high scores obtained on all metrics in all language pairs as satisfactory, especially considering the large test sets used in the shared task.

4. Conclusion

We described our submission to the 2018 BUCC shared task on the extraction of parallel sentences from comparable corpora. Our contribution for this year was twofold. We first applied our STACC$_w$ approach, which is based on weighted set-theoretic operations on expanded lexical sets, to all four language pairs proposed for the task. Additionally, we introduce a variant which further penalizes mismatches in terms of named entities, improving over the already strong weighted variant baseline in most cases. This variant is seamlessly integrated into STACC via a set-based penalty computed over surface-defined named entities.

Our approach reached the highest results on all metrics and in all scenarios, with scores over 80% in terms of f1-measure and 90% in precision. The results from our participation in the BUCC 2018 shared task thus demonstrate the efficiency of the STACC approach in terms of quality of extracted alignments and portability across language pairs.

[4] Note that, for the German-English pair, the penalty was computed with named entity sets that only comprised numbers, as including capitalised words would have also captured common nouns that are not part of the translation tables because of lexical coverage gaps in the corpora.

[5] In the tables, we add an (F) next to each variant name if the alignment threshold was selected to optimise the f1-measure, and a (P) if set for precision.

[6] This claim is based on the results provided by the organisers as of this writing, which include the maximum scores obtained for the task in terms of the three metrics.

[7] See (Zweigenbaum et al., 2017) for an analysis of the improved results obtained via a sample-based complementary human evaluation.

5. Acknowledgements

This work was partially supported by the Spanish Ministry of Economy and Competitiveness and the Department of Economic Development and Competitiveness of the Basque Government via projects AdapTA (RTC-2015-3627-7) and TRADIN (IG-2015/0000347). We would like to thank MondragonLingua Translation & Communication as coordinator of these projects for their support.

6. References

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 16–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Azpeitia, A., Etchegoyhen, T., and Martinez García, E. (2017). Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *In Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Etchegoyhen, T. and Azpeitia, A. (2016). Set-Theoretic Alignment for Comparable Corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2009–2018.

Etchegoyhen, T., Azpeitia, A., and Pérez, N. (2016). Exploiting a Large Strongly Comparable Corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Fung, P. and Cheung, P. (2004). Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and E.M. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 57–63.

Grégoire, F. and Langlais, P. (2017). BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 46–50. Association for Computational Linguistics.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241 – 272.

Munteanu, D. S. and Marcu, D. (2002). Processing Comparable Corpora With Bilingual Suffix Trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 289–295. Association for Computational Linguistics.

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

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Sarikaya, R., Maskey, S., Zhang, R., Jan, E.-E., Wang, D., Ramabhadran, B., and Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In *Proceedings of InterSpeech*, pages 432–435.

Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P. (2016). *Building and Using Comparable Corpora*. Springer Publishing Company, Incorporated, 1st edition.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefănescu, D., Ion, R., and Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 137–144.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 2214–2218.

Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 745–748. IEEE.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67. Association for Computational Linguistics.

UM-pAligner: Neural Network-Based Parallel Sentence Identification Model

Chongman Leong, Derek F. Wong, Lidia S. Chao
NLP^{2}CT Lab, Department of Computer and Information Science
University of Macau, Macau SAR, China
nlp2ct.chongman@gmail.com, {derekfw,lidiasc}@umac.mo

Abstract

This paper describes the UM-pAligner for the parallel sentence identification shared task of BUCC 2018. The proposed UM-pAligner system consists of two main components, alignment candidate identification and classification models. For the identification model, we propose using an orthogonal denoising autoencoder to transform the embedding features of parallel sentences into shared and private latent spaces, with an objective to better capture the translation correspondences of parallel sentences. In classification, a maximum entropy classifier is employed to determine and select the parallel sentences from the candidate list. On Chinese-English track data, the UM-pAligner achieves a retrieval rate up to 83.65% at the identification phase when n-best is set to 80. The classification model obtains an F1-score of 73.47%, 58.54% and 56.00% respectively on `sample`, `training` and `test` data.

