Natural Language Engineering **24** (4): 523–549. © Cambridge University Press 2018
doi:10.1017/S1351324917000481

Measuring bilingual corpus comparability†

BO LI[1], ERIC GAUSSIER[2] and DAN YANG[3]
[1]*Department of Computer Science, Central China Normal University, Wuhan, China*
e-mail: libo@mail.ccnu.edu.cn
[2]*CNRS-LIG/AMA, Université Grenoble Alpes, Grenoble, France*
e-mail: eric.gaussier@imag.fr
[3]*China Electric Power Research Institute, Wuhan, China*
e-mail: yangdan3@epri.sgcc.com.cn

(*Received 2 March 2017; revised 14 December 2017; accepted 15 December 2017; first published online 15 January 2018*)

Abstract

Comparable corpora serve as an important substitute for parallel resources in cases of under-resourced language pairs. Previous work mostly aims to find a better strategy to exploit existing comparable corpora, while ignoring the variety in corpus quality. The quality of comparable corpora affects a lot its usability in practice, a fact that has been justified by several studies. However, researchers have not been able to establish a widely accepted and fully validated framework to measure corpus quality. We will thus investigate in this paper a comprehensive methodology to deal with the quality of comparable corpora. To be exact, we will propose several comparability measures and a quantitative strategy to test those measures. Our experiments show that the proposed comparability measure can capture gold-standard comparability levels very well and is robust to the bilingual dictionary used. Moreover, we will show in the task of bilingual lexicon extraction that the proposed measure correlates well with the performance of the real world application.

1 Introduction

A bilingual corpus is an important resource used to cross the language barrier in multilingual Natural Language Processing (NLP) tasks such as Statistical Machine Translation (SMT) (Och and Ney 2003; Bahdanau, Cho and Bengio 2015) and Cross-Language Information Retrieval (CLIR) (Ballesteros and Croft 1997). Parallel corpora, i.e., document collections comprised of texts that are translation of one another, have been broadly used in cross-language NLP tasks. Their availability remains however limited, especially for minority languages (Markantonatou *et al.* 2006). Parallel corpora publicly available only exist in narrow domains and between a few language pairs. For example, the *Europarl* corpus (Koehn 2005) widely used in SMT research was built by retrieving parallel texts from the proceedings of the

†This work was co-supported by Natural Science Foundation of China (Nos. 61300144 and 61572223), State Language Commission of China (No. YB125-132), Humanity and Social Science Foundation of Ministry of Education of China (No. 15YJC870029) and the Fundamental Research Funds for Central Universities (No. CCNU16A06015, CCNU15A05062, CCNU17GF0005, CCNUSZ2017024).

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

European commission. Other parallel corpora such as the United Nations corpus and the Hansard corpus[1] suffer from the same problem.

As another type of bilingual corpora, a comparable corpus is in general easier to obtain as the only requirement is that documents cover related content in different languages. Previous work has shown that comparable corpora can be successfully used in such applications as bilingual lexicon extraction (for example, refer to Rapp (1999) and Chebel, Latiri and Gaussier (2017)), enhancement of SMT systems (Munteanu, Fraser and Marcu 2004; AbduI-Rauf and Schwenk 2009; Hazem and Morin 2016), enhancement of CLIR systems (Talvensaari *et al.* 2007), as well as the modeling of topics across languages (Boyd-Graber and Blei 2009; Ni *et al.* 2009). There is thus more and more evidence that comparable corpora can be used to bridge the language barrier in applications where parallel resources are not available.

In all these studies, the definition of what is a comparable corpus is rather vague. For example, Ji (2009) defines a comparable corpus as a text collection consisting of documents describing similar topics. In Munteanu *et al.* (2004) and Hewavitharana and Vogel (2008), a comparable corpus is defined as a text collection covering overlapping information. Sharoff, Rapp and Zweigenbaum (2013) state that comparable corpora are less parallel corpora. Based on discussions in former literature, we define here comparable corpora as document sets in different languages that cover similar topics. Comparable corpora can thus be quite different from each other, depending on the extent to which two monolingual parts of the comparable corpus are related to each other. Intuitively, a parallel corpus can be seen as a special case of the comparable corpus, corresponding to the highest comparability level.[2]

Data driven NLP tasks depend a lot on the quality of the resources used, and the better the corpus is, the better knowledge one can extract from it. We thus conjecture that comparable corpora of higher quality[3] will yield better performance of applications relying on them, a fact that has actually been validated in several previous studies (Li and Gaussier 2010; Skadina *et al.* 2010). Existing work mining comparable corpora mostly builds and uses comparable corpora according to humans' simple intuitions. For instance, Robitaille *et al.* (2006) construct a comparable corpus by using the search engine to retrieve web pages highly relative to a set of queries given in two languages. Munteanu and Marcu (2006) build the comparable corpus by using the existing new corpora in the same period and the CLIR system to retrieve related document pairs. These studies are reasonable, but there is still a risk that one might obtain and use a corpus of poor quality, leading to unpredictable performance, since there has not been any method one could use

[1] Both the United Nations corpus and the Hansard corpus are available from http://www.ldc.upenn.edu.

[2] By definition, the parallel corpus could consist of documents in quite different topics/domains. In practice, parallel corpora normally consist of documents in a special domain.

[3] The quality of a multilingual corpus is normally affected by several factors, for instance volume, parallelness and novelty. In this paper, we will mostly concentrate on parallelness.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

to tell the corpus quality in a quantitative way. Without such a measure, one could not build and use a comparable corpus with full confidence.

The task of measuring text similarity has attracted much attention in the NLP field. The Semantic Textual Similarity (STS) shared task, held as part of SemEval, [4] aims to automatically predict the similarity of a pair of sentences. The classic strategies for measuring monolingual text similarity could be extended to the cross-language settings (e.g., Mathieu, Besancon and Fluhr (2004), Luong, Pham and Manning (2015)), which is the most direct approach to realize the comparability measure. The European FP7 project ACCURAT[5] has proposed several comparability measures according to this intuition. Besides, there are several other possibilities to investigate the comparability. These studies will be discussed in more details in Section 2. However, previous work has not been able to propose a systematic approach to measure the corpus quality and to test the comparability measures themselves.

We will first establish in this paper measures to capture different comparability levels. The proposed measures will then be examined on gold-standard comparability levels to show their coherence with gold standards. In addition, the measures will be tested in the task of bilingual lexicon extraction to show its correlation with the performance of real world tasks. The remainder of the paper is organized as follows:

- In Section 2, we define more precisely the notion of comparability and develop several comparability measures. Several other measures are also considered in this section for comparison purposes.
- The comparability measures are evaluated, in Section 3, in terms of correlation with gold-standard comparability levels and robustness to dictionary coverage. In addition, the comparability measures are used in a practical application, bilingual lexicon extraction, to show that the measure correlates well with its usability.
- Several aspects of this study are discussed in Section 4, and the paper is concluded in Section 5.

2 *Comparability measures*

A fine-grained comparability measure is important when one needs to choose the best one from several comparable corpora prepared from different resources or when one needs to enhance the quality of a given corpus of low quality in a systematic way. We thus plan to develop in this section a quantitative measure to capture various comparability levels.

As we have discussed in Section 1, comparable corpora serve as the substitute for parallel resources in case of under-resourced language pairs. The ideal comparability measures should rely, as less as possible, on parallel resources that are expensive to obtain for under-resourced languages. Moreover, computationally cheap measures are preferred, which can lead to feasible solutions for such tasks as corpus quality enhancement (Li, Gaussier and Aizawa 2011).

[4] http://alt.qcri.org/semeval2015/
[5] http://www.accurat-project.eu

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

There have been only a few works trying to investigate the formal definition or quantification of the quality of comparable corpora. Such works as Rayson and Garside (2000) and Kilgarriff (2001) are early attempts to quantify how similar two corpora in the same language are in terms of lexical content. Sharoff (2007) investigates automatic ways of differentiating web corpora in terms of domains and genres. Saralegi, San-Vicente and Gurrutxaga (2008) attempt to measure the degree of comparability of two corpora in different languages by inferring a global comparability from the similarity of all cross-language document pairs. This measure is however computationally infeasible if the corpora contain a large number of documents. Under the seventh European framework,[6] researchers involved in the project ACCURAT (Skadina *et al.* 2010) have studied several measures and metrics for assessing corpus comparability and document parallelism of under-resourced languages. In addition to the above studies devoted to comparable corpora, researchers have recently developed several cross-lingual models of word embeddings (Hermann and Blunsom 2014; Luong, Pham and Manning 2015; Vulic and Moens 2015), which could be used to model cross-lingual semantic similarity. The word embedding approaches however rely on a training procedure on parallel or comparable corpora, which is computationally expensive. We will use some of these measures as baselines for comparison purposes only.

