doi:10.1017/S1351324916000115

Recent advances in machine translation using comparable corpora

REINHARD RAPP[1], SERGE SHAROFF[2] and PIERRE ZWEIGENBAUM[3]

[1]*University of Mainz*
e-mail: reinhardrapp@gmx.de
[2]*University of Leeds*
e-mail: s.sharoff@leeds.ac.uk
[3]*LIMSI, CNRS, Université Paris-Saclay*
e-mail: pz@limsi.fr

(*Received 20 November 2015; revised 15 January 2016; accepted 17 March 2016*)

Abstract

This paper highlights some of the recent developments in the field of machine translation using comparable corpora. We start by updating previous definitions of comparable corpora and then look at bilingual versions of continuous vector space models. Recently, neural networks have been used to obtain latent context representations with only few dimensions which are often called word embeddings. These promising new techniques cannot only be applied to parallel but also to comparable corpora. Subsequent sections of the paper discuss work specifically targeting at machine translation using comparable corpora, as well as work dealing with the extraction of parallel segments from comparable corpora. Finally, we give an overview on the design and the results of a recent shared task on measuring document comparability across languages.

1 What are comparable corpora?

In the announcements and other documentations of the annual editions of the workshop series on *Building and Using Comparable Corpora*, the term *Comparable Corpus* has often been defined as follows (Rapp, Zweigenbaum and Sharoff 2010): ‘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’.

What did the workshop organizers mean by this? Given two text corpora, if we wish so we can always compare them, no matter of their form and content. For the comparison, we may define the dimensions we are interested in. One obvious dimension could be the *language* or *dialect*. But lots of other possible dimensions could be thought of, among them, for example, *topic*, *genre*, *content*, *discourse structure*, *purpose*, *origin of author*, *sex of author*, *employer of author*, *target audience*, *time of writing*, *location of writing*, *length of text*, *text difficulty*, *text type* (e.g.

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

original, summary, or translation), *text style*, *vocabulary*, *collocations*, and *modality* (e.g. written, spoken, sign language). If we try to characterize pairs of texts in terms of such dimensions, we could, for example, say that the so-called *parallel corpora* agree along almost all of the dimensions except for *language*, and that the most commonly used types of comparable corpora agree at least along the dimensions *topic*, *genre*, and *modality*, but not on *language*.

But many other such dimensions could be suggested, and it is the responsibility of a researcher to identify and define them, and to decide which ones are of interest for a particular task. For example, research on machine translation (MT) might be mostly interested in the dimensions *language* and *content*, while research on author identification might focus on style, *vocabulary*, *collocations*, *origin of author*, *sex of author*, *time of writing*, and *location of writing*.

Any two or more corpora can be called comparable corpora if their relationship plays a role in any way. This means that the term *comparable corpora* does not primarily reflect particular properties of the respective corpora, but instead properties of the work which is conducted using them (i.e. it must be work in some way relating the corpora to each other). It is probably for such reasons why (Maia 2003) concluded that 'to a certain degree, comparability is in the eye of the beholder'. For a discussion of some earlier definitions of comparable corpora, see Tang, Wang and Chen (2015).

If we wish to quantify the comparability of two corpora, we must keep in mind that the result of the comparison depends on the choice of dimensions which are taken into account. For example, if we have two texts in different languages but on the same topic, then a comparison along the dimension *language* will result in a low similarity score, but another comparison along the dimension *topic* will result in a high similarity score. And a comparison taking into account both dimensions should result in a similarity score somewhere in between. That is, there is nothing like a single score describing the comparability of two corpora. Instead, each task in mind is likely to require a specifically made up procedure for measuring corpus comparability.

Whereas our definitions of *comparable corpora* and *corpus comparability* are very general, previous authors have provided more practical definitions which were geared toward particular scenarios. For example, Sharoff, Rapp and Zweigenbaum (2013a) define the degrees of comparability in the following way:

(1) Parallel texts: texts which are more or less true and accurate translations.
(2) Strongly comparable texts: heavily edited translations or independent, but closely related texts reporting the same event or describing the same subject.
(3) Weakly comparable texts: texts in the same narrow subject domain and genre, but describing different events, or texts within the same broader domain and genre, but varying in subdomains and specific genres.
(4) Unrelated texts: e.g. random snapshots of the web which, however, can still be used for comparative linguistic purposes.

Another definition is provided by Wu and Fung (2005):[1]

[1] Quoting from `http://www.cs.ust.hk/~dekai/library/WU_Dekai/nonparallel.html` as this page is meant to synthesize and systematize this and two earlier contributions.

