Leveraging Meta-Embeddings for Bilingual Lexicon Extraction from Specialized Comparable Corpora

Amir Hazem **Emmanuel Morin**

LS2N - UMR CNRS 6004, Université de Nantes, France
{amir.hazem,emmanuel.morin}@univ-nantes.fr

Abstract

Recent evaluations on bilingual lexicon extraction from specialized comparable corpora have shown contrasted performance while using word embedding models. This can be partially explained by the lack of large specialized comparable corpora to build efficient representations. Within this context, we try to answer the following questions: First, (i) among the state-of-the-art embedding models, whether trained on specialized corpora or pre-trained on large general data sets, which one is the most appropriate model for bilingual terminology extraction? Second (ii) is it worth it to combine multiple embeddings trained on different data sets? For that purpose, we propose the first systematic evaluation of different word embedding models for bilingual terminology extraction from specialized comparable corpora. We emphasize how the character-based embedding model outperforms other models on the quality of the extracted bilingual lexicons. Further more, we propose a new efficient way to combine different embedding models learned from specialized and general-domain data sets. Our approach leads to higher performance than the best individual embedding model.

1 Introduction

Bilingual lexicons are fundamental resources in multilingual natural language processing tasks such as machine translation (Och and Ney, 2003), cross-language information retrieval (Nie, 2010) or computer-assisted translation (Delpech, 2014). Because a manual compilation of bilingual lexicons requires substantial human efforts, bilingual lexicons are automatically extracted from bilingual corpora. These corpora can be parallel or comparable data sets. Despite good results obtained when compiling bilingual lexicons from parallel corpora, the latter are scarce resources, especially for specialized and technical domains and for language pairs not involving English. In this context, comparable corpora are an interesting and practical alternative to the use of parallel corpora.

Comparable corpora, which gather texts sharing common features such as domain, topic, discourse, etc. without having a parallel source text-target text relationship, allow access to the original vocabulary without falling under the influence of the human translation. Compiling a large comparable corpus is easier, especially for general language (Talvensaari et al., 2007). In contrast, specialized comparable corpora are traditionally of modest size due to the difficulty to obtain many specialized documents in a language other than English. Specialized comparable corpora have a size of around one million words whereas general-domain comparable corpora can gather several million words (Morin and Hazem, 2014).

One way to overcome the small size of specialized comparable corpora is to associate external resources. These resources may be close specialized corpora (e.g. a breast cancer corpus may benefit from contexts derived from a more general oncology corpus), corpora of different types of discourse and gender (e.g. a corpus of popular science discourse supplementing a corpus of scientific discourse), corpora of general language or out-of-domain data. The main challenge is to know how to associate such resources with a comparable specialized corpus.

Proceedings of the 27th International Conference on Computational Linguistics, pages 937–949
Santa Fe, New Mexico, USA, August 20-26, 2018.

According to Jakubina and Langlais (2017), word embeddings are more effective on large comparable corpora than on small comparable corpora. This statement lend support the idea that enriching small specialized comparable corpora may be beneficial to bilingual terminology extraction task. The combination of external resources such as a general-domain comparable corpus with a specialized comparable corpus can be performed using word embedding models. We recently conducted a first attempt in Hazem and Morin (2017) and have shown under which conditions external resources introduced in the form of Skip-gram and CBOW models can be jointly used to improve the performance of bilingual terms extraction. However, our approach was not able to compete with the historical count-based projection approach (Fung and McKeown, 1997; Rapp, 1999). Our current work pursues this direction by contrasting different neural embedding models and by showing how to take advantage of their combination. More specifically, we show that the character-based Skip-gram and CBOW models (Bojanowski et al., 2016) drastically outperform other models including Skip-gram and CBOW. We also propose a new approach based on Ensemble models which combines specialized and general domain embeddings to obtain a unified Meta-Embedding model. Our approach shows significant improvements and obtains the best results on two specialized English/French comparable corpora.

2 Related Work

According to Hermann and Blunsom (2014), methods dealing with bilingual lexicon extraction from comparable corpora can be classified as distributional-based or distributed-based approaches. In the former, words are represented by their context vectors using a distributional count-based approach also known as the standard approach (Fung, 1998; Rapp, 1999). While in the latter, words are embedded into a low-dimensional vector space using neural network models (Bengio et al., 2003; Mikolov et al., 2013b).

The historical standard approach builds a context vector for each word of the source and the target languages, translates the source context vectors to the target language using a bilingual seed lexicon, and compares the translated context vectors to each target context vector using a similarity measure. Different contributions have been proposed in the past few years to improve each step of the standard approach (Gaussier et al., 2004; Gamallo, 2007; Ismail and Manandhar, 2010; Prochasson and Fung, 2011; Hazem and Morin, 2012; Bouamor et al., 2013, among others).

