Data Selection for Bilingual Lexicon Induction from Specialized Comparable Corpora

Martin Laville[1], Amir Hazem[1], Emmanuel Morin[1], Philippe Langlais[2]
[1]LS2N, UMR CNRS 6004, Universite de Nantes, France
[2]RALI-DIRO, Montreal, Canada
[1]firstname.lastname@ls2n.fr, [2]felipe@iro.umontreal.ca

Abstract

Narrow specialized comparable corpora are often small in size. This particularity makes it difficult to build efficient models to acquire translation equivalents, especially for less frequent and rare words. One way to overcome this issue is to enrich the specialized corpora with out-of-domain resources. Although some recent studies have shown improvements using data augmentation, the enrichment method was roughly conducted by adding out-of-domain data with no particular attention given to how to enrich words and how to do it optimally. In this paper, we contrast several data selection techniques to improve bilingual lexicon induction from specialized comparable corpora. We first apply two well-established data selection techniques often used in machine translation that is: Tf-Idf and cross entropy. Then, we propose to exploit BERT for data selection. Overall, all the proposed techniques improve the quality of the extracted bilingual lexicons by a large margin. The best performing model is the cross entropy, obtaining a gain of about 4 points in MAP while decreasing computation time by a factor of 10.

1 Introduction

Mining translation equivalents to automate the process of generating and extending dictionaries from bilingual corpora is a well known Natural Language Processing (NLP) application. Initially conducted on parallel data, this modus operandi has rapidly moved to using comparable corpora, mainly due to their availability and the fact that they are easier to acquire (Sharoff et al., 2013). Comparable corpora gather texts of the same domain, often over the same period without being in a translation relation. Journalistic or encyclopedic websites as well as web crawl data are usually popular for this purpose (Rapp, 1999; Li and Gaussier, 2010; Rapp et al., 2020). In technical and scientific domains, comparable corpora suffer from their modest and limited size compared to general language corpora. This phenomenon is amplified by the difficulty to obtain many specialized documents in a language other than English (Morin and Hazem, 2014). This aspect, which is gradually changing with the deployment of open access data, is nevertheless a major difficulty.

One way to overcome the limited size of specialized (in-domain) comparable corpora for bilingual lexicon induction (BLI) is to associate them with out-of-domain resources as lexical databases (Bouamor et al., 2013) or large general domain corpora (Hazem and Morin, 2018). For instance, by combining a specialized comparable corpus with a general domain corpus, methods based on distributional analysis are boosted (Hazem and Morin, 2018). General domain corpora greatly enhance the representation of specialized vocabulary by adding new contexts. The main drawbacks of this data augmentation approach is the introduction of polysemous information as well as the tremendous increase of computation time. Consider for instance, a French/English comparable corpus in the medical domain and the French terms *os* (*bone*) and *sein* (*breast*). When looking for additional contexts in general corpora, *os* is very likely to recover contexts dedicated to an *operating system* since *OS* is the abbreviation used in French. Similarly, when studying the French term *sein*, many contexts related to the French preposition *au sein de* (*within*)

Proceedings of the 28th International Conference on Computational Linguistics, pages 6002–6012
Barcelona, Spain (Online), December 8-13, 2020

will be added. In the same way, the English medical term *breast* can be associated with food just as in *chicken breast* to mention only the most obvious ones.

Associating out-of-domain data with a specialized corpus is not the mainstream research direction in bilingual lexicon induction from specialized comparable corpora. In computational terminology it is generally accepted that the specialized corpus conveys knowledge of the domain. In the same way, polysemy is often considered as a marginal phenomenon in technical and scientific domains. In this sense, associating out-of-domain data with a specialized comparable corpus can be seen as a "heresy" or, as in the previous examples, an unfortunate way to introduce polysemy.

In the context of Statistical Machine Translation (SMT), Wang et al. (2014) demonstrated that adding out-of-domain data to the training material of a system was detrimental when translating scientific and technical domains. Instead of using a data augmentation strategy, data selection is proposed to improve the quality of SMT systems (Moore and Lewis, 2010; Axelrod et al., 2011; Wang et al., 2014). The basic idea is that in-domain training data can be enriched with suitable sub-parts of out-of-domain data.

