Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms

Jingshu Liu[1,2], Emmanuel Morin[1], and Sebastián Peña Saldarriaga[2]

[1]LS2N - UMR CNRS 6004, Université de Nantes, France
[2]Dictanova, Nantes, France
[1]{jingshu.liu, emmanuel.morin}@ls2n.fr
[2]{jingshu, spenasaldarriaga}@dictanova.com

Abstract

Extracting a bilingual terminology for multi-word terms from comparable corpora has not been widely researched. In this work we propose a unified framework for aligning bilingual terms independently of the term lengths. We also introduce some enhancements to the context-based and the neural network based approaches. Our experiments show the effectiveness of our enhancements over previous works and that the system can be adapted in specialized domains.

Title and Abstract in French

Vers un système unifié pour l'extraction terminologique bilingue de termes simples et complexes

L'extraction d'une terminologie bilingue pour les termes complexes à partir de corpus comparables n'a pas été beaucoup étudiée. Dans ce travail, nous proposons un système unifié pour l'alignement des termes bilingues indépendamment de la longueur des termes. De plus nous introduisons également des améliorations aux approches basées sur l'alignement de contexte et sur un réseau neuronal. Nos expériences montrent l'efficacité de nos améliorations sur les travaux antérieurs, et le fait que le système peut être adapté en domaines de spécialité.

1 Introduction

Bilingual terminology extraction from comparable corpora has aroused a lot of attention since the 1990s (Fung, 1995; Rapp, 1999). Two classes of approach have been developed depending on the nature of the term to be aligned. The first class concerns the alignment of single-word terms using context-based or neural network approaches while the second attempts to align multi-word terms relying on compositional approaches. Few studies have focused on providing a unified framework for aligning single-word and multi-word terms, apart from Delpech et al. (2012) and Taslimipoor et al. (2016). The first requires some specific linguistic information like the morpheme translation table, which makes it difficult to employ for two languages from different linguistic families such as English and Chinese. The second incorporates word embeddings into their collocation alignment system but it is limited to several types of mapping that must match a set of pre-defined syntactic patterns. Our objective is to provide such a unified framework for aligning terms of variable length in specialized domains without specific linguistic knowledge of the source or target language.

Large size comparable corpora in specialized domains are not always available. Consequently, many data driven systems cannot learn from enough information. A possible solution to this problem is to associate external data such as general domain corpora (Hazem and Morin, 2016) to the specialized corpora. Our work adapts this method in order to improve the system performance.

Besides the enrichment of the data, our work focuses on studying and improving different state-of-the-art approaches both for single-word term and multi-word term alignments, which are finally unified into a single framework for alignment of terms of any length. We first describe the context-based projection

Proceedings of the 27th International Conference on Computational Linguistics, pages 2855–2866
Santa Fe, New Mexico, USA, August 20-26, 2018.

approach, and based on discussions in previous works, we introduce two enhancements which improve our final results. Secondly, we cover the bilingual word embedding approach which depends on vector representations of words that can be learned on large text collections using different neural networks (Bengio et al., 2003; Mnih and Hinton, 2008; Collobert and Weston, 2008; Mikolov et al., 2013b), then we also propose an enhancement that re-normalizes the word embedding vector length in order to improve the results.

After studying the approaches for single-word terms which are our fundamental components for processing multi-word terms, following the idea in Morin and Daille (2012), we propose a new system called Compositional Approach with Word Embedding Projection (CMWEP) to combine the advantages of the traditional compositional approach and the bilingual word embedding approach. Our final results show considerable improvements over the state-of-the-art approach for bilingual multi-word term extraction. Moreover, the proposed method is able to align variable length terms in a single process.

2 Single-Word Term Alignment

In this section, we describe the two principal state-of-the-art approaches used for bilingual lexicon extraction from comparable corpora. Furthermore we introduce our enhancements in order to overcome various limitations of these approaches. The reason why we want to study these approaches for single-word terms (SWTs) is that they are fundamental low-level elements in our approaches for multi-word terms (MWTs).

The two approaches for SWTs are known as distributional and distributed semantics (Hermann and Blunsom, 2014). Both use word vector representations. The vectors in the first approach are sparse, high dimensional and explicit (Levy and Goldberg, 2014). The vectors in the second one are dense, low dimensional and generalized. They both rely on the distributional hypothesis (Harris, 1968) which assumes that a word and its translation tend to appear in the same lexical contexts.