We first introduce below a set of measures that are extensions of the ones proposed in our former work (Li and Gaussier 2010). All the measures we are going to consider can be classified into one of the following categories:

- *Vocabulary overlapping*. These measures aim at assessing to which extent the vocabularies of two corpora overlap with each other. They require mapping the vocabulary of one monolingual corpus to the language of the other monolingual corpus, prior to comparing them. In this study, we make use of bilingual dictionaries to bridge the language barrier and detail the measures in Section 2.1.
- *Vector space mapping*. These measures, based on a bag-of-words representation of the two monolingual parts of a comparable corpus, make use of standard similarity measures between vectors (see for example Gaussier, Renders, Matveeva, Goutte and Déjean (2014)). This approach will be described in section 2.2.
- *Machine translation*. An alternative to the above approaches is to use machine translation (MT) system to translate one corpus into the language of another corpus, resulting in two corpora in the same language that can be compared using standard approaches in monolingual settings. This approach will be introduced in Section 2.3.
- *Cross-lingual word embeddings*. Unsupervised learned word embeddings have been exceptionally successful in many NLP tasks. We directly make use here of the cross-lingual version of skip-gram model (Luong, Pham and Manning 2015) as comparability measures.

[6] http://cordis.europa.eu/fp7/

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

MT approaches use parallel corpora for training, which are rare for under-resourced language pairs. *Cross-lingual word embedding* approaches make use of the training procedure on parallel or comparable corpora, which is computationally expensive. Those approaches are presented here as baselines only for the purpose of comparison. For convenience, the following discussions will be made in the context of French–English comparable corpora.

Comparability measures can be defined on various levels such as sentences, documents or whole corpora. Comparability measures on the document or sentence level simply amount to measuring the similarity of sentences/documents across languages, which is a classic task in NLP (e.g., Luong, Pham and Manning (2015)). It is not clear whether these classic techniques can be directly used to capture the differences between various comparable corpora. We intend to define comparability so as to reflect the usability of the comparable corpus in NLP tasks. It is generally considered that different types of multilingual corpora display different levels of comparability. For example, the following comparable corpora have decreasing comparability levels[7]:

(1) Parallel corpora;
(2) Parallel corpora with noise;
(3) Non-parallel corpora covering overlapping topics (i.e., strongly comparable);
(4) Non-parallel corpora covering different topics (i.e., weakly comparable).

A good comparability measure should correlate well with the different comparability levels above. In other words, it should be able to capture (fine-grained) differences on the levels of comparability. We will return to this issue in Section 3. Prior to that, we first introduce several comparability measures below.

2.1 Comparability measures based on vocabulary overlapping

The measures proposed in this section are extensions of those proposed in our former work (Li and Gaussier 2010). They are based on the assumption that it is easier to find translation pairs between documents that are more comparable to each other, since authors tend to use similar words to depict similar topics in different languages (see Morin *et al.* (2007) for a related analysis). This fact is also supported by the four types of comparable corpora presented above, as translation pairs are common in parallel corpora, and not so much in corpora covering different topics.

2.1.1 Context-free measures

A natural way to estimate the number of translation pairs in a bilingual corpus is to rely on the *mathematical expectation* of finding such pairs. Let us assume that we have a French–English comparable corpus $\mathcal{C}$ consisting of a French part $\mathcal{C}_f$ and an

[7] A similar list of decreasing comparability levels can be found in Sharoff *et al.* (2013).

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

English part $\mathcal{C}_e$. If we consider the translation process from the English part to the French part, the comparability measure M_{ef} can be defined as the expectation of finding, for each English word w_e in the vocabulary $\mathcal{C}_e^v$ of $\mathcal{C}_e$, its translation in the vocabulary $\mathcal{C}_f^v$ of $\mathcal{C}_f$. The definition for M_{ef} directly reflects our intuition. As one can note, a general English–French bilingual dictionary $\mathcal{D}$, independent from the corpus $\mathcal{C}$, is required to judge if two words are the translation of each other. Let σ be a function that indicates whether a translation from the translation set $\mathcal{T}_w$ of a word w is found in the vocabulary $\mathcal{C}^v$ of a corpus $\mathcal{C}$, i.e.,:

$$\sigma(w, \mathcal{C}^v) = \begin{cases} 1 & \text{iff } \mathcal{T}_w \cap \mathcal{C}^v \neq \emptyset \\ 0 & \text{else} \end{cases} \tag{1}$$

M_{ef} is then defined as

$$M_{ef}(\mathcal{C}_e, \mathcal{C}_f) = \mathbb{E}(\sigma(w, \mathcal{C}_f^v) | w \in \mathcal{C}_e^v) = \sum_{w \in \mathcal{C}_e^v} \sigma(w, \mathcal{C}_f^v) \cdot \Pr(w) \tag{2}$$

As assumed above, the comparable corpus and the general bilingual dictionary are independent from one another. It is thus natural to assume that the dictionary $\mathcal{D}$ covers a substantial part of $\mathcal{C}_e^v$ and this substantial part can well represent the whole vocabulary. It means the expectation of finding the translation in $\mathcal{C}_f^v$ of a word w in $\mathcal{C}_e^v$ can be approximated by that in $\mathcal{C}_e^v \cap \mathcal{D}_e^v$. This assumption leads to equation (3):

$$M_{ef}(\mathcal{C}_e, \mathcal{C}_f) = \sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \sigma(w, \mathcal{C}_f^v) \cdot \Pr(w) \tag{3}$$

where $\mathcal{D}_e^v$ is the English vocabulary of the given bilingual dictionary $\mathcal{D}$. The derivation from equations (2)–(3) seems trivial. It does help us bring up the notion of dictionary coverage and the robustness measure in Section 3.2.3.

There are several possibilities to estimate $\Pr(w)$ in equation (3). However, the presence of common words suggests that one should not solely rely on the number of occurrences (i.e., term frequency), which is a broadly used approach in other fields like unigram language models, since the high-frequency words will dominate the final results. For example, in the *Europarl* corpus, the English word *Europe* and the French word `Europe` are very common words. It means that even if one piece of English text and one piece of French text are randomly picked from the *Europarl* corpus, we can still expect to find many translation pairs *Europe*–`Europe`. To avoid the bias common words can introduce in the comparability measure, one can weight each word w as ρ_w through TF-IDF or through the simple Presence/Absence (P/A) criterion. In this case, the part $\Pr(w)$ can be estimated as

$$\Pr(w) = \frac{\rho_w}{\sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \rho_w} \tag{4}$$

With the P/A criterion, the weight ρ_w is 1 if and only if $w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v$, otherwise the value is 0. Alternatively, considering the TF-IDF weight for each word w, the weighting function ρ_w can be defined in a standard TF-IDF style (Salton, Wong and Yang 1975).

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Similarly, when considering the translation process from the French part to the English part, the counterpart of M_{ef}, M_{fe}, can be written as

$$M_{fe}(\mathcal{C}_e, \mathcal{C}_f) = \sum_{w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v} \sigma(w, \mathcal{C}_e^v) \cdot \Pr(w) \tag{5}$$

where $\Pr(w)$ is defined in the same way as in equation (4).

The two asymmetric measures M_{ef} and M_{fe} above reflect the degree of comparability when considering the translation process in only one direction. They can be combined to form a comprehensive measure M by reviewing the two directions as a whole. The difference between M_{ef} in equation 3 and M_{fe} in equation (5) comes from the part which estimates the probability of a word in the corpus vocabulary. We denote $\mathcal{C}^v$ as the whole vocabulary of $\mathcal{C}$ and $\mathcal{D}^v$ as the whole vocabulary of $\mathcal{D}$. Following the same idea, considering the English and French vocabularies as a whole, one can directly obtain $\Pr(w)$ for each w in $\mathcal{C}^v \cap \mathcal{D}^v$ by replacing $\mathcal{C}_e^v \cap \mathcal{D}_e^v$ with $\mathcal{C}^v \cap \mathcal{D}^v$ in equation (4). The combined measure M, considering the translation in both directions, can then be written as

$$M(\mathcal{C}_e, \mathcal{C}_f) = \sum_{w \in \mathcal{C}^v \cap \mathcal{D}^v} \sigma(w, \mathcal{C}^v) \cdot \Pr(w) \tag{6}$$

We give here additional comments on the P/A criterion. With this criterion, $\Pr(w)$ for each w in $\mathcal{C}^v \cap \mathcal{D}^v$ is directly $\frac{1}{|\mathcal{C}^v \cap \mathcal{D}^v|}$. One can then obtain the combined measure M as

$$M(\mathcal{C}_e, \mathcal{C}_f) = \frac{\sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \sigma(w, \mathcal{C}_f^v) + \sum_{w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v} \sigma(w, \mathcal{C}_e^v)}{|\mathcal{C}_e^v \cap \mathcal{D}_e^v| + |\mathcal{C}_f^v \cap \mathcal{D}_f^v|} \tag{7}$$

which corresponds to the overall proportion of words for which a translation can be found in the comparable corpus. One can notice from equation (7) that M is a symmetric measure.