Within this context, we apply two well-established data selection techniques often used in machine translation that is: Tf-Idf and cross entropy. We also propose for the first time to exploit BERT for data selection. We show that a subtle selection of the contexts of out-of-domain data allows to improve the representation of specialized domains, while preventing the introduction of polysemy induced by data augmentation. Overall, all the proposed techniques improve the quality of the extracted bilingual lexicons by a large margin. The best performing model is the cross entropy, obtaining a gain of about 4 points in MAP while decreasing computation time by a factor of 10.

2 Related Work

The historical distributional approach (Fung, 1998; Rapp, 1999), known as the standard bag of words approach, builds a context vector for each word of the source and the target languages, translates the source context vectors into the target language using a bilingual seed lexicon, and compares the translated context vectors to each target context vector using a similarity measure. While this approach gives interesting results for large general comparable corpora (Gaussier et al., 2004; Laroche and Langlais, 2010; Vulić et al., 2011), it is rather unsuitable for small specialized comparable corpora due to the context vectors sparsity. (Chiao and Zweigenbaum, 2002; Morin et al., 2007; Prochasson et al., 2009).

More recent distributed approaches, based on deep neural network models (Mikolov et al., 2013), have come to renew traditional ones. Using these approaches, words are embedded into a low-dimensional vector space. For instance, Mikolov et al. (2013) proposed an approach to learn a linear transformation from the source language into the target language. Faruqui and Dyer (2014) used the Canonical Correlation Analysis to project the embeddings in both languages into a shared vector space. Xing et al. (2015) proposed an orthogonal transformation based on normalized word vectors to improve the inconsistency among the objective function used to learn the word vectors and to learn the linear transformation matrix. More recently, Artetxe et al. (2016; 2018a) proposed and then improved an approach that generalizes previous works in order to preserve monolingual invariance using several meaningful and intuitive constraints (i.e. orthogonality, vector length normalization, mean centering, whitening, etc.). Finally, Conneau et al. (2017) and Artetxe et al. (2018b) proposed unsupervised mapping methods getting results close to supervised methods. A thorough comparison of cross-lingual word embeddings has been proposed by Ruder (2017).

A comparison of the approaches of Mikolov et al. (2013) and Faruqui and Dyer (2014) with the standard bag of words approach was carried out by Jakubina and Langlais (2017). The authors showed that embedding approaches perform well when the terms to be translated occur very frequently while the standard approach is slightly better when the terms are less frequent. On specialized domains, Hazem and Morin (2018) compared different approaches and observed that the one described by Bojanowski et al. (2016) (an enhanced variant of the skip-gram and C-BOW models) outperformed the others.

Recently, Peters et al. (2018) proposed a deep contextualized word representation that improves the state-of-the-art across different challenging NLP tasks. This representation is useful to model different types of syntactic and semantic information about words-in-context. It is thus possible from a general

language corpus to have different embeddings for a given word, with each of these embeddings reflecting one of the word's meanings. In specialized domains, it is widely accepted that words can have only one meaning even if their contexts may convey different meanings in a general language corpus.

El Boukkouri et al. (2019) showed that using a combination of out-of-domain contextualized word representation (ELMo) with small in-domain uncontextualized word representation (word2vec) does not improve the results while compared to models trained on large in-domain corpora. For this reason, we focused this study on uncontextualized word embeddings (fastText).

3 Data Selection Techniques

While abruptly adding out-of-domain data to a specialized corpus, with no further thoughts on how to do it improves the results, the computational time is greatly increased and polysemous information is introduced. To leverage the aforementioned issues, we present in this section different data selection techniques that we use in our experiments.

Tf-Idf is the term frequency-inverse document frequency score and is performed at the document level. It is computed for both in-domain and out-of-domain vocabularies, then the out-of-domain documents are ranked by measuring their cosine similarity with the in-domain corpus. Given an in-domain document D_I and an out-of-domain document D_O, the similarity is computed as follows:

$$Cos(D_I, D_O) = \frac{\sum_{i=1}^{n}(D_I)_i(D_O)_i}{\sqrt{\sum_{i=1}^{n}(D_I)_i^2}\sqrt{\sum_{i=1}^{n}(D_O)_i^2}} \quad (1)$$