2.1 Context-Based Projection Approach

The historical context-based projection approach, also known as the standard approach (SA) has been studied in a variety of works (Fung, 1995; Rapp, 1999; Chiao and Zweigenbaum, 2002; Bouamor et al., 2013; Hazem and Morin, 2016; Jakubina and Langlais, 2017). To implement this approach, we first build a co-occurrence matrix for the source and target languages, where each line represents a context vector in an n-word window. These vectors are then normalized using the Mutual Information (MI (Fano, 1961)) for instance. Then we get the word in the target language vector space by projecting each element in the context vector via a bilingual seed lexicon. Finally, the candidate translations are ranked by calculating the similarity of the projected context vector with all the context vectors into the target language. We use the Cosine similarity measure as it is the one most used in previous works. In addition it enables us to parallelize the process of similarity comparison.

Due to the small size of specialized domain corpora, occurrences of words are not always statistically reliable. In order to improve word co-occurrence counts, Hazem and Morin (2016) show that using a general language corpus can significantly improve the standard approach results. They suggest two methods for exploiting external resources. The first adaptation called Global Standard Approach (GSA) consists in building the context vectors from a comparable corpus composed of the specialized and the general comparable corpora. We implement the second adaptation called Selective Standard Approach (SSA) which gives the best results in their experiments. Elements of this adaptation are defined as follows: Let S be the vocabulary of the specialized corpus, G the vocabulary of the general corpus, w the word to represent and c a context word that appears in the window around w such that :

$$\forall w \in S \cap G, \forall c \in S \cap G,\ cooc(w,c) = cooc_S(w,c) + cooc_G(w,c) \quad (1)$$

Distance-Sensitive Co-occurrence

In the standard approach, we note that some context words in the window are not effectively related to the central word. Usually the further the latter is away from a context word, the less they are semantically related. This effect is more obvious especially after stop word filtering. A word originally far from the central word can appear in the context window. This makes the context vector less relevant as a representation for the central word. To reduce this effect, we propose a weighted co-occurrence depending on the distance between the two words, denoted by Distance-Sensitive Co-occurrence (DSC):

$$DSC(w,c) = g(c|w) \times cooc(w,c) \quad \text{where} \quad g(c|w) = \Delta(w,c)^{-\lambda},\ \ \lambda \in [0,1] \tag{2}$$

where w and c respectively denote the central word and the context word, $g(c|w)$ the weight that is distributed to c as the context of w, Δ is the distance between the two words and λ a hyper-parameter that determines the degree of penalization for distant word pairs. Note that $\lambda = 0$ is equivalent to a uniform distribution.

Weighted Mutual Information

Another limitation in the standard approach with MI is that MI overestimates low counts and underestimates high counts. In order to overcome this drawback, we propose the Weighted Mutual Information (WMI) inspired by the work of Pennington et al. (2014) where they introduce a function to smooth word co-occurrences. The original function is a weight function which prevents the overestimation of word co-occurrences. We also use it as a weight function for MI:

$$WMI(w,c) = f(cooc(w,c)) \times MI(w,c) \tag{3}$$

$$f(x) = \begin{cases} (x/x_{max})^{\alpha}, \alpha = 3/4, x_{max} = 20, \text{ if } x < x_{max} \\ 1 \text{ otherwise} \end{cases} \tag{4}$$

We have kept the same value for the hyper-parameter α. Concerning the x_{max}, since our corpora size is much smaller than the one in their work, we decide to make it correspondingly smaller (20). By adding the weight function, the output value for low co-occurrence counts is in fact reduced and the high co-occurrence counts are not impacted because their weight coefficient is always 1.

2.2 Neural Network-Based Approach

The neural network based approach uses neural network models to obtain word representation in low dimensional and dense vectors. These word vectors are also called word embeddings (Mikolov et al., 2013b). In the case of bilingual word embedding, Mikolov et al. (2013a) propose a method to learn a linear transformation from the source language to the target language for the task of lexicon extraction from bilingual corpora. Much research has been focused on this area since then. Faruqui and Dyer (2014) introduce canonical correlation analysis (CCA) to project the embeddings in both languages to a shared vector space. Xing et al. (2015) propose the orthogonal transformation and the vector length normalization during the learning phase. Artetxe et al. (2016) generalize these works and explain the equivalence of different objective functions under orthogonality and different normalization procedures. They show that combining these models effectively improves the results for both monolingual and bilingual tasks. Finally Smith et al. (2017) point out that the mapping should be orthogonal in order to be self-consistent.

Normalization, Mean Centering and Orthogonal Mapping

The method of Artetxe et al. (2016)[1] combines several related studies in this particular order:

1. Normalizing each word vector to unit length.
2. Dimension-wise mean centering for source and target matrices.
3. Learning the transformation matrix by mathematical analysis while constraining the orthogonality of the transformation matrix. In the original work of Mikolov et al. (2013a), the matrix is learned by gradient descent which is an on-line method.