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

(1) Parallel corpus: sentence-aligned corpus containing bilingual translation of the same document.
(2) Noisy parallel corpora: contains non-aligned sentences that are nevertheless mostly bilingual translations of the same document.
(3) Comparable corpus: contains non-sentence-aligned, non-translated bilingual documents that are topic-aligned.
(4) Quasi-comparable corpus: contains non-aligned, and non-translated bilingual documents that could either be on the same topic (in-topic) or not (off-topic).

These two definitions have in common that they put an emphasis on the dimensions required for identifying translations and paraphrases, but neglecting many other dimensions. This is likely to be a very sensible approach for translation-related purposes, but may be completely unsuitable for others (such as author identification). But both definitions fit well into the more general framework sketched above.

To give a few examples of text collections as typically used in comparable corpus research, let us briefly characterize *Wikipedia*, the *International Corpus of English*, the *MLCC Corpus*, and the *WaCky Corpora*.

The articles of the Wikipedia editions in various languages[2] can occasionally be translations of each other (as e.g. a translation can be a starting point for a newly created article), but more typically evolve more or less independently of each other, and are geared toward readerships speaking the respective languages (which to some extend often correlate with regions and nationalities). A specific property of Wikipedia are the so-called interlanguage links, which are author-created connections between articles in different languages, but relating to the same headword (or to translations of the same headword). These interlanguage links make it easy to align Wikipedia editions at the document level.

The International Corpus of English[3] consists of one million word samples in each of many varieties of English around the globe, each following the same collection principles. For example, texts from specific genres had to be collected in particular quantities. Moreover, the original idea had been to gather all texts in the same year.

The MLCC Corpus[4] is an early example of a comparable newspaper corpus. It comprises the contents of a number of financial newspapers, namely The Financial Times (English), Het Financieele Dagblad (Dutch), Le Monde (French), Handelsblatt (German), Il Sole 24 Ore (Italian), and Expansion (Spanish). Although the authors of the different newspapers can in principle be seen as independent of each other, their articles of course often relate to the same world news as e.g. distributed by press agencies. Thus, although it is not as easy as with Wikipedia articles, a document alignment would in many cases be possible by utilizing the publication dates of articles. This should result in a number of alignment candidates, which can be verified by looking at matches of named entities or keywords. Hereby, the matching of keywords across languages requires a dictionary of keywords.

[2] https://www.wikipedia.org/
[3] http://www.ucl.ac.uk/english-usage/projects/ice.htm
[4] http://catalog.elra.info/product_info.php?products_id=764

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

The WaCky corpora[5] are very large text collections in English, French, German, and Italian as opportunistically extracted from the World Wide Web, thus reflecting a very diverse range of documents. This makes them interesting for studies where the number of language phenomena is not supposed to be limited artificially. On the other hand, an alignment at the document level is more difficult for the WaCky corpora as they are very heterogenous and provide few specific alignment clues.

It should be mentioned that our definitions of comparable corpora include parallel corpora as a particular subtype. So popular parallel corpora such as the Europarl corpus could also be listed here. However, as work on parallel corpora has already received an enormous amount of attention elsewhere, we do not focus on them here.

2 Why use comparable corpora for machine translation?

Statistical MT based on parallel corpora has been very successful. For example, the major search engines' translation systems, which are used by millions of people every day, are primarily using this approach, and it has been possible to come up with new language pairs in a fraction of the time that would be required when using more traditional rule-based methods.

In contrast, research on MT using comparable corpora is still at an earlier stage. The subtype of non-parallel corpora most promising for MT are probably monolingual corpora covering roughly the same subject area in different languages but without being exact translations of each other. They are of interest because, despite its tremendous success, the use of parallel corpora in MT has a number of drawbacks:

- It has been shown that translated language is somewhat different from original language, for example Beigman and Flor (2013) showed that 'associative texture' is lost in translation.
- Parallel corpora will always be a far scarcer resource than comparable corpora because only a fraction of all original publications are translated. This is a severe drawback for a number of reasons:
    (1) Among the about 7,000 world languages, of which 600 have a written form, the vast majority are of the 'low resource' type.
    (2) The number of possible language pairs increases with the square of the number of languages. When using parallel corpora, one bitext is needed for each language pair. When using comparable corpora, one monolingual corpus per language suffices.
    (3) For improved translation quality, translation systems specialized on particular genres and domains are desirable. But it is far more difficult to acquire appropriate parallel rather than comparable training corpora.
    (4) As language evolves over time, the training corpora should be updated on a regular basis. Again, this is more difficult in the parallel case.