2.1.2 Context-based measures

The previous development has left aside an important problem, namely the one of *polysemy*. Due to *polysemy*, two words can be treated as the translation of each other according to the dictionary but might hold different senses in the corpus. We can make use of the simple assumption that words in parallel translation usually appear in similar contexts to disambiguate the translation candidates in the dictionary. This same assumption has been broadly exploited in the bilingual lexicon extraction tasks. We will embed the assumption in the function σ in equation (1) to build a context-sensitive version of the previous measures. Let us assume that the English word w_e (resp. French word w_f) appears in the context word set $\mathcal{S}_e$ (resp. $\mathcal{S}_f$) consisting of the words surrounding w_e (resp. w_f) in a certain window in the corpora. Then the similarity of the two context sets is measured by the overlap of the two sets which is directly the proportion of words of which the translation can be found in the counterpart set. Formally, the similarity of w_e and w_f, in terms of their context

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

similarity, can be written as

$$\mathrm{sim}(w_e, w_f) = \frac{\sum_{w \in \mathcal{S}_e \cap \mathcal{D}_e^v} \sigma(w, \mathcal{S}_f) + \sum_{w \in \mathcal{S}_f \cap \mathcal{D}_f^v} \sigma(w, \mathcal{S}_e)}{|\mathcal{S}_e \cap \mathcal{D}_e^v| + |\mathcal{S}_f \cap \mathcal{D}_f^v|}$$

The enhanced version of the function σ in equation 1 is then defined as

$$\sigma_c(w, \mathcal{C}^v) = \begin{cases} 1 & \text{iff } \exists w' \in \mathcal{T}_w \cap \mathcal{C}^v, \ \ \mathrm{sim}(w, w') > \delta \\ 0 & \text{else} \end{cases}$$

where δ, empirically set to 0.3 in our experiments, is the threshold for the similarity. A word w is deemed to be translated, according to the function σ_c, if at least one of its translations w' identified by the function σ in the corpus, is similar to w based on the context similarity measure $\mathrm{sim}(w, w')$. Replacing σ with σ_c in equation (1) will lead to the context versions of the comparability measures above, which are respectively denoted as M_{ef}^c, M_{fe}^c and M^c.

Besides the simple idea used in above context-based measures, there are definitely several other possibilities to improve it. First, the window-based co-occurrence context can be improved by using surface-distance model proposed in Washtell (2009). Second, such techniques as LSA and LDA (Deerwester *et al.* 1900; Blei and Jordan 2003) are usually considered to be more efficient than the simple context vector approaches according to the experiences. However, as we will show, the simple measure M^c or even its context-free version M have already shown very satisfactory performance and are enough for practical usage. Moreover, the simple ideas are computationally much cheaper. The possible enhancements are thus not the prior direction of this paper and will not be investigated further.

The above measures rely on the intersection between bilingual corpora and dictionary. One can tell from the intuition that the corpus vocabulary tends to cover all the words in the dictionary while the corpora are long enough, since the dictionary vocabulary is generally smaller compared to that of the corpus. It is to say, those measures could be sensitive to corpus length. Taking the measure M_{ef} in equation (3) as an example, once the French part $\mathcal{C}_f$ is long enough, it might be able to cover all the French dictionary vocabulary, meaning that M_{ef} could be high even if the French part and the English part are not really comparable. We will review this problem in Section 3.2.2 and give a formal analysis on dictionary coverage in Section 3.2.3.

2.2 Measures based on vector space mapping

It is common practice to represent documents as vectors consisting of words occurring in the documents. The weight of each dimension, i.e., a word, is determined by methods such as TF-IDF. In the cross-language settings, one needs to compare two vectors in different languages, i.e., one vector in the source language needs to be mapped to a vector in the target language. Let us assume one English text set is represented as a vector $\vec{v_e}$, whereas a French text set is represented as a vector $\vec{v_f}$. $\vec{v_f}$ can then be mapped to $\vec{v_e}$ by accumulating the contributions from words in $\vec{v_f}$ that yield identical translations. Let us further assume that $f(w_e)$ (resp. $f(w_f)$) denotes

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

the weight of the word w_e (resp. w_f) in the vector $\vec{v_e}$ (resp. $\vec{v_f}$). In this paper, the weight is defined in a standard TF-IDF style. The dot product between the vectors $\vec{v_e}$ and $\vec{v_f}$ is given by

$$< \vec{v_e}, \vec{v_f} >= \sum_{w_e \in \vec{v_e}} f(w_e) \sum_{w_f \in \mathcal{T}_{w_e} \cap \vec{v_f}} f(w_f) \tag{8}$$

where $\mathcal{T}_{w_e}$ is the translation set of w_e in the bilingual dictionary. Different measures can be derived from the above dot product. We make use here of the standard cosine similarity, which yields the comparability measure M^v:

$$M^v = cos(\vec{v_e}, \vec{v_f}) = \frac{< \vec{v_e}, \vec{v_f} >}{\sqrt{\sum_{w_e \in \vec{v_e}} (f(w_e)^2 + (\sum_{w_f \in \mathcal{T}_{w_e} \cap \vec{v_f}} f(w_f))^2)}}$$

The vector space model has shown rather satisfactory performance in classic tasks such as document classification and clustering. The content we will model in this work is corpus but not document, where the most significant difference is that a corpus usually consists of documents in many different topics. According to the idea of topic models (Blei and Jordan 2003), each document can be seen as a mixture of separate topics. So the vector representation of the whole corpus can then be seen as a rather complicated mixture of various topics, making the vector comparison not as accurate as that in the case of documents.

2.3 Measures based on machine translation

In addition to the previous measures, we propose here a direct approach for measuring the comparability based on MT systems. However, this approach, contrary to the previous ones, cannot be applied on all language pairs as it requires *state-of-the-art* MT systems. The idea here is to translate the source language corpus into the target language prior to comparing the two corpora in the target language. Several comparison measures can be used for this latter purpose. For the translation task, we make use here of the google MT tool[8] that is often considered as state-of-the-art.

The BLEU score (Papineni *et al.* 2002) was initially developed to automatically evaluate MT systems with reference translations, where a good translation candidate should share many n-grams with the reference translation. In order to compare the two target language corpora, we will make use here of a similar idea. We thus represent the corpus as a vector of n-grams. Each dimension of the vector is the weight of the corresponding n-grams computed in the same TF-IDF style as before (by setting n to 1, one will actually obtain a vector space representation). The two vectors of n-grams are then compared with a standard cosine similarity. In order to limit the computational complexity, n is set to 1 and 2, corresponding to the comparability measures M^{g1} and M^{g2}.

[8] http://translate.google.com

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

3 Validation of comparability measures

In this section, we aim at assessing the performance of the proposed comparability measures. The materials used in the experiments are firstly given in Section 3.1. We then design several comparable corpora with gold-standard comparability levels in Section 3.2.1. Following that, we evaluate the performance of the measures in terms of correlation scores in Section 3.2.2 and robustness in Section 3.2.3. Last, the comparability measure is further tested in a real world task to estimate the consistency of comparability levels and the performance of bilingual lexicon extraction.

3.1 Resources in the experiments

For the experiments designed to compare and validate the comparability measures, several corpora are used: The parallel English–French *Europarl*[9] corpus, the TREC[10] *Associated Press* corpus and the corpora used in the multilingual track of CLEF,[11] which includes the *Los Angeles Times*, *Glasgow Herald*, *Le Monde*, *SDA French 94* and *SDA French 95*. In addition to these existing corpora, two monolingual corpora from the Wikipedia dump[12] are built. For English, we obtain the corpus *Wiki-En* by retrieving all the articles below the root category *Society*. For French, the corpus *Wiki-Fr* is built by getting all the articles below the category *Société*. Those two categories are connected by a cross-language link in Wikipedia and consist of articles of similar topics, although the French word *Société* has a stronger focus on the *company* meaning. The information of all the corpora used in the experiments is detailed in Table 1. Since the Europarl corpus in use has been aligned on the sentence level and stored as sentence pairs, the number of documents (Nr. docs) is not available in the table.

The bilingual dictionary used in our experiments is extracted from the google online dictionary. It consists of 33k distinct English words and 28k distinct French words, which constitutes 76k translation pairs. Such standard preprocessing steps as tokenization, POS-tagging and lemmatization are performed on all the corpora using the tool Treetagger (Schmid 1995). We directly work on lemmatized forms of content words (nouns, verbs, adjectives, adverbs).

3.2 A methodology for evaluating comparability measures

We evaluate comparability measures along the following lines:

(1) Whether the designed comparability measures can capture different comparability levels in the corpora;
(2) Whether the proposed measures are robust to dictionary coverage.

[9] http://www.statmt.org/europarl/
[10] http://trec.nist.gov/
[11] http://www.clef-campaign.org
[12] The Wikipedia dump files can be downloaded at http://download.wikimedia.org

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Table 1. *The information of the corpora used in the experiments in Section 3 (k = 1,000, m = 1,000k)*

Name	Short name	Language	Nr. docs	Nr. words
Europarl	Europarl	English	...	51 m
		French	...	55 m
Associated press	AP	English	243 k	126 m
Los Angeles times	LAT94	English	113 k	71 m
Glasgow herald	GH95	English	56 k	27 m
Le monde	MON94	French	44 k	24 m
SDA French 1994	SDA94	French	43 k	13 m
SDA French 1995	SDA95	French	43 k	13 m
Wiki-En	Wiki-En	English	368 k	163 m
Wiki-Fr	Wiki-Fr	French	378 k	169 m

3.2.1 Constructing test corpora

In order to test the comparability measures introduced before, one needs to have some corpora of known comparability levels, which does not exist according to our knowledge. One possibility is to construct such corpora with gold-standard comparability levels by human annotators, which is a costly task. Moreover, it is much harder for people to assign a precise score on corpus level than on sentence/document level as done in the Semantic Textual Similarity shared task of SemEval. Due to the lack of annotated corpora, existing works (e.g., Upadhyay *et al.* (2016)) have to employ such indirect tasks as dictionary extraction, document-classification and syntactic dependency parsing in order to test its semantic similarity models.