Renormalization After Mean Centering

We observe in the implementation above that once the dimension-wise averages are subtracted, each word vector is no longer unit normalized. As a consequence, this could lead to a violation against the

[1] `https://github.com/artetxem/vecmap`

initial intuition that each word should have the same importance for learning the transformation matrix. Therefore we propose a second normalization after the mean centering. Although this would make the expected random product of any dimension not strictly zero, our experiment results show that it is still more beneficial for the bilingual task. Intuition could be that the mean centering was more in consideration of monolingual tasks, whereas in bilingual tasks, the length normalization of each word vector outweighs the mean centering.

Concatenation for Usage of External Data

Word embedding systems usually need a large amount of data in order to obtain reliable word vectors. However the size of domain-specialized comparable corpora is generally very modest (fewer than one million words). The word embedding models trained on our small size specialized corpora are not capable of generalizing meaningful features. To overcome this problem, the idea exploited in Hazem and Morin (2016) which seeks external general data to enrich the training phase can be useful, but it occasionally makes the specialized word representation biased by the general corpus. This is especially the case when the specialized corpus contains some infrequent or ambiguous words that have different meanings in different corpora. Considering these factors, we propose to use a concatenated vector of the one trained on the specialized corpora and on the general corpora as our new word vector. More specifically, we want to concatenate a relatively small size word vector from the specialized corpora and a relatively large size one from the general corpora. In our experiment, the word vector size for the word embedding model trained on the specialized corpus is set to be 100 and the size for the model trained on the general corpus is set to be 300. Hence we preserve the specialized corpus features albeit with less corresponding weight features in the transformation matrix. Consequently the new concatenated word vector carries self-contained information from both corpora (Figure 2). Another advantage is that by doing the concatenation, it is not necessary to retrain our word embedding models over large corpora because we can use existing pre-trained models available. This could save a great deal of time in practice while improving efficiency.

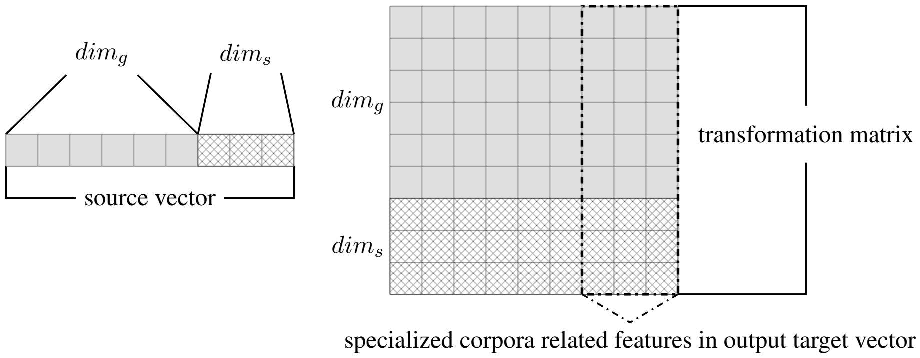

Figure 1: Let dim_g be the dimension size of the word vector from general corpora, and dim_s the size from specialized corpora. The transformation matrix will have a total number of $(dim_g + dim_s)^2$ weight features. Among these, $dim_g \times (dim_g + dim_s)$ will be source or target general corpora related features and only $dim_s \times (dim_g + dim_s)$ source or target specialized corpora related ones.

3 Multi-Word Term Alignment

Regarding MWTs, we first describe the compositional approach and an adapted version which combines the traditional compositional approach and the context-based projection approach. We also introduce a unified representation for MWTs which enables the mapping of variable length terms. Finally we propose a new approach in this family which combines the traditional compositional approach and the neural network approach. To the best of our knowledge, this is the first time that word embedding vectors are used in a compositional approach.

3.1 Compositional Approach

The compositional approach (CA) (Grefenstette, 1999; Tanaka, 2002; Robitaille et al., 2006) is a simple and direct approach that consists in translating each element of an MWT via a dictionary and generating all possible combinations and permutations. The ranking of the candidates is done by their frequency in the target corpora.