With the advent of neural network techniques, Mikolov et al. (2013a) were the first to propose a method to learn a linear transformation from the source language into the target language to improve the task of lexicon extraction from bilingual corpora. Faruqui and Dyer (2014) introduced canonical correlation analysis (CCA) to project the embeddings in both languages to a shared vector space. More recently, Artetxe et al. (2016) presented an approach for learning bilingual mappings of word embeddings that preserves monolingual invariance using several meaningful and intuitive constraints related to other proposed methods (Faruqui and Dyer, 2014; Xing et al., 2015).

Jakubina and Langlais (2017) made a careful comparison of the approaches of Mikolov et al. (2013a) and Faruqui and Dyer (2014) with the standard approach. They have clearly shown that the two previous approaches outperform the standard approach for very frequent terms to be translated (the number of occurrences is at least 250 from an English-Spanish comparable corpus obtained from the 6th workshop on statistical machine translation gathering 2.55 giga words). On the other hand, when the terms are less frequent (the number of occurrences is less than 25 from a French/English comparable corpora built from the Wikipedia dumps gathering 1.53 giga words), the standard approach slightly outperforms the two previous embedding approaches. More recently, we have shown under which conditions Skip-gram and CBOW models can be jointly used to improve the performance of bilingual terms extraction from specialized comparable corpora without exceeding the results of the standard approach (Hazem and Morin, 2017).

Other works performed bilingual word representation without word-to-word alignments of comparable corpora. Chandar et al. (2014) and Gouws et al. (2014) for instance used multilingual word embeddings based on sentence-aligned parallel data whereas Vulić and Moens (2015) and Vulić and Moens (2016) used document-aligned non-parallel data to produce bilingual word embeddings. Theses works are based

on sentence- or document-aligned of general-domain comparable corpora and are outside the scope of this study. It is unlikely to find this type of alignments in a specialized comparable corpus.

3 Embedding Models

In this section, we briefly recall the main and recent word embedding models that we will investigate in this study.

CBOW and Skip-gram are two distributed representations introduced by Mikolov et al. (2013b) that capture linguistic regularities, namely the Continuous Bag-of-Words (CBOW) model and the Skip-gram model. The principle of the CBOW model is to combine the representations of surrounding words to predict the word in the middle, while the training objective of the Skip-gram model is to learn how to predict the surrounding words based on the representations of the middle word. If these models exhibit similar architectures, CBOW is faster and more suitable for large data sets while Skip-gram gives better word representations when monolingual data is small (Mikolov et al., 2013a).

Glove takes advantage of the main benefits of count data while simultaneously capturing the meaningful linear substructures prevalent in prediction-based methods such as word2vec. It is a global log-linear regression model that makes use of a global factorization model and local context window methods to represent words in a global vector space model (Pennington et al., 2014). This model directly captures the global statistics from the corpus based on co-occurrence word probabilities. Its training objective is to learn word vectors such that their dot product equals the log-probability of word's co-occurrence. Glove has shown good results in word analogy, word similarity, and named entity recognition tasks.

Structured Embeddings are two adaptations of CBOW and Skip-gram models that include ordering information[1] (Ling et al., 2015). While word2vec is insensitive to word order, the structured embedding model includes position information in the context representation of words. Given the embedding of the center word w, the Skip-gram model for instance uses a single output matrix to predict every contextual word. In contrast, the structured Skip-gram adapts the model to the positioning of the surrounding words. It defines an output for each relative position to the center word. The adaptation of CBOW is the continuous window model where the input is the concatenation of the embeddings of context words. While in the standard CBOW, the input model is the sum of the embeddings of the context words.

Character n-gram Embeddings is an enhanced variant of the Skip-gram and CBOW models that enrich word vectors with subwords information. It takes into account the internal structure of words which can be very useful for morphologically rich languages. It also incorporates character n-gram embeddings where each word is represented by a bag-of character n-gram (Bojanowski et al., 2016). More precisely, it uses character embedding and word embedding models jointly performing the vector sum of both to form the final embedding representation of words. We refer to the character Skip-gram model by CharSG and the character CBOW model by CharCBOW.

Other models such as the dependency-based model (Levy and Goldberg, 2014) and generalized-based model (Li et al., 2017) were assessed but not presented in this paper for sake of clarity and because of the very low results obtained on the specialized domains when compared to the above presented models.