Cross entropy is used as defined in Moore and Lewis (2010) to select out-of-domain data that is close to the in-domain corpus. Conversely to the Tf-Idf technique, the data selection is performed at the sentence level. Given the in-domain corpus I, a language model is first computed for the source side ($LM_{I,s}$) and the target side ($LM_{I,t}$) of the bilingual corpus. Similarly, we compute a language model for the out-of-domain corpus O ($LM_{O,s}$ and $LM_{O,t}$) which is drawn from a random sample of the same size of the specialized corpus. The cross entropy is computed as follows:

$$H_{LM}(W) = -\frac{1}{n}\sum_{i=1}^{n}\log(P_{LM}(w_i|w_1, ..., w_{i-1})) \quad (2)$$

where P_{LM} is the probability of a LM for the word sequence W and $w_1, ..., w_{i-1}$ represents the history of the word w_i. Formally, $H_{LM_{I,s}}(W)$ represents the cross entropy of the sentence W given the language model $LM_{I,s}$. The cross entropy is computed for every sentence of the out-of-domain corpus given the out-of-domain and in-domain language models. The source and target sentences are then evaluated by $H_{LM_{I,b}}(W_O) - H_{LM_{O,b}}(W_O)$ (where $b \in \{s, t\}$ refers to the side of the corpus) and ranked accordingly. The cross entropy is computed using the *xenC* tool (Rousseau, 2013).

BERT is a supervised learning model that has proven to be efficient in many downstream NLP tasks (Devlin et al., 2019). It has been trained on the Masked Language Model (MLM) and Next Sentence Prediction (NSP) objectives (Devlin et al., 2019). Hence, it can be used whether for word representation or for sentence classification. In order to perform data selection, we use BERT as a binary classifier to predict if a given input sentence is in-domain or out-of-domain. The intuition behind this strategy is that BERT can learn shared features between in-domain sentences. For each positive (in-domain) training sentence, we randomly select a negative sentence from a general domain data set to keep a balanced training set. Class balancing is often a vital setup for classifiers' efficiency. Even if some existing techniques do reduce the impact of unbalanced classes, we only deal with balanced training in this work.

Random Data Selection we also apply a random data selection at the sentence level, where we shuffle randomly the sentences of the out-of-domain corpora.

4 Experimental Protocol

We present in this section the used data that is: the comparable corpora, the seed lexicon, the reference lists and the methodology applied for our experiments.

4.1 Data

We conducted our experiments on two French/English specialized data sets: breast cancer and wind energy. The Breast Cancer corpus (BC) is composed of scientific documents published between 2001 and 2015 available in open access on the ScienceDirect portal[1] where the title or the keywords contain the term *breast cancer* in English and its translation in French. The Wind Energy corpus (WE) has been released within the TTC project[2] and is composed of documents harvested from the Web using a focused crawler based on several keywords such as *wind*, *energy*, *rotor* in English and their translation in French.

We considered two separate out-of-domain data sets in English and French: i) JRC acquis corpus (JRC) composed of legislative texts of the European Union[3] (we used the French/English version at OPUS which is based on the paragraph-aligned corpus provided by JRC (Tiedemann, 2012)) and ii) a fraction of Wikipedia corpus (WIKI)[4]. The used corpora differ in size and quality. On the one hand, for the specialized corpora, BC is of better quality than WE, due to their construction methods (crawled from a scientific portal vs crawled from the web). On the other hand, for the general corpora, WIKI is way bigger and diverse in terms of topics than JRC. Table 1 shows the size of each side of the corpora and their vocabulary.

	English		French	
Corpus	# tokens	# types	# tokens	# types
BC	525,934	14,800	521,262	11,746
WE	311,898	15,344	656,178	15,799
JRC	64.2M	229,836	70.3M	231,126
WIKI	300M	3M	300M	3.1M

Table 1: Content words' size corpora.

The bilingual terminology reference lists required to evaluate the quality of bilingual terminology induction from comparable corpora are the same as used in (Hazem and Morin, 2016) and was derived from the UMLS meta-thesaurus for the breast cancer domain and are provided with the corpora for wind energy domain (see footnote 2). Each word in the reference list appears at least 5 times in the specialized corpus. The reference list for the breast cancer is composed of a set of 248 French/English single word pairs and of 145 single word pairs for the wind energy domain. We are aware that the reference lists are small, but considering the fact that we are in a specialized domain, they remain representative of its vocabulary, and getting larger lists is difficult since the specialized vocabulary is small.