Keywords: parallel sentence classification, orthogonal denoising autoencoder, neural model, maximum entropy

1. Introduction

With a huge success of neural machine translation (NMT) (Bahdanau et al., 2014; Lample et al., 2017; Artetxe et al., 2017; Yang et al., 2017), it requires a reasonable large bilingual (or multilingual) parallel corpus for achieving good translation quality (Koehn and Knowles, 2017). There is also a huge demand of parallel corpora in multilingual natural language processing (NLP) applications, in particular for low-resource language pair (Lu et al., 2010). Automatic construction of parallel corpora has been an important and active research direction in the NLP community (Tian et al., 2014; Chao et al., 2018; Neves, 2017). Comparable corpora is a pair of corpora contain topic aligned documents in two different languages. (Smith et al., 2010). The BUCC2018 shared task is to identify the parallel sentences, which are translations of each other, given a set of comparable corpora in two or more languages. In the shared task, we need to overcome the following issues:

1. Dealing with a large number of candidates: different from the conventional way to extract parallel sentences from comparable documents where the parallel documents are given, in the BUCC shared task, one document holds all the sentences, up to 80,000 sentences in the Chinese-English track. The number of possible combinations is around 6.4 billion, but only 1,900 of them are the gold parallel sentences. To be more manageable, we need a better way to filter out the sentence pairs which are not the strict translations of each other.

2. Identification of plausible candidates: in comparable corpus, the sentences are not strictly parallel, but are loose translations of each other. Thus, the second challenge is how to measure the similarity of sentence in terms of their deep semantic meaning instead of the shallow lexical information. Since those sentences are not literally translated each other.

Corresponding author: Derek F. Wong

In the past years, many approaches have been developed to automatically acquire the parallel sentences from comparable corpora. Munteanu and Marcu (2005) aligned articles by considering the publication date of the documents, and employed a maximum entropy classifier for identifying the parallel sentences from the aligned articles. Various parallel sentence alignment models and strategies have also been applied to induce parallel sentences from the Wikipedia (Adafre and de Rijke, 2006; Yasuda and Sumita, 2008; Smith et al., 2010; Barrón-Cedeño et al., 2015). These systems require the inter-language links to align the multilingual documents in the first step, with the objective to constrain the search complexity by throwing away all possible combinations of sentences across documents. However, these approaches are not suitable for this shared task, since it highly relies on the meta-data of a document. Unfortunately, such meta-data is not officially provided. Thus, one of the challenges of the shared task is to efficiently find out the possible aligned sentences from the large number of sentences. Recent works also try to model the parallel sentences through the use of deep neural networks (DNNs) approach. Chu et al. (2016) exploited neural network features that acquired from a trained NMT system in a classification model. However, the method relies on an external NMT system and the performance of the classifier highly depends on the quality of the NMT model. Grégoire and Langlais (2017) proposed using a recurrent neural network (RNN) for the parallel sentence identification task. Their model takes the advantage of semantic information of a sentence pair that learned by the RNN. However, it does not consider the word alignment and lexical information which have been proven to be very useful (Munteanu and Marcu, 2005; Zamani et al., 2016). In this paper, we describe the UM-pAligner, a parallel sentence alignment system, that we submitted to the BUCC 2018 shared task. The system consists of two main components, the alignment candidate identification and parallel sentence classification models. For the identification, the main task is to filter out the sentence pairs

which are semantically irrelevant by exploiting their deep semantic features. While the classification model takes the features of word alignment and translation probabilities into consideration to further assess the parallelism of the candidates.

2. Proposed Method

2.1. Overview

To solve the problems mentioned in the previous section, the proposed approach consists of two phases: 1) alignment candidate identification that aims to largely filter out the implausible alignment candidates from the comparable corpus; and 2) alignment classification which further evaluates the parallelism of the alignment candidates using additional word-level alignment and lexical features which are more reliable and interpretable. The processing flow of the approach is depicted in Figure 1.

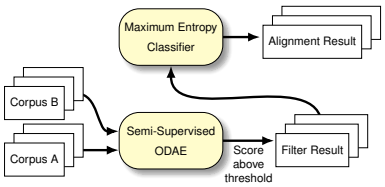

Figure 1: Architecture of UM-*p*Aligner.