[5] http://wacky.sslmit.unibo.it/doku.php

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

For such reasons, it would be a big step forward if it were possible to base statistical MT on comparable rather than on parallel corpora: The acquisition of training data would be far easier, and the unnatural ‘translation bias’ (e.g. source language shining through) within the training data could be avoided.

But is there any evidence that this is possible? Motivation for using comparable corpora in MT research comes from a cognitive perspective: Experience tells that persons who have learned a second language completely independently from their mother tongue can nevertheless translate between the languages. That is, human performance shows that there must be a way to bridge the gap between languages which does not rely on parallel data. Using parallel data for MT is of course a nice shortcut. But avoiding this shortcut by doing MT based on comparable corpora may well be a key to a better understanding of human translation, and to better MT quality.

Work on comparable corpora in the context of MT has been ongoing for two decades. It has turned out that this is a very hard problem to solve, but as it can be considered to be among the grand challenges in multilingual NLP, interest has steadily increased. Apart from the increase in publications, this can be seen from the considerable number of research projects (such as ACCURAT,[6] TTC,[7] and HyghTra[8]) which are fully or partially devoted to MT using comparable corpora. Given also the success of the workshop series on ‘Building and Using Comparable Corpora’ (BUCC), which is now in its ninth year, and following the publication of a related book (Sharoff *et al.* 2013b), the purpose of the current special issue is to collect and make available some of the most advanced work in the field, thus providing insights on the state of the art.

As of course the articles in this special issue can only represent a small fraction of the ongoing work, in the following subsections we also try to highlight some other interesting work. We begin with work describing full MT systems based on non-parallel corpora, and then describe methods for the extraction of parallel segments from comparable corpora. We continue with an innovative topic, namely the induction of continuous vector spaces from multilingual corpora using artificial neural networks. Finally, we describe the setup and results of a recently conducted shared task where the aim was to measure document comparability.

3 Some recent work on MT based on comparable corpora

In recent years, there has been a lot of work related to MT using comparable corpora. Hereby, the focus was typically on three subtopics:

- Development of end-to-end MT systems based on comparable corpora.
- Extraction of parallel segments from comparable corpora for the purpose of providing training material for standard statistical MT systems.
- Extraction of bilingual lexicons from comparable corpora.

[6] http://www.accurat-project.eu/
[7] http://www.ttc-project.eu/
[8] http://www.hyghtra.eu/

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

As the topic of bilingual lexicon extraction has already been covered previously, let us point to the respective paper (Sharoff *et al.* 2013a) and to an online survey of recent publications.[9] The other two topics we describe in the next two subsections, though not comprehensively due to space constraints.

3.1 End-to-end systems

In his well-known memorandum, Warren Weaver (Weaver 1955) had suggested to look at cryptographic methods for dealing with MT. However, this could not be put into practice until more than half a century later. In their pioneering works, Ravi and Knight (2008; 2011) consider MT as a decipherment task, treating a translated text as a cipher of the original text. To put it simply, the aim is to find a way for constructing bilingual vocabulary lists which, when used to replace the words of the translated text, consistently yield readable text of the source language. Although this word substitution decipherment is already demanding due to the large vocabulary sizes of natural languages, extending it to full MT has to also take into account word ambiguity, reordering of words and phrases, and the insertion or deletion of words. The authors propose two methods for doing so: One based on the EM-algorithm, the other based on a Bayesian approach. For two Spanish/English test corpora, one consisting of temporal expressions, the other of movie subtitles, they show that even without parallel training data their decipherment approach achieves accuracies comparable to systems trained on parallel data.

In recent work (Dou *et al.* 2015), further improvements could be achieved by combining the decipherment approach with the standard context vector approach as proposed by Rapp (1995). This is done using a joint inference process. The respective software, which functions as a kind of GIZA for non-parallel data, has been released to facilitate research by others.

Similarly, Nuhn, Schamper and Ney (2015) present a decipherment toolkit. It contains a tool for the decipherment of deterministic cyphers, and another tool for EM decipherment of probabilistic substitution ciphers and simple MT tasks. The toolkit builds on previous work such as Nuhn and Ney (2014).

The work on MT conducted at Google by Mikolov, Le and Sutskever (2013b) received a lot of attention. It uses Mikolov's neural network-based skip-gram and continuous-bag-of-words models to learn distributional vectors (word embeddings). The paper shows how to identify word translations from comparable corpora by using linear transformations of the source and the target language word vector spaces. However, in contrast to the decipherment-based approaches described above, this approach pre-supposes large numbers of translated pairs as extracted from parallel data to train the linear transformations.