For the seed lexicon, we employed the ELRA-M0033 dictionary[5] which contains 243,539 pairs of French/English general terms, from which we remove pairs containing words from our reference lists.

4.2 Methodology

Hereafter, we specify the methodology of our experiments:

- For a given specialized corpus, we first use our data selection methods to find the most similar sentences (or documents for the Tf-Idf approach) of our out-of-domain corpora on both source and target languages. For fine-tuning BERT, we used bert-base-cased for English (Devlin et al., 2019),

[1] www.sciencedirect.com/
[2] http://www.ttc-project.eu/
[3] opus.lingfil.uu.se/JRC-Acquis.php
[4] lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
[5] http://catalog.elra.info/en-us/repository/browse/ELRA-M0033/

and Camembert (Martin et al., 2020) for French with their default parameters. To build the training data set, we used the news commentary[6] corpus as out-of-domain data set.

- Then, we create our training data for both languages by concatenating our full specialized corpus with sub-parts of our general corpus. We build several training data sets by incrementally adding samples of 10%, from the most similar to the least similar data (but also from the least similar to the most similar to study the impact of data selection).
- We use fastText (Bojanowski et al., 2016) to train our source and target word embeddings on the previously created training data. (our parameters are as follows: method: skipgram; minCount: 5; dim: 300; ws: 5; minn: 3; maxn: 6).
- To project our word embeddings of source and target languages in the same space, we use the *VecMap* tool (Artetxe et al., 2018a).
- Finally, for a word to translate, we measure its similarity with every word of the target language. The target candidates are ranked using the Cross-domain Similarity Local Scaling which is an improvement of the cosine similarity that takes into account nearest neighbors (Conneau et al., 2017).

The results are measured in terms of the Mean Average Precision (MAP) (Manning et al., 2008):

$$MAP(Ref) = \frac{1}{|Ref|} \sum_{i=1}^{|Ref|} \frac{1}{r_i} \tag{3}$$

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

5 Results

In this section we present the results obtained for BLI on our corpora as well as the differences in computation time for each method.