Without human annotators, the feasible choice is to build such test corpora automatically. The idea is to introduce noise into high-quality comparable corpora so as to obtain artificial corpora with decreasing comparability levels, which might not be as accurate as human annotation. We construct three such groups of corpora from the parallel *Europarl* corpus and the monolingual *AP* corpus, as follows:

- G_a: All the comparable corpora in G_a are built from the parallel corpus *Europarl*. One starts from the parallel corpus *Europarl*, considered as having the highest comparability level, and gradually decreases its quality by replacing larger parts with some non-parallel parts also present in *Europarl*. In other words, noise is added to the parallel corpus, however covering topics similar with the ones in the original parallel corpus;
- G_b: As before, one starts with the parallel corpus *Europarl*. The difference here is that one exchanges parallel parts of *Europarl* with content extracted from the *AP* corpus. That is to say, the noise brought to the parallel corpus covers different topics from the original parallel corpus.
- G_c: The comparable corpora obtained in G_a correspond to *non-parallel corpora covering similar topics*. In G_c, one starts from corpora with the lowest comparability levels in G_a, i.e., the ones containing no parallel parts, and exchanges some of their parts containing similar topics with content from

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

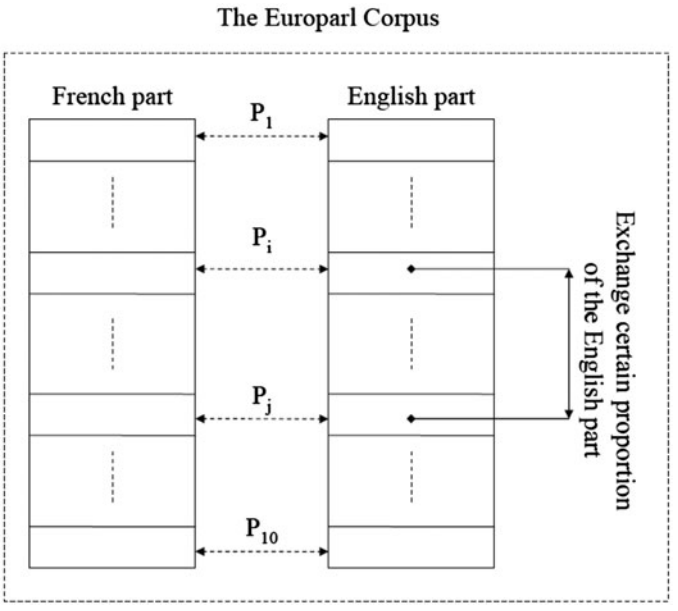


Fig. 1. Constructing the test corpus group G_a with gold-standard comparability levels.

the *AP* corpus, meaning that parts covering different topics are introduced, leading to comparable corpora belonging to the class *non-parallel corpora covering different topics*.

We now give more details on this construction process. The first group G_a is built from the *Europarl* corpus through the following two steps:

(1) The English (and its corresponding French) part of the *Europarl* corpus is split into ten equal parts in terms of number of sentences, leading to ten parallel corpora denoted as $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{10}$. The comparability level of these ten parallel corpora are arbitrarily set to one (i.e., the highest level);
(2) For each parallel corpus, e.g., $\mathcal{P}_i$ $(i = 1, 2, \ldots, 10)$, we replace a certain proportion p of the English part of $\mathcal{P}_i$ with content of the same size, again in terms of the number of sentences, from another parallel corpus $\mathcal{P}_j (j \neq i)$, producing a new corpus $\mathcal{P}_i'$ likely to contain less translation pairs and thus be of lower comparability. For each $\mathcal{P}_i$, as p increases one obtains a series of comparable corpora with decreasing comparability scores. In our experiments, p is increased from 0 to 1 with a step of 0.01. All the $\mathcal{P}_i$ and their respective descendant corpora, according to different values of p, constitute the group G_a, the comparability score of each corpora being set to $1 - p$. As a result, we have 1,000 comparable corpora in G_a. This process is illustrated in Figure 1.

The difference between building the corpora in G_b and in G_a is that, in G_b, the replacement in $\mathcal{P}_i$ is done with documents from the *AP* corpus and not from another parallel corpus $\mathcal{P}_j$ from *Europarl*. Compared with the corpora in G_a, we further degrade the parallel corpus $\mathcal{P}_i$ in G_b since the *AP* corpus covers different topics from *Europarl*.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

In G_c, we start with ten comparable corpora $\mathcal{P}'_i$ from G_a having the lowest comparability score (i.e., 0). They thus contain documents from *Europarl* that are not the translation of each other. Each $\mathcal{P}'_i$ is further altered by replacing certain portions, according to the same proportion p used before, with documents from the *AP* corpus. Although $\mathcal{P}'_i$ itself is comparable and not parallel, its English and French parts are likely to cover similar topics embedded in the *Europarl* corpus. Replacing certain parts of $\mathcal{P}'_i$ with the content from *AP* will thus further degrade the comparability levels of $\mathcal{P}'_i$.

From the process of building the comparable corpora in G_a, G_b and G_c, one can note that the gold-standard comparability scores in different groups, e.g., G_a and G_c, cannot be compared with each other directly, since the comparability scores are normalized between 0 and 1 in each group of corpora. These comparability scores do not represent absolute judgements on the comparability of the corpora considered, but rather relative scores within each group of corpora.

3.2.2 Correlation with gold-standard comparability levels

The goal here is to assess whether the comparability measures we have introduced can capture the differences in comparability levels in the three different groups G_a, G_b and G_c. As the comparability levels are set arbitrarily in these groups, we are interested here in assessing whether the measures yield the correct ranking (in terms of comparability scores) or not. In order to quantify this, we use Pearson Correlation Coefficient, defined as

$$r = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2}\sqrt{\sum_i (Y_i - \overline{Y})^2}}$$

where X_i denotes the comparability score provided by one measure on a given bilingual corpus and Y_i is the arbitrary comparability score (i.e., gold standard) assigned to this corpus in the construction process. $\overline{X}$ represents the average of X_is over all the bilingual corpora considered in G_a, G_b or G_c (and similarly for $\overline{Y}$).

Let us first recall the measures we have proposed in Section 2:

- Measures based on vocabulary overlapping. These measures are defined as the mathematical expectation of finding the translation for each word in the corpus vocabulary. Corresponding to the English/French/whole vocabulary of the corpus, we have the measures M_{ef}, M_{fe}, M and their contextual versions M^c_{ef}, M^c_{fe} and M^c. For the six measures, we will consider both the versions with the P/A weighting criterion and the versions with TF-IDF weighting schema, amounting to twelve measures in total.
- Baseline measures. These measures correspond to the vector spaces mapping and MT approaches discussed before: the measure M^v is based on the bilingual vector space model and the measures M^{g1} and M^{g2} are based on a MT system and n-grams representation. In addition, we use as baselines the recently proposed bilingual skip-gram model *BiSkip* (denoted as M^s) in Luong, Pham and Manning (2015).

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Table 2. *Correlation scores of the vocabulary overlapping measures with the gold standard*

		M_{ef}	M_{fe}	M	M^c_{ef}	M^c_{fe}	M^c
G_a	TF-IDF	0.634	0.724	0.786	0.980	0.974	0.980
	P/A	0.897	0.770	0.936	0.976	0.966	0.972
G_b	TF-IDF	0.950	0.434	0.973	0.989	0.982	0.995
	P/A	0.955	0.190	0.979	0.975	0.977	0.978
G_c	TF-IDF	0.964	−0.292	0.962	0.980	0.934	0.991
	P/A	0.940	−0.595	0.960	0.984	0.968	0.990