For filtering out the semantically irrelevant sentence pairs, we propose a semi-supervised orthogonal denoising autoencoder to detect the parallelism of a given sentence pair. The underlying principle is to transform the embedding of parallel sentences into their shared and private latent spaces that on the other words to capture their aligned and unaligned features of two sentences. The model is efficient in filtering out those of irrelevant sentence pairs and give us a reasonable number of candidates for subsequent classification. For the classification model, we employ a maximum entropy model for the classification task, where we consider the lexical features and the word alignment information of a sentence pair. In brief, the UM-*p*Aligner performs the following steps for identifying the parallel sentences from the comparable corpora:

1. All possible sentence pairs are scored by the semi-supervised orthogonal denoising autoencoder. For those candidates whose score is above a threshold are selected;

2. For those of selected candidates from the first step are scored by the maximum entropy classifier. We use another threshold to determine the final parallel sentences. During the alignment process, one source sentence is only allowed to align to a target sentence once. The candidate with the highest score is considered.

2.2. Semi-Supervised Orthogonal Denoising Autoencoder

To better capture the underlying semantic meanings of parallel sentences, we propose a novel model based on multi-view learning and orthogonal denoising autoencoder for the identification of parallel sentences from a comparable corpus. Those methods have been successfully used in many NLP applications (Zeng et al., 2013a; Wong et al., 2016). In this study, the multi-view technique is employed to treat the source and target sentences as two different interpretations of the same semantic meaning. We believe the bilingual sentence pair which represent the same text's meaning should share the same semantic space, otherwise they should exhibit very different representation. Hence, to differentiate such relationship from a vector representation point of view, we further propose the use of semi-supervised orthogonal denoising autoencoder (Ye et al., 2016) to explicitly impose this constraint by mapping the underlying sentence representation into the shared and private latent spaces. The architecture of the proposed model is illustrated in Figure 2.

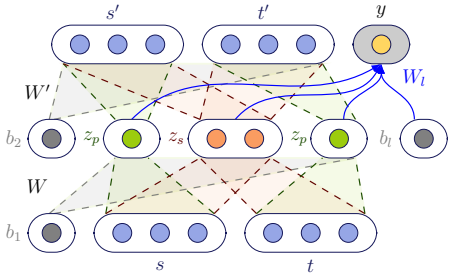

Figure 2: Architecture of the semi-supervised orthogonal denoising autoencoder. The representations of source sentence s and target sentence t are being treated as different input views. The private and shared latent spaces, z_p and z_s represent the common features shared by both sentences and the private features owned by individual sentence. The s' and t' are the reconstructed representations of the source and target sentences, while y is the prediction label of the pair of sentences s and t to see if they are translations of each other or not.

Model Description Given a concatenated representation vector $\mathrm{x} = \{x_1, \ldots, x_m, x_{m+1}, x_n\}$ of a source sentence x_s and its paired target sentence x_t with the sentence lengths of $|\mathrm{x}_s| = m$ and $|\mathrm{x}_t| = n$ respectively, an autoencoder aims to transform it to a hidden space $h = s(Wx + b)$, and the hidden representation h is subsequently transformed back to its reconstructed vector $\mathrm{x}' = g(W'h + b')$ through the activation functions $s(\cdot)$ and $g(\cdot)$ with the weight matrices W and W', and the bias b and b'. The objective is to learn the model parameters that minimizes the reconstruction error $\ell(\mathrm{x}, \mathrm{x}')$, where $\ell(\cdot)$ is a loss function to measure how good the reconstruction performs.
Orthogonal Constraint To accommodate the shared and private latent spaces in the context of multi-view learning, the autoencoder model is revised to connect only the private latent space to its original input view, and disconnect it from the other views, such that the private latent spaces are independent from each other. While the shared space is connected to all of the input views, i.e. the representation of the source and target sentences. The architecture of the

model is depicted in Figure 2. To maintain the orthogonality of the private spaces, the bias is disconnected from the private spaces (Ye et al., 2016). Formally, $I(A|B)$ is defined to denote the indices of columns of matrix A in terms of the matrix B if A is a submatrix of B. The orthogonal constraints on weights is defined as follows:

$$W_{I(z_p^{v_2}|[z_s,z_p]),I(\mathrm{x}^{v_1}|\mathrm{x})} = 0$$

$$W'_{I(\mathrm{x}^{v_1}|\mathrm{x}),I(z_p^{v_2}|[z_s,z_p])} = 0,$$

where $\mathrm{v} = \{v_1, ..., v_k\}$ denotes the different views of an input x, z_s is the shared latent space and $\mathrm{z}_p = \{z_1, ..., z_k\}$ are the private spaces.