Compositional Approach with Context-Based Projection

The main limitation of the traditional compositional approach is the inability to translate a term when one of its composing words is not in the dictionary. To solve this problem, Morin and Daille (2012) propose the Compositional Approach with Context-Based Projection (CMCBP), where the objective is to combine the advantages of the standard and compositional approaches by substituting non-dictionary words with their context vectors obtained by the standard approach. The CMCBP begins by building the co-occurrence matrix as in the standard approach. Then it applies a direct translation reinforced by context alignment. If a word of a term to be translated is not present in the dictionary, it uses the context vector obtained by the standard approach and projects it into the target language, otherwise it takes the context vector of the target language directly. The next step is the generation of all combinations of possible translation representations for a source language term. Finally, the candidate terms are ranked according to their similarity with terms of the same length in the target language, and the final score for each possible translation is defined by the arithmetic or geometric mean of each similarity score.

MWT Representation

CMCBP does not, however, take the mapping of MWTs of variable lengths into account. For example, the English term "*wind vane*" can be translated as "*girouette*" in French and the English term "*wind energy*" by "*Windenergie*" in German. In order to take these cases into account, we propose to modify the representation of the MWTs, inspired by the works of Blacoe and Lapata (2012) in which the representation of a sentence is the sum of the distributional representations of each composing word:

$$vector(term) = \frac{1}{n}\sum_{i}^{n} \frac{vector(w_i)}{||vector(w_i)||}, \quad \text{where } n \text{ is the term length} \tag{5}$$

Notice that this is different from the mean vector introduced in the original work, here the mean vector is calculated from the normalized vectors because we want each component word to have the same impact rather than having the whole meaning influenced by the random vector length which could lead to some unpredictable bias. The MWT representation is then stored in a single vector, giving the ability to handle translations of different lengths while reducing the calculation time. In fact, in CMCBP, aligning an MWT requires calculating all possible permutations, so we must compare a factorial number of vectors for sorting candidates against a single vector with this new representation.

Compositional Approach with Word Embedding Projection

Following the idea of CMCBP, we propose a new method called Compositional Approach with Word Embedding Projection (CMWEP). It can be implemented by applying the following steps:

1. Prepare two word embedding models for the source and the target language with the same vector size.
2. Learn the transformation matrix using the approach we mentioned in 2.2.
3. Translate each word in an MWT via a bilingual seed lexicon. If a word is not in the dictionary, we return the embedding vector projected by the transformation matrix, and if the word is in the lexicon, we return the averaged vector of a list of embedding vectors for each possible translation. Giving an MWT "ABC" for instance, if "A" and "B" are in the lexicon and have their respective translations: $\{$"a_1", "a_2", "a_3"$\}$, $\{$"b_1", "b_2"$\}$ and "C" is not in the lexicon, then for "A" and "B", we return the averaged vector of the word embeddings list of $\{$"a_1", "a_2", "a_3"$\}$ and $\{$"b_1", "b_2"$\}$ directly by the lookup table of the target language word embedding system, for "C" we return its projected embedding vector using the learned transformation matrix. In this way each word in the MWT in the source language is represented by a single vector in the target language space.

4. Generate the representation vector for the whole MWT by applying the method in the section above since each composing word has been represented by one vector.
5. Compare the translated vector to each candidate in the target language using a similarity measure such as Cosine. The candidate translations are ranked according to the scores of the similarity measure.

Here we show an overview which illustrates how the CMWEP turns a MWT into a vector. In addition, as it shares the same idea as in CMCBP, this diagram also represents the process in CMCBP.