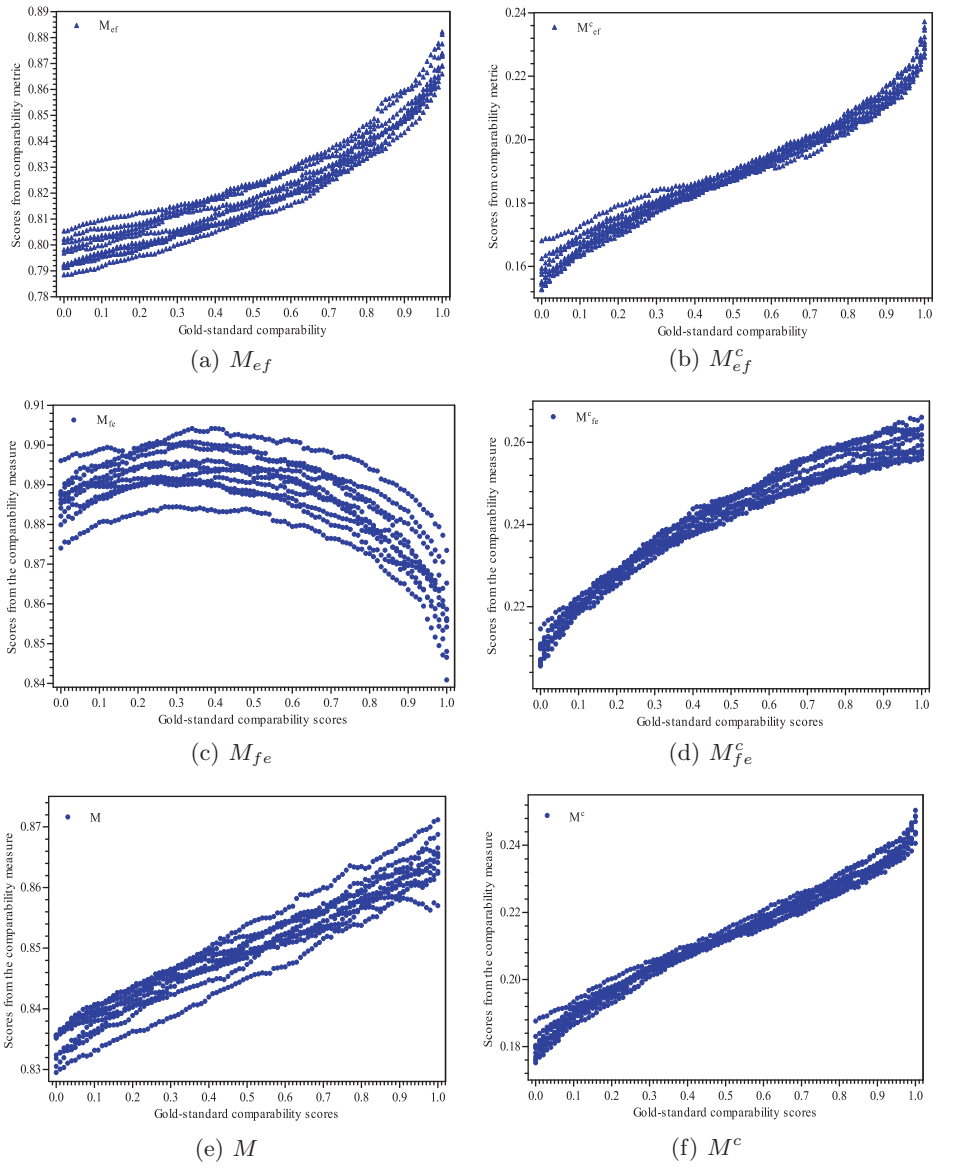
The rows TF-IDF correspond to the TF-IDF weighting schema, and the rows P/A correspond to the Presence/Absence weighting method.

We first make comparisons between measures based on vocabulary overlapping. The correlation scores are listed in Table 2. Each column in the table, e.g., M_{ef}, corresponds to the correlation scores between the specific comparability measure M_{ef} and the gold-standard comparability levels on different corpus groups, namely G_a, G_b and G_c. Let us first pay attention to the measures without context, i.e., M_{ef}, M_{fe} and M. One can find that M together with the P/A weighting schema performs the best and correlates very well with the gold standard on all the three groups of corpora, as the Pearson coefficient is close to 1. M_{fe} performs worst among the three measures with one exception: using TF-IDF weighting, M_{fe} performs better then M_{ef}. One can also conclude, for the best measure M, that using IDF to reduce the effects of frequent words does not help to solve the problem of too frequent words, and the P/A weighting schema works better with M than the TF-IDF one. Although the TF-IDF weighting schema seems to be efficient for M_{fe} on G_b and G_c, the measure M_{fe} together with TF-IDF still performs far from the other two measures.

We then turn to the measures with contextual information, i.e., M^c_{ef}, M^c_{fe} and M^c. One can find from Table 2 that all the context-based measures, weighted by either TF-IDF or P/A, perform very well, and are slightly better than the best performing measure M in the context-free family. Among the three measures with context, M^c performs slightly better than M^c_{ef} and M^c_{fe}. The measures M^c_{ef} and M^c_{fe}, with contextual information, are better than their corresponding context-free versions M_{ef} and M_{fe}. These findings are in agreement with the assumption that using contextual information helps to disambiguate translation candidates and thus leads to better performance of the measures.

Deeper analysis is given here in order to explain the different performance of the different measures. We will only consider here the P/A weighting scheme so as to simplify the discussion. Figure 2 plots the measures M, M_{ef}, M_{fe}, M^c, M^c_{ef}, M^c_{fe} on ten comparable corpora and their descendants in G_c with respect to their gold-standard comparability scores. We first compare here the three context-free measures M_{ef}, M_{fe} and M. One can notice from Figure 2(c) that the comparability scores from M_{fe} decrease at a certain point even though the gold standard scores increase. The reason for the different performances is that asymmetric measures M_{ef} and M_{fe} are sensitive to the relative length of the corpus. Given a single English

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press



(a) M_{ef} (b) M^c_{ef}

(c) M_{fe} (d) M^c_{fe}

(e) M (f) M^c

Fig. 2. (Colour online) Evolution of the measures M_{ef}, M_{fe}, M, M^v_{ef}, M^v_{fe} and M^v w.r.t. gold standard on the corpus group G_c (x-axis: gold-standard comparability scores, y-axis: comparability scores from the measures). (a) M_{ef}. (b) M^c_{ef}. (c) M_{fe}. (d) M^c_{fe}. (e) M. (f) M^c.

document and a large French document collection, it is very likely that one can find the translations for most of the English words, even though the two corpora are lowly comparable, because the English vocabulary in consideration is very small and it is easy to find all the translations even if the French corpora are not really comparable to the English content. In our case, since the average sentence length in *AP* is larger than that of *Europarl*, we increase the length of the English part of the test corpora remarkably when degrading the corpora in G_b and G_c, leading to poor performance of M_{fe}. In order to further support our judgement, we have also tried

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Table 3. *Correlation scores of the baseline comparability measures with the gold standard*

	M^v	M^{g1}	M^{g2}	M^s
G_a	0.698	0.724	0.492	0.824
G_b	−0.611	0.479	0.228	0.762
G_c	−0.744	0.311	0.210	0.703

to manually control the sentence length in the AP corpus, namely that we intend to choose in AP those sentences of which the length is similar to that in the original corpus. The results then show that M_{fe} works well under the new settings without a biased length. The length related problem can be overcome by M that considers the translation in both directions.

We now turn to the contextual versions M^c_{ef}, M^c_{fe} and M^c. All of the three measures perform very well on the three groups of corpora. Let us pay attention to the measure M_{fe} and its context version M^c_{fe}. The former one is sensitive to the corpus length, whereas the latter one is far less sensitive. We conjecture here that contextual information helps identify the correct translations in cases where they are a lot of possible translations (as is the case with unbalanced, in terms of size, collections), and thus can still capture different comparability levels. Furthermore, as the results show, M and all the context-based measures are able to capture all the differences in comparability artificially introduced in the degradation process we have considered in Section 3.2.1. Last, one can conclude from the results that it is easier to capture the different comparability levels in G_b than in G_a and G_c. This is due to the fact that the differences in G_a (based on the same corpus) are less marked, and thus more difficult to identify. On the opposite, G_c comprises corpora with low levels of comparability, also more difficult to separate out. G_b based on a parallel corpus with additions from a different corpus, displays comparability levels easier to identify.

We finally list in Table 3 the results from the baseline measures. The results shown here are obtained in the same way as the ones for the vocabulary overlapping measures in Table 2. From the results one can find that the recently introduced bilingual skip-gram measure M^s performs the best on all the three groups of corpora, the measures M^{g2} and M^v performing worse. All these baseline measures do not perform as well as the measures M (weighed by P/A) or the three context-based measures only relying on the vocabulary overlapping approach. The standard vector space is broadly used to represent textual elements in previous studies aiming at capturing similarity at the sentence or document levels. What our experiments reveal here is that this approach does not yield interesting performance (compared to other approaches) for measuring comparability on the corpus level. The reason is probably that noise involved in dictionary-based mapping in equation (8) destroys the capability of monolingual vector space model. This finding is also partially supported by a recent report of the ACCURAT project[13] that vector representation

[13] Related materials can be found on the deliverables of the project. The project website is: http://www.accurat-project.eu.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Table 4. *Real world comparable corpora and experimental results*

Group no.	First corpus	Second corpus	Results
1	Wiki-En+Wiki-Fr	Wiki-En+MON94	Y
2	Wiki-En+Wiki-Fr	Wiki-En+SDA94	Y
3	Wiki-En+Wiki-Fr	LAT94+Wiki-Fr	Y
4	Wiki-En+Wiki-Fr	GH95+Wiki-Fr	Y

One could conjecture that the first corpus should be more comparable than the second one in each group according to the intuition, which constitutes the gold standard. The last column corresponds to the experiment results where Y denotes a coherence with gold standard.

is an appropriate method for capturing document-level similarity in a cross-language environment. The measure based on cross-lingual word embeddings is inferior to the measure M, which is probably due to the same reason that it is weak at measuring similarity on corpus level. We will give more discussion on these measures in Section 4.