Several pre-trained embedding models are publicly available such as CBOW and Skip-gram models (Mikolov et al., 2013b), global word representation-based models (Pennington et al., 2014), character skip-gram-based models (Bojanowski et al., 2016), etc. If it is interesting to study the impact of pre-trained embeddings on bilingual terminology extraction from comparable corpora. The major part of the above cited pre-trained embedding models exist solely in English. We only use the character skip-gram (CharSG) model (Bojanowski et al., 2016) which is available in both French and English.

[1] Word order in context word representation.

4 Approach

The task of bilingual terminology extraction from specialized comparable corpora consists of acquiring for each term of the source language its translation in the target language. The basic idea using embedding models is to first (1) build word embeddings of the source and the target languages, then, (2) to build a mapping matrix (Artetxe et al., 2016) that allows to obtain for each source term, its representation in the target language and finally, (3) to measure the similarity between the target representation of the source term and all the word candidates of the target language to extract the most similar term as the correct translation (Mikolov et al., 2013a; Artetxe et al., 2016). Our approach follows these three steps (Mikolov et al., 2013a) while acting at the word embedding level representation. Our idea is to enrich the word embedding representation of the source and target languages in order to improve the mapping matrix and so, bilingual terminology extraction from specialized comparable corpora. To do so, we present several ways to take advantage of word embedding models and ensemble approaches.

The principle of ensemble approaches is to combine different models in order to capture the strengths of each individual model. The main combination techniques that have shown their effectiveness so far are vectors addition (Garten et al., 2015) and vectors concatenation (Garten et al., 2015; Yin and Schütze, 2016). For vectors addition, given two embedding models, the procedure consists in applying a simple dimension-wise vectors addition[2]. For vectors concatenation, given two embedding models of dimensions $dim1$ and $dim2$, the resulting concatenated embedding vector will be of size $dim1+dim2$. The vectors have to be normalized before concatenation. Usually L2 norm is applied[3]. Yin and Schütze (2016) performed a weighted concatenation of five embedding models. They also experienced the SVD (Singular Value Decomposition) on top of weighted concatenation vectors of dimension 950. This resulted in a reduced model of 200 dimensions.

In the line of the above cited approaches, we first explore the ensemble modeling (additive and concatenation) on a large scale over the multiple word embedding models presented in Section 3. This is done exclusively on the small specialized comparable corpora. We then explore different ways to supply each specialized comparable corpus with external resources based on ensemble approaches. We show new ways to take advantage of external data and embedding models in order to efficiently extract bilingual lexicons from specialized comparable corpora. Our methodology is two-fold. In the subsection 4.1 we first describe ensemble approaches and their application on one type of corpora (here the specialized comparable corpora), then in subsection 4.2, we introduce the adaptation of ensemble approaches while combining the specialized corpus with external resources.

4.1 Specialized Meta-Embeddings

While each embedding model captures some specific context word information, it is natural and straightforward to seek for their complementarity. The specialized meta-embeddings approach consists in combining different embedding models learned from the specialized corpus. We basically use an ensemble approach to represent each word, which means that each word has its own meta-embedding. This is illustrated in the following equation:

$$Ensemble(w) = f(v_w^1, v_w^2, ..., v_w^n) \quad (1)$$

with $Ensemble(w)$ the meta-embedding representation of a given word w and f the ensemble approach used to combine the different embedding models. f can be the additive or the concatenation technique. Finally, v_w^n represents a given embedding model of the word w and n the number of used embedding models. For instance, given the Glove, the CBOW and the Skip-gram models trained on a specialized corpus, the ensemble model of a given word w would be the concatenation (or addition) of its three embedding representations ($f(v_w^{Glove}, v_w^{CBOW}, v_w^{Skip-gram}$).

[2]This technique can not be applied when embeddings are not of the same dimension size (unless using padding).

[3]L2 norm can be applied either at dimension level (as suggested by Glove authors) or at vector length level.

4.2 Mixed Meta-Embeddings

While specialized comparable corpora suffer from the lack of data, one good alternative is to enrich them with external resources. The remaining question is how to take advantage of out-of-domain data to increase specialized corpus size without degrading its specific properties. A basic way to combine word embeddings from two different data sets (here the specialized and the external corpora) is to apply an ensemble approach while fixing the type of the embedding model. For instance, using the CBOW model, each word can be jointly represented by its embedding vector issued from the specialized corpus (noted v_w^s) and its embedding model issued from the general domain corpus (noted v_w^g). This is illustrated in the following equation:

$$Ensemble(w) = f(v_w^s, v_w^g) \tag{2}$$

The Mixed Meta-Embedding representation can be generalized over several embedding models as follows:

$$Ensemble(w) = f'(f(v_w^{s^1}, v_w^{g^1}), f(v_w^{s^2}, v_w^{g^2}), ..., f(v_w^{s^n}, v_w^{g^n})) \tag{3}$$

where f and f' represent the ensemble functions (concatenation or addition). n represents the number of embedding models. One condition while combining embeddings built from different corpora is to ensure that a given word w is present in all the combined corpora. If not, we can choose to discard this word or to replace the missing vector by zeros (padding).