Semi-Supervised Model The denoising autoencoder was originally proposed to enforce the autoencoder in learning robust features (Ye et al., 2016). In our case, we want the model to be able to learn the latent features which are best to distinguish if a pair of sentences are the translations of each other. To this extend, we further modify the model to guide the training towards this objective. The latent spaces are leveraged by adding a feed-forward NN layer in addition to the reconstruction layer, and defined as:

$$y = \sigma(W_l[z_s, z_p] + b_l),$$

where $\sigma(\cdot)$ is the sigmoid function, W_l and b_l are the weight matrix and the bias.

Model Training The model parameters are optimized by minimizing the loss function:

$$J = \alpha J_{rec} + (1 - \alpha)J_{label},$$

where J_{rec} and J_{label} are reconstruction and cross-entropy loss. The hyper-parameter α is used to weight the reconstruction and cross-entropy error in controlling the preference of the learned model:

$$J_{label} = \frac{1}{n} \sum [y' \log(y) + (1 - y') \log(1 - y)]$$

$$J_{rec} = \frac{1}{2n} \sum ([\mathrm{x}_s; \mathrm{x}_t] - [\mathrm{x}_{s'}; \mathrm{x}_{t'}]).$$

2.3. Maximum Entropy Classifier

Previous works have shown the effectiveness of acquiring parallel sentences using a maximum entropy model (Berger et al., 1996; Munteanu and Marcu, 2005; Wong et al., 2009; Zeng et al., 2013b). Thus, we employ it for our classification problem and define it as:

$$p(c|\mathrm{s}, \mathrm{t}) = \frac{\exp(\sum \lambda_i f_i(y,\mathrm{s},\mathrm{t}))}{Z(\mathrm{s},\mathrm{t})},$$

where $p(c|\mathrm{s}, \mathrm{t}) \in [0, 1]$ is the probability where a value close to 1.0 indicates that the paired sentences are translations of each other, $y \in (0, 1)$ is a class label representing where the sentences (s, t) are parallel or not parallel, $Z(\mathrm{s}, \mathrm{t})$ is the normalization factor, f_i are the feature functions, and λ_i are the feature weights to be learned. The features we considered in this task include the length-based features (Gale and Church, 1993), alignment-based features (Munteanu and Marcu, 2005; Dyer et al., 2013) and the anchor text (Patry and Langlais, 2011).

3. Experiments

3.1. Pre-train of Sentence & Word Embeddings

In training the proposed model, the embeddings of words and sentences can either be trained from scratch jointly with the model or pre-trained prior to the training of the model. To be more manageable, we prefer constructing the word and sentence embeddings separately. The word embeddings are constructed using the Global Vectors (Glove) (Pennington et al., 2014), and the sentence embeddings are trained with the Smooth Inverse Frequency scheme (SIF) (Arora et al., 2017). The embeddings are trained on the Chinese-English parallel corpora of `casict2011`, `casict2015`, `casia2015`, `datum2015`, and `neu17` of the CWMT datasets (Wong and Xiong, 2017). [1] There are 8 million parallel sentences in total, covering a wide range of different genres such as newswire, law, technical documents and on-line publications (web-pages).

3.2. Datasets

Preprocessing First, we observed that the Chinese dataset is a mixture of Simplified and Traditional Chinese texts. To unify it, we convert all the Traditional Chinese texts into the Simplified ones (Wong et al., 2009), to ensure that all the texts are in the same encoding scheme. Secondly, for those of the official training data, the sentences are translated using an on-line translation system. Thus, we have collected 147,930 "parallel" sentences of the training data of zh-en track and the additional 500,000 parallel sentences of `neu17` from the CWMT (Wong and Xiong, 2017). The constructed parallel data are then used to train the orthogonal denoising autoencoder and the maximum entropy classifier. Thirdly, for those of Chinese data, texts are segmented into words, as known as Chinese word segmentation (Wang et al., 2012; Zeng et al., 2013a; Zeng et al., 2013b).

Negative Samples In training the autoencoder and the maximum entropy classifier, we need false training instances. In this work, for each of the positive samples, we randomly produce 5 negative samples. In total, the data used for training the models consists of 647,930 positive and 3,239,650 negative samples.