Artificial test corpora built using the above methodology may not fully reflect characteristics of all comparable corpora, although it is the most direct approach by which one could obtain quantitative levels. In addition to those experiments, we make use here of the comparability measure on *real world* comparable corpora without quantitative levels. The pairs of comparable corpora we will consider are listed in Table 4, which are in accordance with one's direct intuition. In comparison to artificial corpora constructed in Section 3.2.1, gold-standard comparability levels in Table 4 are qualitative rather than quantitative.

We use the best-performing measure M to test if the differences between each pair of comparable corpora in Table 4 can be differentiated. The results are listed in the last column of Table 4 where Y denotes experimental results agree with one's intuition (i.e., the gold standard). One can find from the results that all the comparability levels can be captured correctly by the measure M, which further proves the reliability of M in real world cases.

3.2.3 *Robustness of comparability measures*

Since the two measures M and M^c perform the best in their respective classes of measures, i.e., measures without context and measures with context, we choose to compare them in terms of the robustness w.r.t. the change of the dictionary that is the only resource used to cross the language barrier. To simplify the discussion here, we follow the same consideration as in Section 3.2.2 and only use the P/A weighting schema. It is important that the comparability measure one retains remains consistent when the dictionary coverage of the corpus changes slightly, as this is a necessary condition to distinguish between different comparability levels. Indeed, if a slight change in the dictionary coverage entails an important change in the comparability score, then it becomes impossible in practice to compare different corpora, as they will likely have a different coverage with respect to the dictionary.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

We say, informally, that a comparability measure is robust at certain dictionary coverage range if the measure can distinguish between different comparability levels when the dictionary changes in this range. The experiments and analysis below try to assess the robustness of the retained measures.

To do so, several dictionaries of different sizes, corresponding to different coverages on the corpus vocabulary, are built by randomly choosing the subparts from the original dictionary. The coverage here is simply defined as the proportion of unique words in the corpus vocabulary that are covered by the dictionary. This definition is coincident with the definition of M and M^c, which corresponds to the proportion of words translated in the part of vocabulary covered by the dictionary in the whole vocabulary. In order to bridge the language barriers, a dictionary that is sufficient large is necessary. We thus choose to randomly pick certain proportions, from 50% to 99% with a step of 1%, of our original dictionary. For each proportion, 30 different dictionaries are built by randomly sampling the original dictionary 30 times at this same proportion. These 1,500 dictionaries (i.e., 50×30) are then used to compute M and M^c on different corpora with decreasing comparability scores in G_a, G_b and G_c.

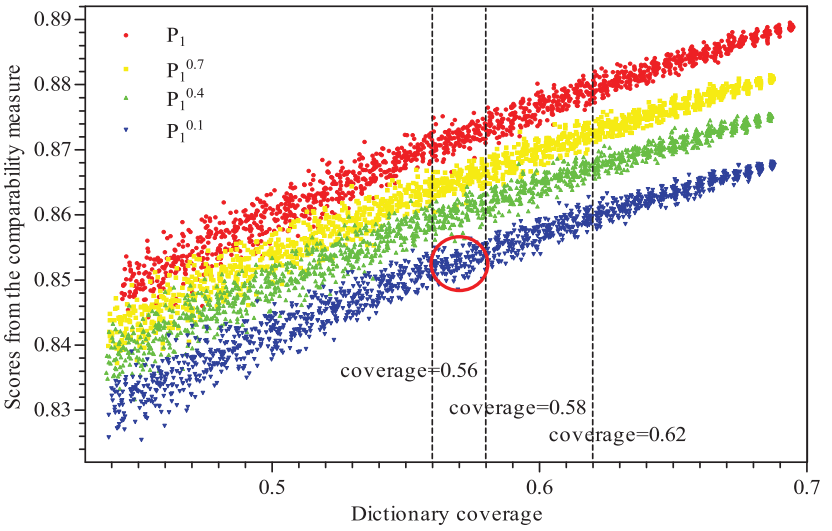
As we have discussed in the corpus construction process in Section 3.2.1, in each of G_a, G_b and G_c, one tries to obtain a series of decreasing comparability levels starting from ten parallel corpora (in G_a and G_b) or ten high quality comparable corpora (in G_c). In order to estimate the robustness, it is natural to perform several times the experiments in Section 3.2.2 together with various dictionaries. Then one can judge the robustness directly from correlation scores with different dictionaries. The problem exists in the fact that the dictionary coverage might be different on various test corpora, which prevents us from drawing a clear conclusion based on a uniform coverage level.

We will design here another set of experiments to clarify the impact of dictionary coverage on measure performance. For the clarity of analysis, we only take here the first parallel corpus $\mathcal{P}_1$ from $(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{10})$ together with its ten descendant corpora that are built by setting the proportion p to 0.1, 0.2, ..., 1.0 in G_a. That is to say, we exchange 10%, 20%, ..., 100% of the content in the high quality corpora with noise from other corpora. In this case, we obtain eleven comparable corpora $\mathcal{P}_1$, $\mathcal{P}_1^{0.9}$, $\mathcal{P}_1^{0.8}$, ..., $\mathcal{P}_1^{0}$ with the gold comparability scores from 1 to 0, with a step of 0.1. Last, for readability reasons, we only plot in Figure 3 the comparability scores for some of the eleven comparable corpora, i.e., $\mathcal{P}_1$, $\mathcal{P}_1^{0.7}$, $\mathcal{P}_1^{0.4}$ and $\mathcal{P}_1^{0.1}$, w.r.t. the different coverages (but the results are the same for the other corpora).

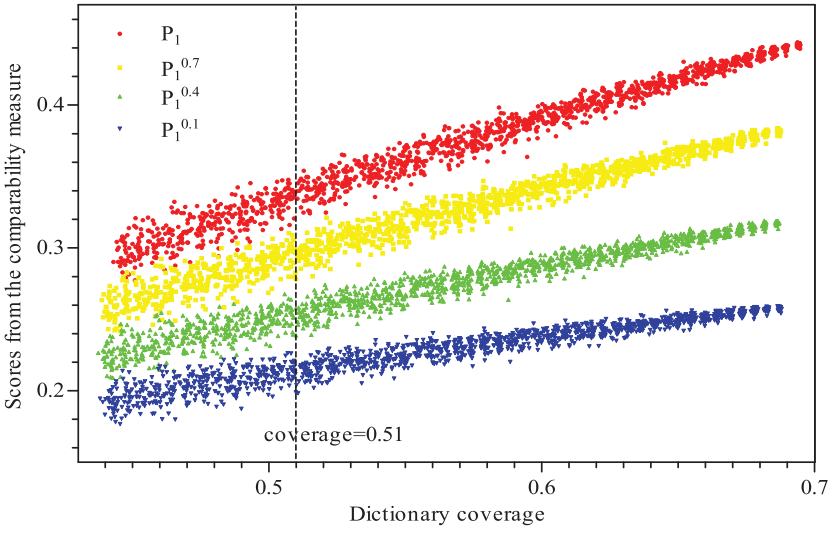
From Figure 3(a), one can find that when the dictionary coverage lies above a certain threshold (inspected from the figure, roughly set to 0.62), the differences between the four different comparability levels[14] can be captured very well, as the different data points are well separated. In other words, the different comparability levels can be captured very well by the measure M when the dictionary

[14] This is also true for all the eleven comparability levels, although we only plot 4 in the figure.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press



(a) M



(b) M^c

Fig. 3. (Colour online) Evolution of M and M^c w.r.t. different dictionary coverages on comparable corpora $\mathcal{P}_1$, $\mathcal{P}_1^{0.7}$, $\mathcal{P}_1^{0.4}$ and $\mathcal{P}_1^{0.1}$ in G_a (x-axis: different dictionary coverages; y-axis: comparability scores from M or M^c). (a) M. (b) M^c.

coverage is roughly above 0.62. The same conclusion can be drawn from the inspection of Figure 3(b). One can thus conclude from this qualitative analysis that both M and M^v are robust to the changes of the dictionary after a certain point.

We have drawn for comparability measures the intuitive conclusion regarding robustness from Figure 3. In order to analyze the results quantitively, we will first try to define the *degree of robustness* of a comparability measure.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

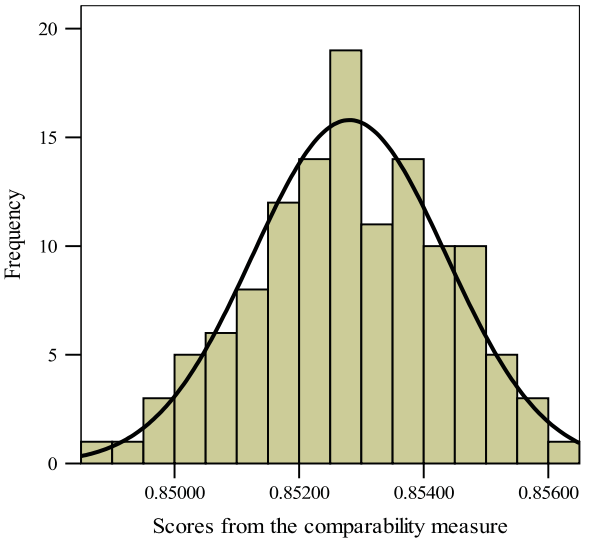


Fig. 4. (Colour online) The frequency histogram of the comparability scores (from M) on $\mathcal{P}_1^{0.1}$ between the coverage 0.56 and 0.58 (x-axis: comparability scores; y-axis: frequency), with the associated normal approximation.