5 Data and Resources

In this section, we describe the different textual resources used for our experiments: the comparable corpora, the bilingual dictionary and the terminology reference lists.

5.1 Comparable Corpora

The specialized comparable corpora were selected in terms of bilingual terminology access in technical domains. For our experiments, we used two specialized comparable corpora:

Breast cancer corpus (BC) is composed of documents collected from scientific and medical portals such as the ScienceDirect[4]. The documents were taken from the medical domain within the sub-domain of "breast cancer". We have selected the documents published between 2001 and 2015 where the title or the keywords contain the term *breast cancer* in English and its translation in French.

Wind energy corpus (WE) has been released in the TTC project[5]. This corpus has been crawled from the Web using *Babouk* crawler (Groc, 2011) based on several keywords such as *wind*, *energy*, *rotor* in English and its translation in French.

In addition, we use three corpora of general language as external resources:

JRC acquis corpus (JRC) is a collection of legislative texts of the European Union[6]. We used the French-English version at OPUS which is based on the paragraph-aligned corpus provided by JRC (Tiedemann, 2012).

Common crawl corpus (CC) is an open repository of data collected over 7 years of web crawling sets of raw web page data and text extracts[7].

[4] www.sciencedirect.com/
[5] www.ttc-project.eu/index.php/releases-publications
[6] opus.lingfil.uu.se/JRC-Acquis.php
[7] commoncrawl.org

Wikipedia corpus (Wiki) The English wikipedia corpus[8] is a dump which was released on 03-Feb-2018 and the French wikipedia corpus [9] was released on 02-Feb-2018.

Even if JRC, CC are parallel corpora, we didn't explicitly exploit their parallel relationship. We considered these external data sets as if they were comparable corpora. The documents were normalized through tokenisation, part-of-speech tagging, and lemmatisation using the TTC TermSuite[10]. Finally, the function words were removed and the hapax were discarded. Table 1 shows the number of documents (# doc.) and the number of content words (# words) for each corpus.

Corpus	# content words		# distinct words	
	FR	EN	FR	EN
BC	521,262	525,934	6,630	8,221
WE	313,954	314,551	5,346	6,378
JRC	70.3M	64.2M	100,004	93,104
CC	91.3M	81.1M	250,999	259,226
WIKI	740.2M	2,669M	1,067,095	2,443,866

Table 1: Characteristics of the corpora.

5.2 Bilingual Dictionary

The bilingual dictionary used in our experiments is the French/English dictionary ELRA-M0033 (243,539 entries) available from the ELRA catalogue[11]. This resource is a general language dictionary which contains only a few terms related to the medical and wind energy domains.

5.3 Gold Standard

To evaluate the quality of bilingual terminology extraction from specialized comparable corpora, a bilingual terminology reference list is required. In the general domain, the reference list is randomly composed of a sub-part of the bilingual dictionary (Gaussier et al., 2004; Jakubina and Langlais, 2017). In the specialized domain, this list is usually composed of few words that reflect the terminology of the specialized comparable corpus. For instance, Chiao and Zweigenbaum (2002) used a list composed of 95 single words, Morin et al. (2007) used 100 single words and Bouamor et al. (2013) used 125 and 79 single words. For breast cancer, the lists are derived from the UMLS[12] meta-thesaurus. Concerning wind energy, the lists are provided with the corpora (see footnote 5). Each word composing a pair of terms of a reference list appears at least 5 times in the comparable corpus. The bilingual terminology reference list is composed of 248 French/English single words for the Breast cancer corpus and 150 French/English single words for the Wind energy corpus.

6 Experiments and Results

We conducted two sets of experiments. The first one aims at providing insights into the behaviour of each state-of-the-art embedding model on the specialized comparable corpora. The second one aims at studying the contribution of ensemble models.