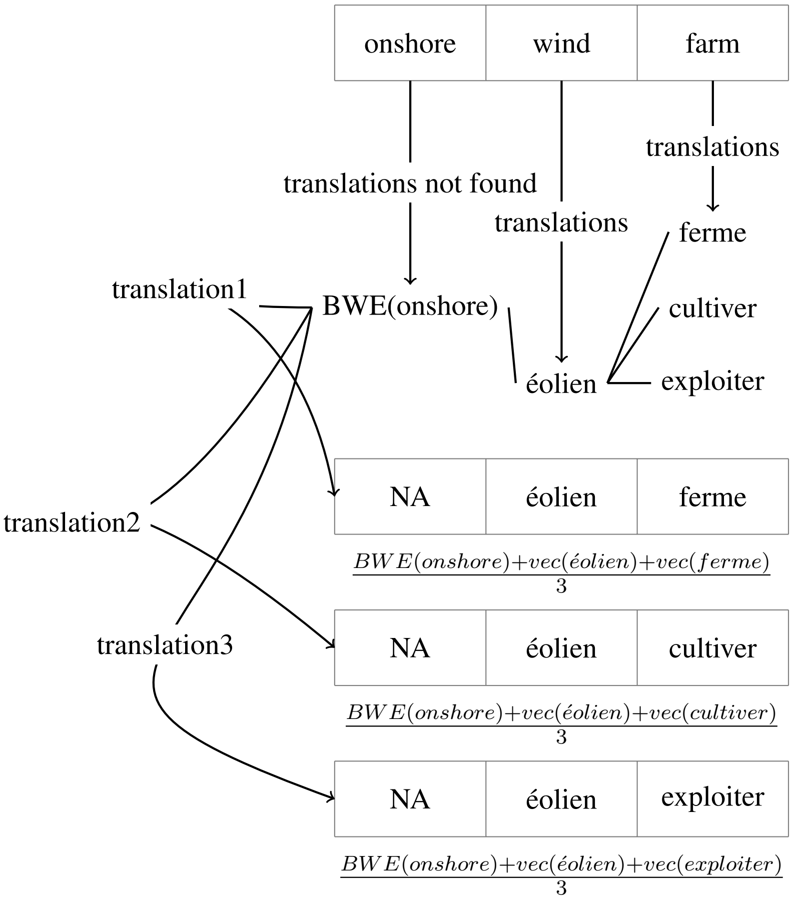

Figure 2: An example for translating the MWT *onshore wind farm* into French. The translations for each composing word are limited to those that exist in the comparable corpus and the bilingual dictionary. The final similarity between the MWT and a candidate is the maximum of the three similarities between each of the three possible representation vectors and the candidate vector.

4 Experiment Data and Resources

In this section, we outline the different textual resources used for our experiments: the comparable corpora, the reference lists, the bilingual seed lexicon and the pre-trained word embedding models.

4.1 Comparable Corpora

For our experiments, we used two specialized comparable corpora that have been used in previous works for bilingual terminology extraction in technical domains:

Breast Cancer Corpus (BC) is composed of documents collected from the Elsevier website[2]. The documents were taken from the medical domain within the subdomain of breast cancer.

[2] http://www.elsevier.com

Wind Energy Corpus (WE) has been released by the TTC project. The corpus has been crawled using the Babouk crawler (Groc, 2011) based on several keywords such as "*wind*", "*energy*", "*rotor*" in English and their translations in French.[3].

News Commentary (NC) consists of political and economic commentaries crawled from the web[4]. We use this corpus as our external data.

4.2 Gold Standard

Our reference lists for SWTs in BC and WE corpora are the same as used in Hazem and Morin (2016) which consist of 248 SWTs for BC and 139 for WE. The reference list for MWTs in WE is built based on the term list provided. Finally this list contains 73 MWT pairs but each pair has multiple variant translations and in our settings, we consider them to be also the gold translations[5]. If we deploy the 73 MWTs to a list where each pair contains only one translation, the list would have a size of 277 pairs. Because some MWTs have far more possible translations than others, we decide not to use the deployed version list to prevent the result from being biased by these MWTs. The reference list for the Italian/English task is the same as in Artetxe et al. (2016) which contains 1,500 entries. We would like to mention that the candidate list for one MWT includes all the words in the vocabulary plus all the MWTs extracted by a symbolic terminology extraction system, which generally extracts three times as many MWTs as words in the vocabulary. The terms are extracted following some pre-defined syntactic patterns, for example the pattern *NOUN NOUN* could lead to the extraction of *shop assistant*. So the one-to-many or many-to-one mapping is theoretically findable. Moreover, our reference list for the MWT task contains only out-of-dictionary MWTs. Table 1 presents the principal characteristics of the data.

Corpus	**# distinct words**		**# content words**		**Reference List**
	FR	EN	FR	EN	
BC	521,262	525,934	6,630	8,821	SWT: 248
WE	314,549	313,943	6,038	7,134	MWT: 73, SWT: 139
NC	5.7 M	4.7 M	23,597	29,489	

Table 1: Characteristics of comparable corpora in our experiments

4.3 Bilingual Lexicon

For our French/English experiments, we use the French/English dictionary ELRA-M00337[6] (243,539 entries). From this dictionary we select a subset of 3,007 entries from the BC corpora and a subset of 2,745 entries from the WE corpora based on a word frequency threshold of 5. These two subsets are used as the training data in our word embedding mapping experiments. For our Italian/English experiments, we only use the same seed lexicon[7] as used in Artetxe et al. (2016) where 5,000 entries are manually selected.

4.4 Pre-trained Word Embedding Models

To be fully comparable, in the Italian/English experiments we use the same model as in Artetxe et al. (2016). This could be retrieved by following the instructions they provide. The English embeddings were trained on a 2.8 billion word general corpus (ukWaC + Wikipedia + BNC)[8], while a 1.6 billion word general corpus itWaC was employed to train the Italian embeddings. As for the French/English embeddings, the vectors in dimension 300 were obtained using the skip-gram model described in Bojanowski et al. (2016) with default parameters[9].