In his MSc thesis, Ramtin Mehdizadeh Seraj (2015) tries to improve standard phrase-based MT by providing information on phrases which are missing in the parallel data. He does so by looking at paraphrases. In particular, he tries to replace unseen phrases by paraphrases which can be found in the parallel corpus. Then, it

[9] http://www.statmt.org/survey/Topic/DictionariesFromComparableCorpora

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

is assumed that the translation of the paraphrase can also serve as a translation for the unseen phrase. For paraphrase identification, two methods are considered: One is based on distributional profiles as taken from monolingual corpora. Here, like in bilingual lexicon extraction, it is assumed that phrases with similar meanings should co-occur with similar context words. The other is based on bilingual pivoting and requires parallel corpora. The underlying assumption is that source language phrases translating to the same target language phrase are likely to be paraphrases. Note that this is true for any target language, so if for a particular corpus translations into many languages are available, then the findings from all these translations can be combined. The author shows that by using paraphrases based on bilingual pivoting the BLEU score of an SMT system could be improved by 1.79 percent points.

Avneesh Saluja and his co-authors (2014) start from the observation that in standard SMT systems translation candidates for words and phrases, are derived from parallel texts, and only the selection among these (as well as their order) is influenced by the language model as derived from monolingual data. To make better use of the source and target language monolingual data, they construct phrase graphs for both languages. Next, via semi-supervised graph propagation, they identify translations of phrases which do not occur in the parallel data, whereby it is assumed that similar phrases have similar translations. In effect, this is similar to identifying paraphrases of phrases whose translations are known (see above). The approach is used to enhance state-of-the-art phrase-based MT systems, resulting in improvements of between 1 and 4 BLEU points.

3.2 Mining parallel segments from comparable corpora

As parallel corpora are a very valuable resource (e.g. they are fundamental for statistical MT) but for most language pairs quite scarce, there have been attempts to extract parallel sentences or sentence fragments from comparable corpora. This could potentially offer a solution to the data acquisition bottleneck as comparable corpora tend to be far more abundant.

A pioneering role in this type of work had Dragos Stefan Munteanu and Daniel Marcu. In Munteanu and Marcu (2002), starting from a small bilingual dictionary that was derived from a parallel corpus, they use bilingual suffix trees in order to extract a parallel from a comparable corpus. The suffix trees are a technical device to efficiently compare strings of varying length. Thus, it is possible to take into account the full literal context of a word. Roughly speaking, given a sequence of words *abc* in the source language, and a sequence *xyz* in the target language, when the seed dictionary indicates that x is a translation of a and z a translation of c, then this would be taken as evidence that y might well be a translation of b. This evidence would be strengthened if other matching triplet-pairs would also include b and y in the middle positions. Given a sufficient amount of such evidence, the seed dictionary can be expanded by the bilingual word pair $b - y$. This expansion of the dictionary will improve the chances to find new triplets that match in the first and the third positions, which again leads to dictionary expansion. These iterative

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

expansions indicate that what we have here is a bootstrapping approach which from iteration to iteration identifies more and more word translations as well as more and more parallel sentence fragments. A limitation of the algorithm is that it can only find word alignments that are monotonic, i.e. the system can only be applied to language pairs which are similar in word order (such as English–French but also English–Chinese).

In their later seminal paper, Munteanu and Marcu (2005) improved their method by training a maximum entropy classifier which for a given pair of sentences can reliably determine whether or not they are translations of each other. They also showed empirically that a statistical MT system can be built from scratch by starting with a small parallel corpus of only 100,000 words and by expanding it using parallel segments as extracted from pairs of the very large Gigaword-Corpora (Arabic–English and Chinese–English). The Gigaword corpora are newsticker text collections as provided by the Linguistic Data Consortium.

Whereas newsticker texts in different languages, for a given date, typically cover the same world news and thus offer a good chance to find parallel sentences, for very non-parallel corpora this chance is much slimmer. Munteanu and Marcu (2006) therefore extended their method to the detection of sub-sentential fragments using a signal processing inspired approach.

Abdul-Rauf and Schwenk (2009) describe a system for the extraction of parallel data from comparable corpora which uses a statistical MT system built from a small parallel corpus. This system is used to translate the source language side of a large comparable corpus. The resulting sentence translations are then utilized to find corresponding sentences on the target language side of the comparable corpus using information retrieval techniques and filters such as WER (Levenshtein distance) and TER (translation edit rate). WER measures the number of insertions, deletions, and substitutions which are required to transform one sentence into the other, but has the disadvantage that it does not allow for acceptable variations in word order. TER takes this into account by allowing block movements of words, thus allowing reordering of words and phrases.