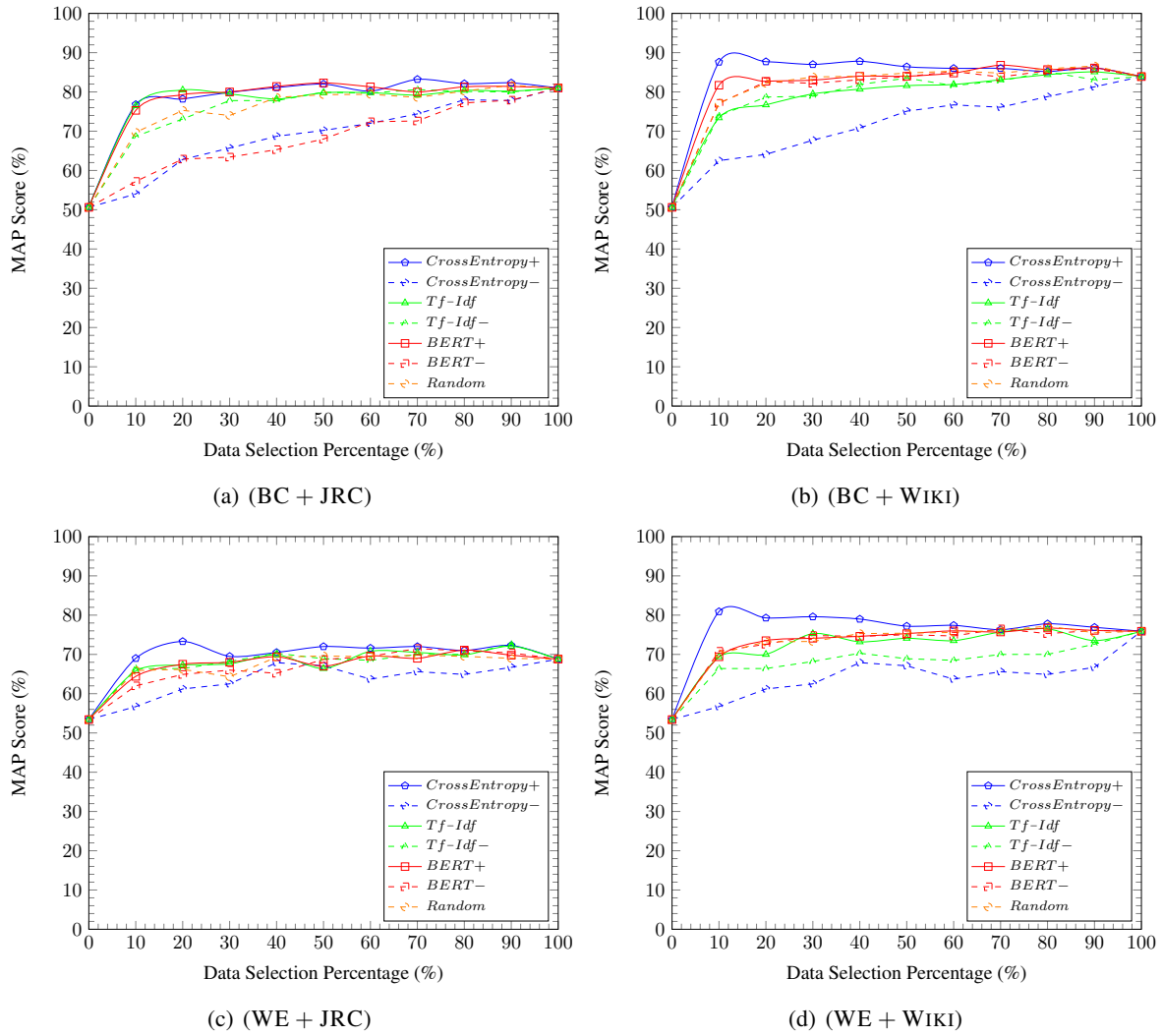
5.1 BLI Evaluation

Figure 1 presents the obtained results for the four data selection methods ($CrossEntropy$, $Tf\text{-}Idf$, $BERT$ and $Random$) for our two specialized corpora and the different combinations with data from the two general corpora. The curves labeled $+$ mean that we take the documents/sentences from the most similar to the least similar, while the curves labeled $-$ indicate that we select the documents/sentences from the least similar to the most similar. This distinction allows to see the real impact of data selection. The first point of all the curves (x axis = 0%) corresponds to the results obtained with the specialized corpus only, and each following point corresponds to a combination of the specialized corpus with a different percentage of data selected from the general corpus, up to the full corpus.

We first note that for every corpus and configuration, adding general data improves the results. If we look at the trends of the $CrossEntropy+$ curves, we see a great improvement at 10% and then a slower improvement for JRC and a slow degradation for WIKI. These results are particularly interesting because they show the usefulness of a good data selection, but they also show the need of a corpus that is large enough to see a real impact on the results. WIKI being larger than JRC, it has more various contexts and 10% of it corresponds to 30M words while it is only 6M words for JRC. Conversely, the $-$ presents the opposite trend and shows slow improvements. The best results are obtained with the full corpora. The curves $+$ for $Tf\text{-}Idf$ and $BERT$ are not as interesting as the $CrossEntropy+$ (with the exception of the BC+WIKI for $BERT$) as there is not a great improvement followed by a slow degradation. Also, we do note observe much differences between the $-$ and $Random$ curves.

Table 2 presents some major points of our experiments. First, we present the results on the general and specialized corpora only. Then we show three results for each combination of specialized and general corpora. The first result (column 100%) is the concatenation of the specialized corpora with the full

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



(a) (BC + JRC) (b) (BC + WIKI)

(c) (WE + JRC) (d) (WE + WIKI)

Figure 1: Contrasting several data selection techniques on the specialized and out-of-domain corpora.

	Specialized Corpus	JRC	*Spec.* + $n\%$ JRC			WIKI	*Spec.* + $n\%$ WIKI		
			100%	70%+	70%-		100%	10%+	10%-
BC	50.6	59.8	81.0	83.2	74.4	82.7	83.9	**87.6**	62.5
WE	53.4	66.4	68.8	72.0	65.6	69.7	75.9	**80.9**	55.5

Table 2: Results (MAP %) on the corpora and their combination for the *CrossEntropy* selection.

general corpora, and illustrates a strategy of data augmentation. The two other columns illustrate the optimal percentage for the *CrossEntropy* data selection approach. The + (resp. −) columns represent the most (least) similar documents of the general corpora.