[3] Corpus available here: `http://www.lina.univ-nantes.fr/?Reference-Term-Lists-of-TTC.html`

[4] `http://opus.lingfil.uu.se`

[5] The reference list and evaluation software are available here: `https://github.com/Dictanova/term-eval`

[6] `http://catalog.elra.info/product\_info.php?products\_id=666`

[7] `https://github.com/artetxem/vecmap`

[8] `http://clic.cimec.unitn.it/georgiana.dinu/down/`

[9] `http://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

5 Experiment Settings and Results

We conducted two sets of experiments. The first one aims at validating our enhancements to the state-of-the-art approaches by comparing the results of the bilingual SWT extraction in general and specialized domains. The second experiment aims at studying the application of the proposed approaches of the bilingual MWT extraction focusing on a specialized domain which is our main interest. This kind of task often lacks specialized data. As pointed out in Mikolov et al. (2013a), applications to low resource domains is a very interesting topic with much to be explored.

5.1 Standard Approach

In the standard approach, the window size is 3 (a total of 7 words are considered), which is the same as in Hazem and Morin (2016). The distance weight parameter λ in distance-sensitive co-occurrence is empirically set to be 0.25.

5.2 Word Embedding on Specialized Corpora

We use the implementation of *word2vec* from *deeplearning4j*[10] to train the word embedding model on our specialized corpora. Considering the relative low frequency of some words in specialized corpora, we apply the skip-gram algorithm which is supposed to work better with infrequent words[11]. We set the negative sample to 20, the window size to 5 and the training epoch to 20.

5.3 Results on SWT Task

Table 2 shows the results of the bilingual SWT extraction using the standard approach (SA) and its variant, the Selective Standard Approach (SSA) for the breast cancer corpus (BC) and the wind energy corpus (WE). It also includes the results of our two enhancements: Weighted Mutual Information (WMI) and Distance-Sensitive Co-occurrence (DSC).

Standard Approach	**BC**	**WE**
Hazem and Morin (2016)†	25.9	15.6
SA + WMI	28.9	21.4
SA + DSC	27.4	15.8
SA + WMI + DSC	**29.5**	**21.8**

(a) Result (MAP%) of standard approaches for the bilingual SWT extraction

Selective Standard Approach	**BC**	**WE**
Hazem and Morin (2016)†	56.5	44.4
SSA + WMI	55.0	33.9
SSA + DSC	**57.3**	**45.3**
SSA + WMI + DSC	55.8	35.7

(b) Result (MAP%) of selective standard approaches with NC corpus as the external data for the bilingual SWT extraction

Table 2: Context-Based Projection approaches results for bilingual SWT extraction. † indicates results obtained by our implementation of the approach.

We observe in Table 2a that WMI alone improves the results compared to those obtained by Hazem and Morin (2016) with MI but also compared to those when using the *Discounted Odds Ratio* (Evert, 2005) as the normalization method, where the MAP is 0.270 for BC and 19.4 for WE. This shows the interest of penalizing small occurrences to compensate the overestimation of the original MI. The second observation is that DSC alone also improves the results. This confirms our intuition that the further a context word is from its central word, the less relevant it is. Finally, we see that combining both enhancements gives the best result. So we could consider that the two enhancements are not mutually exclusive. Another observation is that the improvement of DSC for WE is indeed poorer than for BC, but the tendency is always improving. So we think our hypothesis above holds true for the WE data. Again we can see the same tendency when combining WMI. Note that, despite the difference in improvement between the two datasets, the enhancement that our approach offers is still relevant.

The results in Table 2b shows that WMI is less efficient when the data is enriched because the overestimation of small occurrences is smoothed by the addition of the enlarged overall data as discussed in

[10] `http://deeplearning4j.org/`

[11] `https://code.google.com/archive/p/word2vec/`

Hazem and Morin (2016). Moreover, since some term elements to be translated are quite infrequent or even non-existent in the general corpus, it is possible that penalizing all the small occurrences reduces discriminative features in the general corpus. Besides, although the results of SSA are different the two corpora share again the same tendency: they both reach their best when using DSC without WMI. However DSC always improves the results whether it is applied alone or combined with WMI. Again, it shows that the two enhancements are not mutually exclusive with external data.

Bilingual Word Embedding	**Accuracy**
Mikolov et al. (2013a)‡	34.93
Artetxe et al. (2016)‡	39.27
Our method (Renorm.)	**39.60**

(a) Accuracy % (p@1) of bilingual word embedding approaches for the Italian/English general word mapping task. The word embedding vectors are trained on wiki corpora mentioned in 4.4.