Quirk *et al.* (2007) propose a generative model to extract parallel fragments from comparable corpora. For this purpose, they extend standard (IBM type) word alignment models to account for very noisy translations. While the standard models allow only for systematic deviations concerning the translations of sentences, in the case of comparable corpora much more flexibility is required as, if at all, bilingual sentence pairs extracted from comparable corpora typically show only partial overlap. The authors describe two models to deal with this problem: a conditional model of loose translations and a joint model of simultaneous generation. They show that the parallel fragments extracted in this way produce good improvements when added to the training data of an SMT system.

4 Bilingual spaces induced from parallel and comparable corpora

Parallel and comparable corpora have been used to induce representation spaces (typically vector spaces) where similar words have similar representations.

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

Mono-lingual representations have used context vectors where context size is defined as a syntactic dependency (Grefenstette 1992) or approximated with a window of words (Rapp 1995) possibly extending to a whole document (Gabrilovich and Markovitch 2007), and each cell *i* in a vector contains a co-occurrence count (or association measure) of context word *i* with the represented word. More recently, latent representations with few dimensions (also called word embeddings) obtained by training neural network predictors on monolingual corpora (e.g. Mikolov *et al.* (2013a)) have been created with similar properties. These monolingual representations have been extended to parallel (e.g. Mikolov *et al.* (2013a)) and comparable corpora (e.g. Klementiev, Titov and Bhattarai (2012a); Gouws, Bengio and Corrado (2015); Vulic and Moens (2014a); Dou *et al.* (2015)).

To obtain bilingual representations for a pair of languages (henceforth called source and target language without assuming a specific direction in processing), one needs information to map the source and target languages. This can come from a seed bilingual dictionary (Rapp 1995; Fung and McKeown 1997). This can also be obtained from aligned words in parallel corpora (Klementiev, Titov and Bhattarai 2012b; Apidianaki, Ljubešić and Fišer 2013; Zou *et al.* 2013), or simply from aligned sentences (Chandar *et al.* 2014; Gouws, Bengio and Corrado 2015), or even aligned documents (Bouamor *et al.* 2013; Vulic and Moens 2014b). To the best of our knowledge, no method so far used absolutely no hint of bilingual mapping. Haghighi *et al.* (2008) were very close to doing so but still used a small seed dictionary of hundred word pairs to bootstrap their process. Nevertheless, the general objective of many publications on comparable corpora is to induce additional word translations based on initial bilingual mappings.

A bilingual representation space supports representations of words in two languages in the same space: representations of words in these two languages can then be compared directly, for instance, to look for word translations. The most common method to obtain a bilingual representation consists in first building monolingual representation spaces independently, for instance, with context vectors, and then creating a bilingual space from them. The standard model of bilingual lexicon induction from comparable corpora uses a seed bilingual dictionary with one-to-one translations to prune source and target word representations into the shared subspace of the seed dictionary (Rapp 1999). Canonical correlation analysis can be used to create a new space in which the representations of source and target words which are translations of one another are maximally correlated (Haghighi *et al.* 2008; Faruqui and Dyer 2014). Word mappings (i.e. translation relations) are induced by an EM algorithm in Haghighi *et al.* (2008) whereas they are directly given by word alignment in parallel corpora in Faruqui and Dyer (2014). Mikolov, Le and Sutskever (2013b) assume that a linear transformation can map from the source space to the target space and learn a translation matrix to do so. They evaluate this method on a WMT 2011 word-translation task, where they obtain a better precision for the top 1 and top 5 translation candidates than methods based on edit distance or word-count context vectors.

Another possibility consists of building a monolingual representation for the source corpus, then transferring it to the target language through the word

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

alignments of a parallel corpus (Täckström, McDonald and Uszkoreit 2012; Zou *et al.* 2013) and finally adapting it to take into account word distribution statistics in the target corpus (possibly iterating back and forth). Zou *et al.* (2013) test the contribution of these representations to phrase-based MT by adding a semantic similarity feature to the decoder: the distance between bag-of-word representations (i.e. the average of word representations) of the two phrases in a bilingual phrase pair. This improves by 0.48 BLEU points its Chinese–English translations in the NIST 2008 dataset.