***Definition** 1*

Let us assume that we have different comparable corpora $\mathcal{C}_1$, $\mathcal{C}_2$, …, $\mathcal{C}_k$, with gold-standard comparability levels that are increasing, which can be written as: $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \ldots \prec \mathcal{C}_k$ (the symbol $\prec$ is used here to denote the relation less *in the gold-standard comparability levels). We further assume to have a bilingual dictionary $\mathcal{D}$ such that the coverage of $\mathcal{D}$ on all the k corpora $\mathcal{C}_i (i = 1, 2, \ldots, k)$ belongs to a range $[r, r+\epsilon]$, with ϵ being a small fixed value. Then, we define the degree of robustness of a comparability measure M w.r.t the dictionary coverage r as*

$$\chi(M, r) = \underset{i \in 1,2,\ldots,k-1}{avg} Pr(M(\mathcal{C}_{i+1}) > M(\mathcal{C}_i))$$

Through the above definition, one measures to which extent a certain measure, associated with a certain coverage value r, can well separate different comparability levels in the average case.

We now turn to the problem of estimating the degree of robustness of the comparability measures M and M^v. Let us first focus on the measure M. We notice from Figure 3 that, for each of the test corpora, e.g., $\mathcal{P}_1^{0.1}$ in Figure 3(a), the comparability scores corresponding to a certain coverage range (e.g., from 0.56 to 0.58, identified by the circle in the figure) follow a normal distribution, according to the Shapiro–Wilk test (Shapiro and Wilk 1965) at the significance level 0.05. This fact is also illustrated by the frequency distribution histogram in Figure 4. Thus, in a coverage range of size 0.02 (i.e., the value ϵ in the Definition 1), the comparability scores of M for a specific corpus can be modeled as a normally distributed variable Z. Hence, on each span, the scores of M on two bilingual corpora, say $\mathcal{P}_1^{0.1}$ and $\mathcal{P}_1^{0.2}$, can be described as two normally distributed variables denoted as $Z_{0.1}$ and $Z_{0.2}$ of

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Table 5. *The degree of robustness with two measures M, M^c and different dictionary coverages on all the eleven comparability levels*

r	0.46	0.48	0.50	0.52	0.54	0.56
$\chi(M,r)$	0.70	0.69	0.73	0.76	0.79	0.83
$\chi(M^c,r)$	0.81	0.83	0.86	0.89	0.93	0.92
r	0.58	0.60	0.62	0.64	0.66	0.68
$\chi(M,r)$	0.82	0.85	0.90	0.92	0.95	1.00
$\chi(M^c,r)$	0.95	0.97	0.98	0.99	1.00	1.00

The coverage range is set to $[r, r+\epsilon]$ as in Definition 1 with ϵ being fixed to 0.02.

which the parameters (i.e., the mean μ and variance σ^2) can be estimated from the samples, i.e., all the comparability scores on $\mathcal{P}_1^{0.1}$ and $\mathcal{P}_1^{0.2}$ in the specific coverage range. With the estimated parameters, we can write:

$$Z_{0.1} \sim N(\mu_{0.1}, \sigma_{0.1}^2), \;\; Z_{0.2} \sim N(\mu_{0.2}, \sigma_{0.2}^2)$$

To compute the degree of robustness in Definition 1, one needs to obtain all the probability $\Pr(M(\mathcal{C}_{i+1}) > M(\mathcal{C}_i))$. Taking here the corpora $\mathcal{P}_1^{0.2}$ and $\mathcal{P}_1^{0.1}$ as an example, one needs to estimate $\Pr(Z_{0.2} > Z_{0.1})$. With the independence assumption between variables $Z_{0.1}$ and $Z_{0.2}$, the new variable $Z_{0.2} - Z_{0.1}$ satisfies:

$$Z_{0.2} - Z_{0.1} \sim N(\mu_{0.2} - \mu_{0.1}, \sigma_{0.1}^2 + \sigma_{0.2}^2)$$

and computing $\Pr(Z_{0.2} > Z_{0.1})$ amounts to the computation of $\Pr(Z_{0.2} - Z_{0.1} > 0)$, which can be done through tabulations of the normal distribution. In Table 5, we list, for different coverage ranges, the robustness values for both comparability measures M and M^c on all the eleven comparability levels. Let us first pay attention to robustness of the measure M. One can find from the results that the higher the dictionary coverage is, the more reliable M is to distinguish between different comparability levels. Furthermore, when the dictionary coverage is above a certain threshold (e.g., 0.60), we have a high confidence (≥ 0.85) that the different comparability levels between the corpus pairs can be reliably captured by M, so that the measure is robust in a given coverage range, here set to 0.02. The same conclusions can be drawn for the other comparable corpus pairs we have constructed (from G_b and G_c).

One can further conclude from Table 5 that, compared to M, it is easier to achieve high confidence, and thus higher robustness, with M^c. Even when the dictionary coverage is only 0.50, the confidence is larger than 0.86 that M^c captures the different comparability levels. However, the computation complexity of the context-based measures is usually very expensive. For this reason, we will only make use, in the remainder of this study, of the measure M weighted by the P/A criterion.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

3.3 Experiments of bilingual lexicon extraction

In addition to experiments in Section 3.2, we will use here the measure M in a real world task to show its efficiency. We will make use of bilingual lexicon extraction to show that comparable corpora of higher quality, measured by M, truly result in better lexicons extracted.

For bilingual lexicon extraction, we will make use of the approach developed in Li and Gaussier (2010) to produce several comparable corpora of different quality. This approach consists of two steps: (1) extract a high-quality subpart from the *original corpus*; (2) enhance the low-quality subpart, the part left in the original corpus after removing the high-quality subpart, with the *external corpus*. We do not plan to go into the details of this algorithm which is not the focus of the paper. Two comparable corpora are needed in order to use the algorithm in Li and Gaussier (2010):

- Original corpus: the comparable corpus of which the quality to be improved;
- External corpus: the comparable corpus used to enhance the low-quality subpart of the original corpus.

The test corpora here are built in a manner different from that used to evaluate the comparability measures in Section 3.2.1, intending to show that the measure M is not over-fitting test corpora artificially introduced in Section 3.2.1.

Corpora used in this part have been introduced in Table 1 in Section 3.1. We use here *GH95* and *SDA95* as the original corpus $\mathcal{C}^0$. In order to enhance the low-quality subpart, we consider two external resources: (a) $\mathcal{C}_T^1$ made of *LAT94*, *MON94* and *SDA94*, and (b) $\mathcal{C}_T^2$ consisting of *Wiki-En* and *Wiki-Fr*. We run the algorithm in Li and Gaussier (2010) with two groups of input corpora, namely $\mathcal{C}^0 + \mathcal{C}_T^1$ and $\mathcal{C}^0 + \mathcal{C}_T^2$, resulting in two corpora $\mathcal{C}^1$ and $\mathcal{C}^2$, respectively. Considering the comparability scores by M, the comparability of $\mathcal{C}^1$ is 0.912 and the one of $\mathcal{C}^2$ is 0.916. Both of them are more comparable than the original corpus $\mathcal{C}^0$ of which the comparability is 0.882.

In order to measure the performance of the bilingual lexicon extraction on $\mathcal{C}^0$, $\mathcal{C}^1$ and $\mathcal{C}^2$, we follow the standard practice in previous work and rely on the approach proposed in Fung and Yee (1998). In this approach, each word w is represented as a context vector consisting of the weight $a(w^c)$ of each context word w^c, the context being extracted from a window running through the corpus. Once context vectors for English and French words have been constructed, a general bilingual dictionary $\mathcal{D}$ can be used to bridge them by accumulating the contributions from words that are translation of each other. Standard similarity measures, as the cosine or the Jaccard coefficient, can then be applied to compute the similarity between vectors.