Bilingual Word Embedding	**BC**	**WE**
Hazem and Morin (2017)‡	82.3	-
L^2 + MC + Orth.	27.4	21.8
Concat + L^2 + MC + Orth.	82.4	**83.1**
Concat + L^2 + MC + Orth. + Renorm.	**83.2**	**83.1**

(b) Result (MAP%) of bilingual word embedding approaches for bilingual SWT extraction

Table 3: Bilingual Word Embedding approaches results for bilingual SWT extraction. ‡ indicates results reported by the authors.

Table 3 shows the results of the bilingual word embedding approaches on SWTs of the specialized BC and WE corpora along with the general WIKIPEDIA corpus. Our final proposed method is in the last line in each table. We can see that without external data, the results are not better than the traditional approach.

In Table 3a the three factors (denoted by L^2, *MC, Orth.*) resumed in Artetxe et al. (2016) improve the original results of Mikolov et al. (2013a). By adding an additional operation between the mean centering and the orthogonal mapping, our method brings a small but consistent improvement that can also be found in the special domain task in Table 3b. Other than that, the external data information stored in the pre-trained word embeddings significantly improves the results when carefully concatenated to those trained on special corpora.

5.4 Results on MWT Task

Model	**Accuracy**	**MAP**
CA	59.0	61.5
CMCBP	49.3	61.4
CMCBP + WMI + DSC	46.6	63.2
CMCBP + SSA	50.7	66.0
CMCBP + SSA + WMI + DSC	53.4	66.3
CMWEP + L^2 + MC + Orth.	52.1	67.8
CMWEP + L^2 + MC + Orth. + Concat.	**61.6**	73.3
CMWEP + L^2 + MC + Orth. + Concat. + Renorm.	**61.6**	**73.4**

Table 4: Accuracy (p@1) and MAP % of different approaches for MWTs in the WE corpus

Table 4 shows results with different approaches on MWTs in the special domain WE corpus. The traditional compositional approach (CA) has better accuracy than CMCBP because the candidates are ranked by their frequency in the target language corpus while the candidates in CMCBP are ranked by the similarity measure, so theoretically those translated by CA are also translated by CMCBP but with the same similarity score (1.0). Therefore, the final rank in CMCBP is to some degree randomized as multiple top candidates have the same score. Nevertheless, when using external data (+SSA), the MAP is considerably improved by CMCBP, this shows that the CMCBP can effectively find many out-of-dictionary translations that can not be found by CA. The same problem also exists in CMWEP.

Another remarkable point is that unlike the SWT bilingual extraction with SSA, combining WMI and DSC improves the results in MAP with or without external data.

Our word embedding methods (CMWEP) notably improve the results when using information carried by external data (+ *concat*) by almost 12 points in MAP and 2.6 points in accuracy. Combining the usage of external data and the renormalization gives the best result. The characteristics of the WE corpus determine the fact that for most MWTs, each component word are included in the dictionary. So CA already provides a high result especially for the accuracy, if we took a less academic or less cleaned corpus for example the Amazon Reviews where there are more out-of-vocabulary words because of spelling mistakes and web languages, we expect the result of CA to be much lower while the result of other approaches would be less impacted. This is also why the renormalization improves the result very slightly, since most of the translations are projected by a CA-like process.

5.5 Error analysis

The error analysis is made by manual observation: we first extract the 28 MWTs for which no translation has been found in the in the top@1 of our best performing model (the last line in Table 4), then we study these MWTs one by one. Among the 28 MWTs, we observe three categories:

1. **Weak compositionality**. The translation of the whole is not a combination of all the translations for each element (Melamed, 2001). 10 MWTs (35.7%) belong to this type. For example "*pitch angle*" is translated as "*angle d'inclinaison* (lit. *angle of tilt*)" by the system while the gold one should be "*angle de calage* (lit. *angle of sitting*)" where the word "*calage*" is not the translation of any word in the bilingual dictionary. Similarly, the gold translation for "*rotor shaft*" is a single word "*arbre*" which is basically the translation of one composing word "*shaft*". This problem could be even worse between two further languages, Baldwin and Tanaka (2004) report that at least 50% of the Japanese *NN* compounds are not translated through a compositional strategy into English.
2. **Ambiguous translations**. Multiple possible translations have the same similarity at the top@n list because each composing word has multiple translations in the dictionary and some combinations exist in the terminology candidate list. 12 MWTs (42.9%) belong to this type. For example, "*low frequency*" is translated by "*rotation inférieure* (lit. *inferior rotation*)" while it should be "*basse fréquence* (lit. *low frequency*)". Both "*inférieure*" and "*basse*" are translation of "*low*", and "*rotation*" and "*fréquence*" are translation of "*frequency*". This problem also relates to the monolingual word disambiguation.
3. **Different orders with same words**. Since our method ignores the word order, two sequences of the same set of words have the same representation vector. 6 MWTs (21.4%) belong to this type. For example, "*power installation*" is translated as "*puissance d'installation* (lit. *power of installation*)" while it should be "*installation de puissance* (lit. *installation of power*)". We could have taken the word order into consideration but it would increase a huge amount of calculation time. Moreover, the alignment of variable length terms would not be manageable.