A series of methods have been proposed to learn source and target word representations jointly in a common space (Klementiev *et al.* 2012b; Chandar *et al.* 2014; Gouws *et al.* 2015) from monolingual and parallel corpora. Klementiev *et al.* (2012b) frame the problem as multitask learning where the interaction between tasks is based on word alignments computed from a parallel corpus. Chandar *et al.* (2014) and Gouws *et al.* (2015) do not require word alignments and directly process parallel sentences instead, which they represent by their average word vector. Chandar *et al.* (2014) jointly optimize four objectives for bilingual autoencoders which, from the representation of a sentence in a source language, can reconstruct both the original source sentence and its translated sentence. In Gouws *et al.* (2015), monolingual training is based on Mikolov *et al.* (2013a)'s negative sampling skipgram model, while bilingual synchronization is obtained by minimizing the distance between the bag-of-word representations of parallel sentences. All three methods are tested on a cross-language document classification task. Source and target documents are represented by the average of their word representations. Since they belong to the same space, this enables training a classifier on the source language and applying it to the target language by direct transfer. All three outperform a classifier trained on source documents (represented as bags of words) and applied to target documents which have been machine-translated to the source language, and successively gain in accuracy and speed. Gouws *et al.* (2015) also tackle the same WMT 2011 word-translation task as Mikolov *et al.* (2013b) and outperform its results.

In these methods, two monolingual corpora are 'connected' by a parallel corpus or a seed bilingual dictionary. However, very few of the cited references discuss the comparability of their monolingual corpora (Li and Gaussier 2010; Su and Babych 2012) and their compatibility with the parallel corpus.

5 A benchmark for measuring comparability

The increasing interest in comparable corpora research led to a considerable number of methods for dealing with its fundamental problems. However, it is often very hard to compare the performance of these methods as up to now there has been no agreement on common test data. In this situation, in the framework of the BUCC workshop series, three shared tasks have been envisaged: One for measuring the comparability of bilingual documents, another for extracting parallel segments from comparable corpora, and a third for bilingual lexicon extraction from comparable corpora. Of these, only the first has already been conducted as part of the

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

BUCC-2015 workshop[10] which was co-located with ACL-IJCNLP 2015.[11] In the following, we describe the design and the results of this first shared task which aimed at detecting the most similar documents in a large multilingual text collection. This provided a benchmark for evaluating different approaches for identifying more or less parallel documents.

5.1 Data set description

The dataset is derived from static Wikipedia dumps of the main articles. A feature of Wikipedia is that it provides so-called inter-language links between many corresponding articles of different languages, i.e. between articles describing the same or corresponding headwords. These inter-language links are provided by the authors of the articles, i.e. they are based on expert judgement. For the shared task, we selected bilingual pairs of articles which fulfilled the following requirements:

(1) The inter-language links between the articles had to be bidirectional, i.e. not only an article in Language_1 needs to be linked to the corresponding article in Language_2, but also vice versa. This ensured a page in one language is not linked only to a portion of a page in another one.
(2) The size of the textual content of the two articles within a pair (i.e. their length measured as the number of characters) had to be similar.

Note that this selection procedure for the article pairs implies that an article pair selected for one language pair may or may not be selected for another language pair. All articles which satisfied the selection conditions have been considered for the evaluation run.

The data for each language pair has been split randomly into two sets:

Training set: articles with information about the correct links for the respective language pairs provided to the participants;

Test set: articles without the links.

The task was for each article in the test set to submit up to five ranked suggestions to its linked article, assuming that the gold standard contains its counterpart in another language. The languages in the shared task were Chinese, French, German, Russian, and Turkish. Pages in these languages needed to be linked to a page in English. For each source page, there exists exactly one correct linked page in the gold standard.

5.2 Evaluation

Evaluation has been done using standard TREC evaluation measures,[12] modeling the task as the retrieval of a ranked list of links from a source page. The *Success*

[10] The second and the third shared task will be conducted as part of upcoming BUCC events.
[11] `https://comparable.limsi.fr/bucc2015/bucc2015-task.html`
[12] See `http://trec.nist.gov/trec_eval/`

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

measures correspond to commonly used measures when evaluating term translations in comparable corpora. We use them here to evaluate the proposed inter-language links between the articles.

Success@1 determines the proportion of source articles for which the correct target article has been ranked in the top position; Success@5 determines the proportion of source articles for which the correct target article has been ranked among the top five positions. *Mean Reciprocal Rank* is also a relevant measure: If the correct target article is ranked at position N, a score of $1/N$ is given to this source article. Then, these scores are averaged over the set of source articles. Mean Reciprocal Rank yields the same score as success@1 when the top ranked article is correct, but also scores decreasing fractions of one when the correct article is found anywhere in the ranking: this results in a higher average score than success@1.