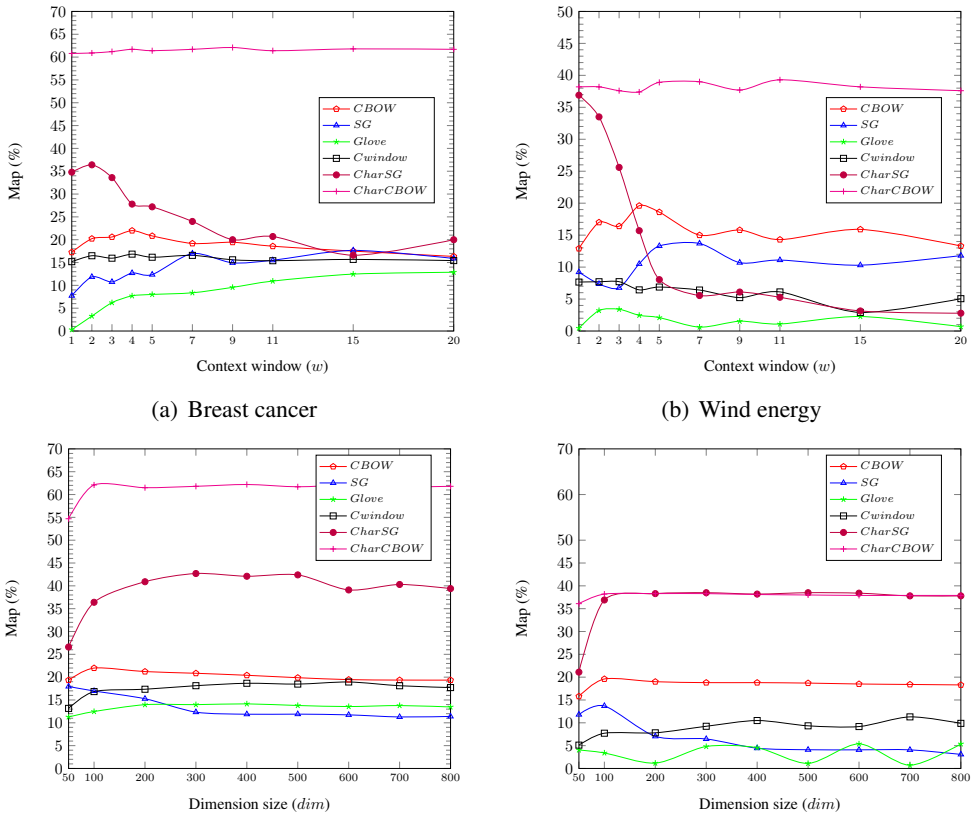
We present the results obtained for the terms belonging to the reference list for English to French direction measured in terms of the Mean Average Precision (MAP) (Manning et al., 2008) as follows:

$$MAP(Ref) = \frac{1}{|Ref|} \sum_{i=1}^{|Ref|} \frac{1}{r_i} \quad (4)$$

[8] dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
[9] dumps.wikimedia.org/frwiki/latest/frwiki-latest-pages-articles.xml.bz2
[10] code.google.com/p/ttc-project
[11] www.elra.info/
[12] www.nlm.nih.gov/research/umls

where $|Ref|$ is the number of terms of the reference list and r_i the rank of the correct candidate translation i.

Figure 1 shows the results of each embedding model for the task of bilingual terminology extraction from the two specialized comparable corpora. We report the results of the continuous bag-of-word model ($CBOW$), the Skip-gram model (SG), the glove model ($Glove$), the structured continuous window model ($Cwindow$) and the two character n-gram models, namely the character skip-gram model ($CharSG$) and the character CBOW model ($CharCBOW$)[13]. For each specialized comparable corpus, we varied the context window size (w) (see sub-figures 1(a) and 1(b)) and the embeddings dimension size (dim) (see sub-figures 1(c) and 1(d)).



(a) Breast cancer (b) Wind energy

(c) Breast cancer (d) Wind energy

Figure 1: Contrasting several embedding representations on the breast cancer and wind energy corpora.

The first remarkable point is the performance of the character n-gram embedding models which far outperform other models. This is an important finding since to our knowledge, no previous evaluation of character n-gram-based models have been conducted so far for bilingual terminology extraction from specialized corpora. The first competitive model which is $CharCBOW$, is not sensitive to the context window size as well as the embeddings dimension size. The second competitive model which is $CharSG$, is also not sensitive to the embeddings dimension size but sensitive to the context window size. $CBOW$ which is the third competitive model, turned out to be the best word-based model. Nevertheless, its performance is far below the character-based models. $CBOW$ is also not very sensitive to the context window and the embeddings dimension sizes.