We clearly see improvements for both specialized corpora with the WIKI corpus when going from a full data augmentation (100%) to data selection (10%). In terms of MAP, this makes us gain 3.7 points for BC and 5 points for WE. However, for the smallest JRC, the gain is less important. If we still have an interesting MAP improvement of 2.2% and 3.2% for BC and WE respectively, it needs way more data to be effective (70%). These observations confirm the usefulness of data selection, but under the condition of using large enough corpora with various content.

Based on the results of *CrossEntropy*+ at 10% of data for WIKI, we conducted one more experi-

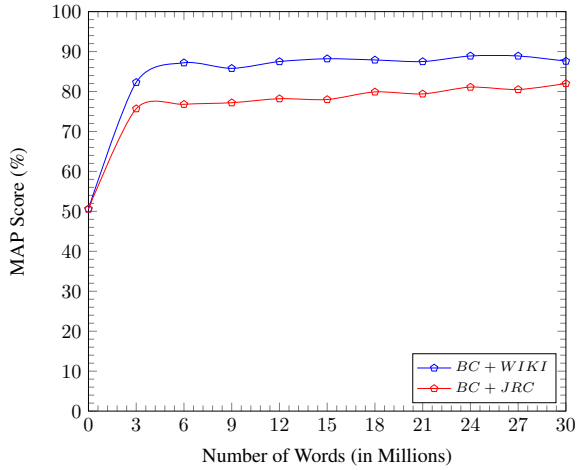


Figure 2: Results on $CrossEntropy$, on lower percentage of the general corpus.

ment on smaller percentages to get better insights on smaller data selection effect. This experiment is illustrated in Figure 2 where the evolution of data quantity is represented in terms of number of words, to allow the comparison with the JRC corpus (3M words represent 1% of WIKI and 5% of JRC). Overall, we see that the two curves keep the same trend, with better MAP scores for WIKI. Interestingly, we observe that the selection of only 3M words from WIKI (1% of the corpus) allows to reach a MAP score of 82.3% which is almost the same a the whole corpus (83.9%). This result is however surpassed when selecting 6M words (2% of the corpus) reaching a MAP score of 87.2%. The best results for WIKI are achieved at 8 and 9% with a MAP score of 88.9%. For JRC, the improvement with 3M words (5% of the corpus) remains interesting (75.7%) but does not surpass the previous best results (83.2%). These observations show the necessity of using a corpus related to many diverse topics to obtain the best results.

5.2 Computation Time

In this section, we report the computation time for all the configurations. All our experiments have been carried out with an Intel Core i9-9900K and a GeForce RTX 2080.

Data Selection Type		Selection Time (s)	Optimal Data Percentage[7]	Embeddings Training Time (s)	MAP Score
No Selection	BC	-	0	**180**	50.6
Augmentation	BC+WIKI	-	100	65,521	83.9
Selection	$CrossEntropy$	420	10	6,552	**87.6**
	$Tf\text{-}Idf$	339	90	58,968	85.1
	$BERT$	147,467†	70	45,865	86.8

Table 3: Computation time in seconds for the selection and word embeddings training time for BC and WIKI (†GPU time).

Table 3 shows the computation time needed to perform data selection and to train the word embeddings for BC and WIKI. The embeddings training time corresponds to the computation time of the source and target embeddings. We see that using in-domain corpus only (No Selection) is really fast ($180s$) but does not present good results (MAP of 50.6%) while a data augmentation approach improves the results (83.9%) but also drastically increase the computation time ($65,521s$). Finally, with a good data selection approach ($CrossEntropy$), the augmentation of computation time is smaller ($6,972s$) and even improves the MAP score (87.6%). $BERT$ and $Tf\text{-}Idf$ selections also present better results than the data augmentation approach, but no real improvement in terms of computation time is observed and the MAP increase is less interesting than for $CrossEntropy$.