5.3 Comparison of methods used by participating systems

The approach used by the system CCNUNLP is described in Li and Gaussier (2013). In essence, it uses a bilingual dictionary for converting the word feature vectors between the languages and for estimating their overlap. The other systems are discussed in detail in the proceedings of BUCC'15 (Morin *et al.* 2015; Zafarian *et al.* 2015), and full evaluation results are available there as well (Sharoff, Zweigenbaum and Rapp 2015). The LINA system (Morin *et al.* 2015) is based on matching hapax legomena, i.e. words occurring only once. In addition to using hapax legomena, the quality of linking in one language pair, e.g. French–English, is also assessed by using information available in pages in another language pair, e.g. German–English. The AUT system (Zafarian *et al.* 2015) uses the most complicated setup by combining several steps. First, documents in different languages are mapped into the same space using a feature transformation matrix. This helps in selecting a relatively small subset of pages to detect possible links. Second, document similarity is assessed using three pipelines, namely, a polylingual topic model, a named entities detection tool, and a word feature mapping procedure using MT.

Although the number of different runs is not sufficient to draw general conclusions, we can compare the same methods across different language pairs and different methods on the same language pairs.

CCNUNLP obtained better results on Chinese than on French, probably because of the quality of the underlying dictionaries. LINA.CL worked better on German than for French, while the reverse was true for LINA.P.[13] After the evaluation run, it occurred that the submissions of AUT had a data processing bug.

Overall, the CCNUNLP method obtained the best results on Chinese and French, followed by the LINA.CL method (second best on French, and best on German).

[13] LINA.CL and LINA.P are the cross-lingual and the pigeonhole-apprach as described by Morin *et al.* (2015).

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

5.4 Discussion

The results are encouraging. Success@1 rates reach 0.71 for Chinese and 0.61 for French and German. However, this level of accuracy is still far from a reliable identification of comparable Wikipedia pages. Given the small number of participating systems and an uneven coverage of the language pairs involved it is difficult to make predictions about which methods are more or less successful. A dictionary-based method (CCNUNLP) is slightly ahead of a method based on hapax legomena (LINA.*). A multi-stage method like the one used by AUT is promising, but its complexity makes it prone to errors.

Another question concerns the evaluation scenario. The shared task has been evaluated by using gold standard data in intrinsic evaluation. Given that the purpose of collecting comparable corpora is to provide more data for terminology extraction or MT, we need to evaluate text collections by referring to their successful use in such tasks. The limitation in using extrinsic evaluation is the lack of gold-standard methods and resources.

Acknowledgment

This work was in part supported by a Marie Curie Career Integration Grant within the seventh European Community Framework Programme.

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 ACL (EACL 2009)*, Athens, Greece, Association for Computational Linguistics.

Apidianaki, M., Ljubešić, N., and Fišer, D. 2013. Vector disambiguation for translation extraction from comparable corpora. *Informatica (Slovenia)* **37**(2): 193–201.

Beigman Klebanov, B., and Flor, M. 2013. Associative texture is lost in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*,, Sofia, Bulgaria, Association for Computational Linguistics, pp. 27–32.

Bouamor, D., Popescu, A., Semmar, N., and Zweigenbaum, P. 2013. Building specialized bilingual lexicons using large scale background knowledge. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, Association for Computational Linguistics, pp. 479–489.

Chandar, A. P. S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. C., and Saha, A. 2014. An autoencoder approach to learning bilingual word representations. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, pp. 1853–1861, Montreal, Quebec, Canada.

Dou, Q., Vaswani, A., Knight, K., and Dyer, C. 2015. Unifying bayesian inference and vector space models for improved decipherment. 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)*, Beijing, China, Association for Computational Linguistics, pp. 836–845.

Faruqui, M., and Dyer, C. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, Association for Computational Linguistics, pp. 462–471.

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

Fung, P., and McKeown, K. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pp. 192–202, see http://anthology.aclweb.org/W/W97/W97-0100.pdf.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, pp. 1606–1611.

Gouws, S., Bengio, Y., and Corrado, G. 2015. BilBOWA: fast bilingual distributed representations without word alignments. In F. Bach, D. and Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, JMLR Workshop and Conference Proceedings*, vol. 37, Lille, France.

Grefenstette, G. 1992. SEXTANT: exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware, USA, Association for Computational Linguistics, pp. 324–326.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, Association for Computational Linguistics, pp. 771–779.

Klementiev, A., Titov, I., and Bhattarai, B. 2012a. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, Mumbai, India, The COLING 2012 Organizing Committee, pp. 1459–1474.

Klementiev, A., Titov, I., and Bhattarai, B. 2012b. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, Mumbai, India, The COLING 2012 Organizing Committee, pp. 1459–1474.