3.3. Experimental Results

Table 1 presents the statistical information of the used `sample` and `training` data of the zh-en track provided by the BUCC2018 organizer for evaluation.

Dataset	Source	Target	Gold
`Sample`	8,624	13,589	257
`Training`	94,637	88,860	1,899

Table 1: Statistical information of the `sample` and `training` data.

Model Setting The proposed autoencoder is implemented using Tensorflow (Abadi et al., 2016). The dimension of the

[1] The parallel corpora are available at: `http://nlp.nju.edu.cn/cwmt-wmt/`

sentence embedding is set to 300. We use 2048 nodes for the hidden state, in which 1024 of them are for the shared latent space and the private latent space for each view is set to 512 nodes. For the training, the model is optimized by the Adam optimizer (Kingma and Ba, 2014) with a batch size of 2048. We train the model for 200 epochs in our experiments. The model is evaluated using the method provided by the organizer, where the precision (P), recall (R) and F_1-score (F_1) are calculated as:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F_1 = \frac{2\times P\times R}{P+R}$$

Alignment Candidate Identification Table2 reports the identification results on the `sample` and `training` datasets respectively, by varying the selection n-best. However, during the selection, we also apply the constraints of length ratio and a threshold of model scores ($t_{ode} = 0.99$) strictly to filter out those of loose translations of each other. We can see that around 80% gold pairs are retrieved when we consider the 60-best.

Dataset	n-best	Recall (%)	# Candidates
`Sample`	1	36.18%	7,945
	5	69.26%	35,819
	10	78.21%	64,624
	20	82.87%	108,549
	40	83.65%	165,050
	60	83.65%	199,522
	80	83.65%	221,784
	100	84.04%	237,124
	∞	84.43%	282,641
`Training`	1	12.74%	92,726
	5	39.02%	451,385
	10	52.39%	879,390
	20	65.92%	1,682,456
	40	75.40%	3,115,220
	60	79.98%	4,357,690
	80	82.46%	5,445,018
	100	83.51%	6,402,092
	∞	86.25%	17,357,720

Table 2: Identification results on `sample` and `training` dataset, constrained by a model score threshold, $t_{ode} = 0.99$

Parallel Sentence Classification After the first step, we now have a candidate list of manageable size. In which, we further access the parallelism of the paired sentences using the maximum entropy classifier. We use the model scores to determine the final candidates of parallel sentences. In defining the threshold, we have conducted two sets of experiments on `training` dataset. In the first experiment, we set the model threshold to $t_{me} = 0.999$, and the model obtains 47.41%, 63.30% and 54.21% of precision, recall and F_1-score respectively. When we vary the model threshold to $t_{me} = 0.9999$, the classifier obtains a better F_1-score of 58.54%. The results are reported in Table 3. Hence, we use the model threshold of $t_{me} = 0.9999$ for our subsequent experiments.

Dataset	Threshold	Precision	Recall	F1 Score
`Sample`	0.9999	74.20%	72.76%	73.47%
`Training`	0.9999	67.00%	51.97%	58.54%

Table 3: Classification performance of the maximum entropy classifier on the candidates.

4. Shared Task Result

In the `test` dataset, there are 91,824 Chinese sentences and 90,037 English sentences, among which there are only 1,896 gold parallel sentences. We adjust selection criterion, n-best, in the phase of candidate identification. For the classification, we apply the model threshold of $t = 0.9999$ to determine the final results. The results given by the identification and classification models are reported in Table 4 and 5 respectively.

Dataset	n	Retrieved(%)	Selected pairs
`Test`	60	79.06%	4,253,884
	80	81.69%	5,333,848

Table 4: Identification performance on BUCC2018 `test` set, with decision threshold $t = 0.99$.

Dataset	n	Precision	Recall	F1 Score
`Test`	60	73%	45%	55%
	80	72%	46%	56%

Table 5: Classification performance on BUCC2018 `test` set, with classification threshold $t = 0.9999$.