6 Conclusion and future work

This work proposes a unified framework for bilingual terminology extraction of multi-word terms, improving on several previous works. Our experiments demonstrate the effectiveness of our enhancements of previous works. Generally speaking, the bilingual word embedding method with external data information plus the three factors resumed in Artetxe et al. (2016) and the renormalization outperforms previous works and brings us the best result. Moreover we observe that without external information, the traditional approach with our proposed enhancements still has very competitive results. For future work, we would like to restructure the representation of MWTs which is now based on a naive hypothesis that all the notional words contribute equally to the whole meaning of an MWT. In addition, as we mentioned in section 5.5, the performance for two further languages would be interesting to see and the disambiguation for words with multiple possible translations could be very useful.

Acknowledgments

The authors would like to thank Joseph Lark for feedback on earlier version of this paper, and the anonymous reviewers for their valuable comments. Jingshu Liu is partially supported by the CIFRE grant N 2016/0659.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 2289–2294, Austin, Texas.

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain.

Yoshua Bengio, R'ejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, pages 1137—1155.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP'12)*, pages 546–556, Jeju Island, Korea, July. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 759–764, Sofia, Bulgaria.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1–5, Stroudsburg, PA, USA.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pages 160–167, New York, NY, USA.

Estelle Delpech, Béatrice Daille, Emmanuel Morin, and Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In *Proceedings of the 24rd International Conference on Computational Linguistics (COLING'12)*, pages 745–762, Mumbai, India.

Stefan Evert. 2005. *The statistics of word cooccurrences : word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

Manaal Faruqui and Chris Dyer. 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 (EACL'14)*, pages 462–471, Gothenburg, Sweden.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora (VLC'95)*, pages 173–183, Cambridge, MA, USA.

Gregory Grefenstette. 1999. The world wide web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer 21*, London, UK.

Clément De Groc. 2011. Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 497–498.

Zellig Sabbettai Harris. 1968. *Mathematical structures of language*. Interscience Publishers.

Amir Hazem and Emmanuel Morin. 2016. Efficient data selection for bilingual terminology extraction from comparable corpora. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pages 3401–3411, Osaka, Japan.

Amir Hazem and Emmanuel Morin. 2017. Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP'17)*, pages 685–693, Taipei, Taiwan.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 58–68, Baltimore, Maryland.

Laurent Jakubina and Phillippe Langlais. 2017. Reranking translation candidates produced by several bilingual word similarity sources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 605–611, Valencia, Spain.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (ACL'14)*, pages 171–180, Ann Arbor, Michigan.

I. Dan Melamed. 2001. *Empirical Methods for Exploiting Parallel Texts*. MIT Press, Cambridge, MA, USA.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIP'13)*, pages 3111–3119, Lake Tahoe, Nevada, USA.

Andriy Mnih and Geoffrey Hinton. 2008. A scalable hierarchical distributed language model. In *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS'08)*, pages 1081–1088, Vancouver, British Columbia, Canada.

Emmanuel Morin and Béatrice Daille. 2012. Revising the compositional method for terminology acquisition from comparable corpora. In *Proceedings of the 24rd International Conference on Computational Linguistics (COLING'12)*, pages 1797–1810, Mumbai, India.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1532–1543, Doha, Qatar.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, Maryland, USA.

Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, and Takehito Utsuro. 2006. Compiling French-Japanese terminologies from the web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 225–232, Trento, Italy.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations (ICLR'17)*.

Takaaki Tanaka. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1–7, Stroudsburg, PA, USA.

Shiva Taslimipoor, Ruslan Mitkov, Gloria Corpas Pastor, and Afsaneh Fazly. 2016. Bilingual contexts from comparable corpora to mine for translations of collocations. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'16)*, Konya, Turkey.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'15)*, pages 1006–1011, Denver, Colorado.