Li, B., and Gaussier, E. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, Coling 2010 Organizing Committee, pp. 644–652.

Li, B., and Gaussier, E. 2013. Exploiting comparable corpora for lexicon extraction: measuring and improving corpus quality. In S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung (eds.), *Building and Using Comparable Corpora*, pp. 131–149. Springer-Verlag.

Maia, B. 2003. What are comparable corpora. In *Multilingual Corpora: Linguistic Requirements and Technical Perspectives. Workshop at the Corpus Linguistics Conference*, Lancaster, UK.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the Workshop at ICLR'13.*

Mikolov, T., Le, Q. V., and Sutskever, I. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Morin, E., Hazem, A., Boudin, F., and Loginova-Clouet, E. 2015. LINA: identifying comparable documents from wikipedia. In *Proceedings of the Workshop on Building and Using Comparable Corpora at ACL 2015.*

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 (EMNLP)*, Philadelphia, PA, USA. Association for Computational Linguistics.

Munteanu, D. S., and Marcu, D. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* **31**(4): 477504.

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 (COLING/ACL 2006)*, Sydney, Australia, Association for Computational Linguistics.

Nuhn, M., and Ney, H. 2014. Em decipherment for large vocabularies. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA, pp. 759–764.

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

Nuhn, M., Schamper, J., and Ney, H. 2015. Unravel - a decipherment toolkit. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, pp. 549–553.

Rapp, R. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd ACL*, Cambridge, MA, pp. 320–322.

Rapp, R. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th ACL*, Maryland, Association for Computational Linguistics, pp. 395–398.

Rapp, R., Zweigenbaum, P., and Sharoff, S. 2010. Preface. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora at LREC 2010*, page V, Valletta, Malta. European Language Resources Association (ELRA).

Ravi, S., and Knight, K. 2008. Attacking decipherment problems optimally with low-order ngram models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Honoluli, Hawaii, Association for Computational Linguistics, pp. 812–819.

Ravi, S., and Knight, K. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, Association for Computational Linguistics, pp. 812–819.

Saluja, A., Hassan, H., Toutanova, K., and Quirk, C. 2014. Graph-based semi-supervised learning of translation models from monolingual data. In *Proceedings of the 52nd ACL,* Baltimore, MD, June.

Seraj, R. M. 2015. *Paraphrases for Statistical Machine Translation*. PhD Thesis, Simon Fraser University.

Sharoff, S., Rapp, R., and Zweigenbaum, P. 2013a. Overviewing important aspects of the last twenty years of research in comparable corpora. In S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung (eds.), *BUCC: Building and Using Comparable Corpora*, pp. 1–17. Springer.

Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P. (eds.) (2013b. *BUCC: Building and Using Comparable Corpora*. Springer.

Sharoff, S., Zweigenbaum, P., and Rapp, R. 2015. Bucc shared task: cross-language document similarity. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, Beijing, China, Association for Computational Linguistics, pp. 74–78.

Su, F., and Babych, B. 2012. Development and application of a cross-language document comparability metric. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, European Language Resources Association (ELRA), pp. 3956–3962.

Täckström, O., McDonald, R., and Uszkoreit, J. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 477–487.

Tang, L., Wang, T., and Chen, Y. 2015. Problems of alignment in paraconc for a case study. In *Ally Hu (ed.): Computer Science and Applications. Proceedings of the 2014 Asia-Pacific Conference on Computer Science and Applications*, Shanghai, China, Taylor & Francis, London, UK, pp. 57–62.

Vulic, I., and Moens, M.-F. 2014a. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *EMNLP*, pp. 349–362.

Vulic, I., and Moens, M.-F. 2014b. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *EMNLP*, pp. 349–362.

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press

Weaver, W. 1955. Translation. In W.N. Locke, D. B., editor, *Machine Translation of Languages*, pp. 15–23. MIT Press.

Wu, D., and Fung, P. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Natural Language Processing–IJCNLP 2005*, Springer, pp. 257–268.

Zafarian, A., Agha Sadeghi, A. P., Azadi, F., Ghiasifard, S., Ali Panahloo, Z., Bakhshaei, S., and Mohammadzadeh Ziabary, S. M. 2015. AUT document alignment framework for BUCC workshop shared task. In *Proceedings of the Workshop on Building and Using Comparable Corpora at ACL 2015.*

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*, Seattle, Washington, USA, Association for Computational Linguistics, pp. 1393–1398.

https://doi.org/10.1017/S1351324916000115 Published online by Cambridge University Press