Tables 2 and 3 show a comparison of several combinations of word embedding models for the two specialized comparable corpora. Each combination is based on L2 normalization at vector length level (len) or at vector dimension level (dim). We chose the three embedding models that have shown the best performance individually (according to the Figure 1: $CharSG$, $CharCBOW$ and $CBOW$) and

[13]We don't report the results of dependency-based and structured-based models due to the low performance of these approaches.

we combine them with all the models. For each Table, the first line is a quick reminder of the three embedding models and their performance in terms of MAP scores shown between brackets. The following two lines also remind the embedding models used for combination and their MAP scores when used individually. The following four lines present the results of the combination of these models by addition or concatenation for both normalizations. For instance in Table 2, the $CharSG$ and $CBOW$ embedding models used individually have respectively 36.4% and 21.9% of MAP and the combination of the two models gives 22.9% of MAP by addition and 34.9% of MAP by concatenation with L2 normalization at vector length level. Combining embedding models is very effective when using concatenation in most cases. The best combination model is obtained by the concatenation of $CharSG$ and $CharCBOW$ using length L2 Norm (70.3% of MAP in Table 2) and by the concatenation of $CharCBOW$ and $CBOW$ (49.5% of MAP in Table 3) followed by the same model using dimension L2 Norm for breast cancer corpus (68.1% of MAP in Table 2) and using $CharSG$ with SG concatenation for wind energy corpus (45.4% of MAP in Table 3). We observe that both types of L2 normalization are in general useful for concatenation with a better performance when using length normalization.

	CharSG (36.4)					CharCBOW (60.8)				CBOW (21.9)			L2 Norm.
	CBOW	SG	Glove	CharCBOW	Cwindow	SG	Glove	CBOW	Cwindow	SG	Glove	Cwindow	
	21.9	16.9	12.4	60.8	16.1	16.9	12.4	21.9	16.1	16.9	12.4	16.1	
Addition	22.9	15.9	19.7	59.3	18.9	47.6	40.5	57.3	55.8	20.3	14.6	18.4	Len
Concat.	34.9	34.6	29.8	**70.3**†	34.1	65.6†	48.3	**64.3**†	64.7†	**26.2**†	18.2	23.9†	
Addition	21.9	17.2	20.2	56.0	18.8	40.9	40.8	55.1	48.5	17.4	15.5	18.5	Dim
Concat.	37.7†	37.0	31.4	68.1†	33.8	58.8	50.9	62.2†	59.2	23.9†	18.9	**26.9**†	

Table 2: Embeddings combination using the breast cancer corpus (MAP %) (†:t-test significance at 0.05).

	CharSG (36.9)					CharCBOW (38.2)				CBOW (19.6)			L2 Norm.
	CBOW	SG	Glove	CharCBOW	Cwindow	SG	Glove	CBOW	Cwindow	SG	Glove	Cwindow	
	19.6	13.7	4.81	38.2	19.6	13.7	4.81	19.6	19.6	13.7	4.81	19.6	
Addition	35.7	21.1	18.6	33.2	23.7	29.5	23.1	41.9†	25.5	11.8	13.4	10.4	Len
Concat.	**45.4**†	42.8†	24.1	38.8†	30.0	45.6†	28.0	**49.5**†	40.1†	21.1†	15.9	22.0†	
Addition	20.9	25.5	13.5	33.2	12.8	26.5	12.1	25.6	12.6	11.9	13.2	11.1	Dim
Concat.	31.3	43.4†	18.2	39.0†	32.1	44.0†	17.4	32.3	41.6†	20.2	17.1	**23.5**†	

Table 3: Embeddings combination using the wind energy corpus (MAP %)(†:t-test significance at 0.05).

Model	BC	JRC	CC	WIKI	ALL	BC + JRC	BC + CC	BC + WIKI	ALL
CharCBOW	**60.8**	35.3	57.4	60.7	62.7	52.9	73.9	73.1	**74.3**
CharSG	36.4	41.1	61.6	58.8	61.8	65.2	76.5	69.2	70.3
(Bojanowski et al., 2016)	-	-	-	**72.4**	-	-	-	-	-
Model	WE	JRC	CC	WIKI	ALL	WE + JRC	WE + CC	WE + WIKI	ALL
CharCBOW	**38.2**	42.8	60.1	67.8	70.1	43.8	65.9	71.3	71.4
CharSG	36.9	49.9	61.9	71.0	**70.8**	55.1	65.7	72.2	**72.4**
(Bojanowski et al., 2016)	-	-	-	68.2		-	-	-	-

Table 4: Results (MAP %) of different embedding models on the specialize corpora as well as several out-of-domain corpora and their combinations ((Bojanowski et al., 2016) is the pre-trained CharSG model).

Table 4 shows the results of different combinations[14] of specialized corpora with external resources.

[14]Combination means that we first merge different data sets in one single corpus and then, we learn an embedding model

We represent in the 1^{st} column the results for the specialized corpora taken individually (BC and WE) and then, from the 2^{nd} to the 4^{th} column we show the results of the external resources taken individually and their combination represented in the 5^{th} column (ALL). Finally, from the 6^{th} to the 8^{th} column we combine the specialized corpora with each external data and then their entire combination[15] (All).