We then divide the original dictionary in Section 3.1 into two parts: 10% of the English words (3k words) together with their translations are randomly chosen and used as the evaluation set, the remaining words being used to compute context vectors and similarity between them. For each English word in the evaluation set, all the French words in the corpus are then ranked according to their similarity with the English word. To evaluate the quality of the lexicons extracted, we first retain for each English word its N first translations, and then measure the precision of the lists

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Table 6. *Precision of bilingual lexicon extraction on different corpora*

	$\mathcal{C}^0$	$\mathcal{C}^1$	$\mathcal{C}^2$
W_L	0.114	0.136	0.181
W_M	0.233	0.345	0.401
W_H	0.417	0.568	0.633
All	**0.205**	**0.258**	**0.310**

The comparability scores (by M) of $\mathcal{C}^0$, $\mathcal{C}^1$ and $\mathcal{C}^2$ are 0.882, 0.912 and 0.916, respectively.

obtained, which amounts in this case to the proportion of lists containing the correct translation (in case of multiple translations, a list is deemed to contain the correct translation as soon as one of the possible translations is present). This evaluation procedure has been used in previous work (e.g., Gaussier *et al.* (2014)) and is now standard for the evaluation of lexicons extracted from comparable corpora. In this study, we consider bilingual lexicon extraction with potential usage in CLIR and set N to 20 following previous studies as Gaussier *et al.* (2014). Furthermore, several studies have shown that it is easier to find the correct translations for frequent words than for infrequent ones (Pekar *et al.* 2006). To take this fact into account, we distinguished different frequency ranges to assess the impact of corpus quality measured by M for all frequency ranges. Words with frequency less than 100 are defined as low-frequency words (W_L), whereas words with frequency larger than 400 are high-frequency words (W_H), and words with frequency in between are medium-frequency words (W_M).

The results obtained are displayed in Table 6. They show that the standard approach performs significantly better on the improved corpora $\mathcal{C}^1/\mathcal{C}^2$ than on the original corpus $\mathcal{C}^0$. The overall precision is increased by 5.3% on $\mathcal{C}^1$ (corresponding to a relative increase of 26%) and 9.5% on $\mathcal{C}^2$ (corresponding to a relative increase of 51%). It should also be noticed that the performance of the standard approach is better on $\mathcal{C}^2$ than on $\mathcal{C}^1$. When considering different frequency ranges, one could come to the same conclusion. In a word, those results reveal that performance of lexicon extraction is coincident with the comparability scores computed for the three corpora, which further proves that the comparability measure M is able to quantify the corpus quality precisely.

4 Discussion

The quality of the comparable corpus is an important feature affecting its usability in NLP tasks. Our work presented in this paper is a first attempt that systematically investigates a set of approaches to measure the corpus quality. We would like to give additional comments on some parts of the work described in previous sections.

In order to evaluate the comparability measures, we have designed in Section 3.2 novel evaluation schema where corpora of gold-standard comparability levels are produced based on the parallel corpus and another monolingual corpus. We believe this design to be the most direct one that one can think of in order to obtain

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

quantitative comparability levels. We are aware of no former studies trying to build gold-standard comparability levels. With the test corpora obtained through this design, we found that the comparability measure *M*, based on vocabulary overlapping and the P/A weighting schema, can capture different comparability levels very well and is robust to changes in dictionary coverage. The measures we have developed could capture the differences between various corpora introduced manually.

As for the evaluation of proposed measures in Section 3.2.2, we have mostly cared about relative differences among various test corpora rather than the absolute comparability scores. The reason is that one only knows the relative rather than the absolute comparability levels of test corpora constructed by gradually introducing noise to existing corpora in Section 3.2.1. In addition, as we have discussed before, it is hard for people to assign precise comparability scores on corpus level than on sentence and document level.

The standard vector space approach for text representation has been used successfully in previous studies dealing with documents and sentences (Salton *et al.* 1975). It is however inferior, in the context of comparable corpora, to the simple measure *M* we have proposed. In our opinion, the cosine measure on top of vector representation is not appropriate to deal with long texts covering a broad collection of topics. Documents and sentences cover a single or only a few topics, making the vector representation comparable with each other, which is not the case for sets of documents. In addition, the direct vector mapping *via* bilingual dictionary brings in much noise affecting performance of the classic vector space model. This judgment has also been supported by such studies as Deshmukh and Hegde (2012).

We have also noticed that the MT-based measures do not perform well, which is contrary to our intuitions. The reason is the same as the above analysis: the vector space model is weak in dealing with text at corpus level. It is possible to alert the original VSM by considering the impact of text length. However, as we have shown, the simple measure *M* we have proposed shows quite satisfactory performance. There is not a strong drive to further explore the possibility of using VSM-based methods.

5 Conclusion

In this paper, we have first reviewed the notion of comparability in light of the usage of various bilingual corpora. This notion motivates the introduction of several comparability measures, as well as experiments designed to validate them. We find from those experiments that the measure M, together with the P/A weighting scheme, correlates well with gold-standard comparability levels and is robust to the dictionary coverage. Moreover, this measure has a low computational cost. The measure *M* is then validated in bilingual lexicon extraction to show its correlation with the performance of real tasks. We have made in this paper two main contributions: (1) A systematic approach to test if a comparability measure is reliable. We could not find in previous work such strategies that provide quantitative evaluation for a comparability measure, since it is difficult to construct comparable corpora of

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

known comparability levels. (2) A comparability measure that correlates well with gold-standard comparability levels and is robust to dictionary coverage.

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*, pp. 16–23.

Bahdanau, D., Cho K., and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, pp. 1–15.

Ballesteros, L., and Croft, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th ACM SIGIR*, Philadelphia, Pennsylvania, USA, pp. 84–91.

Blei, A., and Jordan, I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3**: 993–1022.

Boyd-Graber, J., and Blei, D. M. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-2009)*, pp. 75–82.

Chebel, M., Latiri, C., and Gaussier, E. 2017. Bilingual lexicon extraction from comparable corpora based on closed concepts mining. In *Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Jeju, Korea, pp. 586–598.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41**(6): 391–407.

Deshmukh, A., and Hegde, G. 2012. A literature survey on latent semantic indexing. *International Journal of Engineering Inventions* **1**(4): 1–5.

Fung, P., and Yee, L. Y. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Quebec, Canada, pp. 414–20.

Gaussier, E., Renders, J. M., Matveeva, I., Goutte, C., and Déjean, H. D. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp. 526–33.

Hazem, A., and Morin, E. 2016. Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pp. 3401–11.

Hermann, K. M., and Blunsom, P. 2014. Multilingual Models for Compositional Distributional Semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Maryland, USA, pp. 58–68.

Hewavitharana, S., and Vogel, S. 2008. Enhancing a statistical machine translation system by using an automatically extracted parallel corpus from comparable sources. In *Proceedings of the LREC 2008 Workshop on Comparable Corpora*.

Ji, H. 2009. Mining name translations from comparable corpora by creating bilingual information networks. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC-2009)*, pp. 34–7.

Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics* **6**: 97–133.

Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*.

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*, Beijing, China, pp. 644–52.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Li, B., Gaussier, E., and Aizawa, A. 2011. Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 473–8.

Luong, T., Pham, H., and Manning, C. D. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the NAACL Workshop on Vector Space Modeling for NLP*.

Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, G., Vassiliou, M., and Yannoutsou, O. 2006. Using patterns for machine translation. In *Proceedings of the European Association for Machine Translation*, pp. 239–46.

Mathieu, B., Besancon, R., and Fluhr, C. 2004. Multilingual document clusters discovery. In *Proceedings of RIAO*. pp. 116–25.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, pp. 664–71.

Munteanu, D. S., Fraser, A., and Marcu, A. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the HLT-NAACL 2004*, Boston, MA., USA, pp. 265–72.

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 the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, pp. 81–8.

Ni, X., Sun, J. T., Hu, J., and Chen, Z. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*. WWW '09, pp. 1155–6.

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

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*, pp. 311–8.

Pekar, V., Mitkov, R., Blagoev, D., and Mulloni, A. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation* **20**(4): 247–66.

Rapp, R. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, Maryland, USA, pp. 519–26.

Rayson, P., and Garside, R. 2000. Comparing corpora using frequency profiling. In *Proceedings of the ACL Workshop on Comparing Corpora*, pp. 1–6.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., and Utsuro, T. 2006. Compiling French-Japanese terminologies from the web. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, pp. 225–32.

Salton, G., Wong, A., and Yang, C. S. 1975. A vector space model for automatic indexing. *Communications of the ACM* **18**: 613–20.

Saralegi, X., SanVicente, I., and Gurrutxaga, A. 2008. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the 6th International Conference on Language Resources and Evaluations - Building and Using Comparable Corpora Workshop*.

Schmid, H. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.

Shapiro, S. S., and Wilk, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika* **52**(3): 591–611.

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

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press

Sharoff, S., Rapp, R., and Zweigenbaum, P. 2013. Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In S. Sharoff, R. Rapp, P. Zweigenbaum, P. Fung (eds.), *Building and Using Comparable Corpora*. Berlin: Springer-Verlag, pp. 1–17.

Skadina, I., Vasiljevs, A., Skadins, R., Gaizauskas, R., Tufis, D., and Gornostay, T. 2010. Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora (LREC-2010)*, pp. 6–14.

Talvensaari, T., Laurikkala, J., Järvelin, L., Juhola, M., and Keskustalo, H. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems* **25**(1): 4.

Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. 2016. Cross-lingual models of word embeddings: an empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 1661–1670.

Vulic, I., and Moens, M. F. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 719–725.

Washtell, J. 2009. Co-dispersion: a windowless approach to lexical association. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 861–9.

https://doi.org/10.1017/S1351324917000481 Published online by Cambridge University Press