5. Conclusion

In this shared task, we have proposed a parallel sentence identification and classification model, UM-pAlignerṪhe system consists of two main components: 1) we propose the use of semi-supervised orthogonal denoising autoencoder to determine if a source and target sentences are parallelism or not, by considering their deep semantic meaning; and 2) we construct a maximum entropy based classifier using the symbolic features of texts, as complementary to the neural network based autoencoder, to further assess if the sentences are the translations of each other. The model achieves the F_1-score of 73.47%, 58.54% and 56% on the `sample`, `training` and `test` dataset respectively.

6. Acknowledgements

We thank the reviewers for their valuable and insightful comments and suggestions. This work was supported in part by the National Natural Science Foundation of China (Grant No. 61672555), a Multiyear Research Grant from the University of Macau (Grant Nos. MYRG2017-00087-FST, MYRG2015-00175-FST and MYRG2015-00188-FST) and the Joint Project of Macao Science and Technology Development Fund and National Natural Science Foundation of China (Grant No. 045/2017/AFJ).

7. References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

Adafre, S. F. and de Rijke, M. (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.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Barrón-Cedeño, A., España-Bonet, C., Boldoba, J., and Màrquez, L. (2015). A factory of comparable corpora from wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, BUCC@ACL/IJCNLP 2015, Beijing, China, July 30, 2015*, pages 3–13.

Berger, A. L., Pietra, S. D., and Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Chao, L. S., Wong, D. F., and Ao, C. H. (2018). Umpcorpus: A large portuguese-chinese parallel corpus. In *Proceedings of the First Workshop on the Belt and Road Language Resources and Evaluation (B&R LRE)*.

Chu, C., Dabre, R., and Kurohashi, S. (2016). Parallel sentence extraction from comparable corpora with neural network features. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Grégoire, F. and Langlais, P. (2017). A deep neural network approach to parallel sentence extraction. *CoRR*, abs/1709.09783.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39.

Lample, G., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Lu, B., Tsou, B. K., Jiang, T., Kwong, O. Y., and Zhu, J. (2010). Mining large-scale parallel corpora from multilingual patents: An english-chinese example and its application to smt.

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

Neves, M. L. (2017). A parallel collection of clinical trials in portuguese and english. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora, BUCC@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 36–40.

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, BUCC@ACL 2011, Portland, OR, USA, June 24, 2011*, pages 87–95.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 403–411.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., and Yi, L. (2014). Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 1837–1842.

Wang, L., Wong, D. F., Chao, L. S., and Xing, J. (2012). Crfs-based chinese word segmentation for micro-blog with small-scale data. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 51–57, Tianjin, China, 20-21 December, 2012.

Derek F. Wong et al., editors. (2017). *Machine Translation - 13th China Workshop, CWMT 2017, Dalian, China, September 27-29, 2017, Revised Selected Papers*, vol-

ume 787 of *Communications in Computer and Information Science*. Springer.

Wong, F., Chao, S., Hao, C. C., and Leong, K. S. (2009). A maximum entropy (me) based translation model for chinese characters conversion. *Advances in Computational Linguistics, Research in Computer Science*, 41:267–276. 10th Conference on Intelligent Text Processing and Computational Linguistics - CICLing, Mexico City. http://www2.dc.ufscar.br/ helenacaseli/.

Wong, D. F., Lu, Y., and Chao, L. S. (2016). Bilingual recursive neural network based data selection for statistical machine translation. *Knowl.-Based Syst.*, 108:15–24.

Yang, B., Wong, D. F., Xiao, T., Chao, L. S., and Zhu, J. (2017). Towards bidirectional hierarchical representations for attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1432–1441.

Yasuda, K. and Sumita, E. (2008). Method for building sentence-aligned corpus from wikipedia. In *2008 AAAI Workshop on Wikipedia and Artificial Intelligence (WikiAI08)*.

Ye, T., Wang, T., McGuinness, K., Guo, Y., and Gurrin, C. (2016). Learning multiple views with orthogonal denoising autoencoders. In *MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I*, pages 313–324.

Zamani, H., Faili, H., and Shakery, A. (2016). Sentence alignment using local and global information. *Computer Speech & Language*, 39:88–107.

Zeng, X., Wong, D. F., Chao, L. S., and Trancoso, I. (2013a). Co-regularizing character-based and word-based models for semi-supervised chinese word segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 171–176.

Zeng, X., Wong, D. F., Chao, L. S., and Trancoso, I. (2013b). Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 770–779.