Overall, we see that combining different resources gives significant improvements over the two specialized corpora. We also notice the usefulness of external data used individually which performs better than small specialized corpora, except for CharCBOW which shows the strength of this model and its usefulness over other types of embedding models. Also, using the pre-trained embedding model of Bojanowski et al. (2016) obtained good results (72.4% for BC and 68.2% for WE) but if we compare it to the same model (CharSG) learned while combining specialized and external data sets, we observe better performance (76.5% for BC and 72.4% for WE) which shows that using jointly specialized and external data is more efficient than the use of external data only. This statement is confirmed by the results obtained by individual external data sets which are always lower than their combination with specialized corpora.

	Individual corpus			Corpus combination (GSA)	
Models	BC	JRC	CC	BC + JRC	BC + CC
SA	27.0	**52.0**	**75.5**	61.7	**80.2**
CBOW	17.1	40.3	60.9	49.9	67.7
SG	12.8	40.5	56.0	46.5	63.2
CharCBOW	**60.8**	35.3	57.4	52.9	73.9
CharSG	36.4	41.1	61.6	**65.2**	76.5
				Vector concatenation	
SCBOW	-	-	-	53.7	70.7
SSG	-	-	-	36.3	40.2
SCBOW+SSG	-	-	-	56.1	70.9
SSA	-	-	-	66.6	**82.3**
Our approaches					
SCharCBOW	-	-	-	64.9	74.9
SCharSG	-	-	-	73.0	77.4
SCharCBOW+SCharSG	-	-	-	67.0	**80.7**
Meta-Emb (Best)	-	-	-	**74.8**	**83.1**

Table 5: Results (MAP %) of the *Standard Approach* (SA), the *Global Standard Approach* (GSA) and the *Selective Standard Approach* (SSA) and our approaches using CharCBOW and CharSG and their combinations (SCharCBOW, SCharSG and Meta-Emb) for the breast cancer corpus (BC) using the different external data (the improvements indicate a significance at the 0.05 level using the Student t-test).

In Table 5, we report the results obtained in Hazem and Morin (2017) (SA, CBOW, SG, SCBOW, SSG ,SCBOW+SSG and SSA) and our results using combination and meta-embeddings of character n-gram models (SCharCBOW, SCharSG, SCharCBOW+SCharSG and Meta-Emb (Best)[16]). The main conclusion in Hazem and Morin (2017) is that the best embedding combination (SCBOW+SSG) couldn't outperform the selective standard approach (SSA). According to our results, character n-gram models and their combination obtained better results than the best CBOW and Skip-gram combination (SCBOW+SSG) and also outperformed the selective standard approach (SSA obtained 66.6% using BC + JRC and 82.3% using BC + CC while our best model obtained **74.8%** and **83.1%** on the same corpora). The results of Table 5 provide strong support for data combination and meta-embeddings using character n-gram models. Also, we highlight the fact that character ngram models and their combination is much faster than CBOW and Skip-gram models. In addition, the dimension size of embedding models

on the merged corpus.

[15] ALL in the 5^{th} column means that we combine JRC, CC and WIKI, while ALL in the 9^{th} column means that we combine JRC, CC and WIKI and the specialized corpora BC or WE.

[16] Meta-Emb (best) stands for the combination of SCharSG on specialized corpus with SCharCBOW+SCharSG on external data.

is very low (around 300) while in the standard approach it corresponds to the vocabulary size which leads to sparse vectors and high computational cost.

7 Discussion

The first important finding of this work is the efficiency of the character n-gram models (CharCBOW and CharSG) which drastically outperform other models, whether on specialized or general domain data sets. Their better performance can be explained by the fact that both models are based on characters to build the embedding models. While CBOW and Skip-gram suffer from the lack of data to build efficient models over specialized corpora, the character-based approaches benefit of much more training examples as they use characters for their models. The second important finding is the performance of external data when applied for extracting bilingual terms. This is not surprising as external data sets such as wikipedia or common crawl for instance, contain several scientific and specialized documents. In addition, we could observe the complementarity of external resources with the specialized domain data sets. More precisely, concatenating embeddings of specialized and external resources significantly improves the results. This can be explained by the nature of the captured information which can be resumed in the concatenated embedding vectors. If a single generic embedding model is difficult to obtain, character n-gram and word-based embedding models can be efficiently combined to improve bilingual terminology extraction from comparable corpora.

We also conducted an error analysis of the different proposed models and we couldn't find a strong relation between the embedding models and the non captured terms. However, we observed that the CharCBOW model with the meta-embedding combination using BC with CC corpora improves overall the rank of the translations obtained for each individual corpus. By looking at the 115 translations of BC and the 80 of BC+CC not found in the first rank, we found 58 terms in common including 44 terms with a better rank with BC+CC corpora. In the same way, from the 75 translations of CC and the 80 of BC+CC not found in the first rank, we found 62 terms in common including 46 terms with a better rank with BC+CC corpora. It might seem surprising that we found more terms outside the first rank with CC corpus than BC+CC (80 versus 75) since the MAP is lower with CC than BC+CC (57.4 versus 73.9). In fact, the CharCBOW model with BC+CC improves overall the rank of all the translations of CC and more particularly the rank of the first hundred translations taken into account in the calculation of the MAP. We also observed that the more frequent terms are the best translated for SA and SSA approaches. For the embedding approaches, this observation is not relevant. For instance, the translations of frequent terms such as *cancer* and *breast* are found in the first ranks and the translations of infrequent terms such as *lumpectomy* and *fibroadenoma* are found in the last ranks with SA and SSA approaches. With embedding approaches, *cancer* and *fibroadenoma* are found in the first ranks and *breast* and *lumpectomy* in the last ranks. We have not been able to better characterize terms in the last ranks with embedding approaches. This is strongly dependent on embedding parameters and also context size and embedding dimensions. Our best system, however obtained 90% of precision on top 5, in the perspective of providing first translation terms for translation aided systems, our proposed approach is certainly more appropriate than the standard approach whether in terms of computational cost or in terms of accuracy.

8 Conclusion

In this paper we have explored a variety of embedding models and there impact on the task of bilingual terminology extraction from specialized comparable corpora. We have also proposed meta-embedding representations and have shown under which conditions they can be jointly used for better performance. If further investigations are probably needed, our findings strengthens the idea that using meta-embeddings based on specialized and general domain data sets improves the performance of mining bilingual specialized lexicons.

Acknowledgments

The research leading to these results has received funding from the French National Research Agency under grant ANR-17-CE23-0001 ADDICTE (Distributional analysis in specialized domain).

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, TX, USA.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/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.

A. P. Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'14)*, pages 1853–1861, Montreal, Quebec, Canada.

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 1208–1212, Tapei, Taiwan.

Estelle Maryline Delpech. 2014. *Comparable Corpora and Computer-assisted Translation*. John Wiley & Sons, Inc.

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 and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, pages 192–202, Hong Kong.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98)*, pages 1–17, Langhorne, PA, USA.

Pablo Gamallo. 2007. Learning bilingual lexicons from comparable english and spanish corpora. In *Proceedings of the 11th Conference on Machine Translation Summit (MT Summit XI)*, pages 191–198, Copenhagen, Denmark.

Justin Garten, Kenji Sagae, Volkan Ustun, and Morteza Dehghani. 2015. Combining Distributed Vector Representations for Words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 95–101, Denver, CO, USA.

Eric. Gaussier, Jean-Michel Renders, Irena. Matveeva, Cyril. Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *CoRR*, abs/1410.2455.

Clément De Groc. 2011. Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In *Proceedings of 10th International Conferences on Web Intelligence (WIC'11)*, pages 497–498, Lyon, France.

Amir Hazem and Emmanuel Morin. 2012. ICA for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 126–133, Istanbul, Turkey.

Amir Hazem and Emmanuel Morin. 2017. Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora. In *Proceedings of the 8th 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, MD, USA.

Azniah Ismail and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 481–489, Beijing, China.

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. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 302–308, Baltimore, MD, USA.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. Investigating Different Syntactic Context Types and Context Representations for Learning Word Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 2421–2431.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, pages 1299–1304, Denver, CO, USA.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, 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 S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Emmanuel Morin and Amir Hazem. 2014. Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 1284–1293, Baltimore, MD, USA.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Jian-Yun Nie. 2010. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

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)*, pages 1532–1543, Doha, Qatar.

Emmanuel Prochasson and Pascale Fung. 2011. Rare Word Translation Extraction from Aligned Comparable Documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, pages 1327–1335, Portland, OR, USA.

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, MD, USA.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. 2007. Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 25(1).

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Ivan Vulić and Marie-Francine Moens. 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 (ACL'15)*, pages 719–725, Beijing, China.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research (JAIR)*, 55(1):953–994.

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-HLT'15)*, pages 1006–1011, Denver, CO, USA.

Wenpeng Yin and Hinrich Schütze. 2016. Learning Word Meta-Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 1351–1360, Berlin, Germany.