6 Qualitative analysis

BC + n% WIKI	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
En Occ (Breast)	$3.6k$	$6.5k$	$7.3k$	$7.8k$	$8.2k$	$8.5k$	$8.7k$	$8.8k$	$8.9k$	$8.9k$	$9.0k$
Fr Occ (Sein)	$2.9k$	$7.3k$	$12.3k$	$17.8k$	$23.8k$	$30.2k$	$37.1k$	$44.3k$	$51.5k$	$57.0k$	$59.4k$
Rank	2	3	12	604	399	933	≥ 1000	≥ 1000	≥ 1000	≥ 1000	≥ 1000
En Occ (Calcium)	23	$1.8k$	$2.2k$	$2.3k$	$2.4k$	$2.5k$	$2.6k$	$2.6k$	$2.7k$	$2.7k$	$2.7k$
Fr Occ (Calcium)	14	983	$1.3k$	$1.5k$	$1.7k$	$1.8k$	$1.8k$	$1.9k$	$1.9k$	$2.1k$	$2.2k$
Rank	140	1	1	1	1	1	1	1	1	1	1
En Occ (Back)	27	$7.3k$	$19.5k$	$33.8k$	$49.7k$	$66.6k$	$84.0k$	$102.1k$	$120.2k$	$139.8k$	$153.9k$
Fr Occ (Dos)	7	883	$2.0k$	$3.3k$	$4.6k$	$6.0k$	$7.4k$	$8.8k$	$10.3k$	$11.9k$	$14.0k$
Rank	≥ 1000	37	12	4	5	9	14	50	43	27	99
En Occ (Lymphoscintigraphy)	20	20	20	20	20	20	20	20	20	20	20
Fr Occ (Lymphoscintigraphie)	27	28	28	28	28	28	28	28	28	28	28
Rank	4	1	1	1	1	1	1	1	1	1	1

Table 4: 4 interesting translation pairs over several data selection percentages from $CrossEntropy+$. The optimal selection is for 10% as seen in the previous section.

The results showed in Section 5 establish a relationship between data selection and the quality of bilingual lexicons. More precisely, they show that we can improve the results while reducing the computation time in comparison with data augmentation. The results obtained when contrasting one document-level selection ($Tf\text{-}Idf+$) and two sentence-level selections ($CrossEntropy+$ and $BERT+$) confirmed the usefulness of using data selection and, more importantly, stressed the fact that a small amount of data could be sufficient for that purpose. If the main issue is to better characterize specialized terms while preserving the domain characteristics, we assume that the representations of common (i.e. from a general domain) words present in the general corpus are also enriched. In order to measure the augmentation impact on BLI, we selected several translation pairs that were affected either positively or negatively by data increase and discuss each case.

Table 4 shows for each data selection amount, the rank of a given translation term as well as the occurrence of the translation pair after data augmentation for BC. To each rank we assign a color according to the impact of the selection on the results. Hence, green means that the obtained rank after adding n% of data equalized or improved the translation rank while red means that the rank was degraded. This table presents the results for the $CrossEntropy+$ lists where we had optimal results at 10% of data selected.

First, we see that the term *breast* is -almost- correctly translated to *sein* (Rank = 2) when no data is added. However, with 100% of the general corpus, the rank is hugely degraded (Rank ≥ 1000). A closer look at the added contexts revealed that the majority of documents contained the French expression *au sein de* (appeared $47,025$ times) which can be translated as *within* or *at the heart of*. This pair is one of the few that doesn't improve with the addition of out-of-domain data, but we can see that taking only 10% (Rank = 3) leads to a significant improvement compared to the full general corpus (Rank ≥ 1000).

Contrariwise, *calcium* showed the benefits of adding correct contexts on both sides. If its correct translation *calcium* was ranked at 140 when only specialized corpus was used, the rank is immediately improved by the addition of out-of-domain data to rank 1. It is also important to note that the source and the target words remains close in terms of frequency. This pair also shows that the selection succeeded to find most of the interesting occurrences of the words in the first 10%.

The pair *back* - *dos* shows the same trend as *breast* - *sein*, mainly due to the frequent use of *back* in English. However, adding out-of-domain data didn't decrease the rank. On the contrary, it had a quite positive impact, mostly because at first, there were few occurrences of *dos* in BC.

Finally, our last pair shows that the addition of out-of-domain data remains interesting even if the words pair are -almost- not found in the general corpora. Here, with only one new occurrence and the improved representation of other words of the specialized vocabulary, rank one could be reached.

These examples reflect the benefits of a data selection approach, which helps enriching the specialized vocabulary as a whole without degrading the representation of the terms of the reference list.

In what follows, we briefly illustrate some examples of the type of contexts that have been selected. Table 5 presents the result of the sentences selected by the $CrossEntropy$ on WIKI for the specialized corpus BC. It is interesting to see that the best selected sentences ($CrossEntropy+$) do not look like real sentences, but are closer to phrases related to the domain (*tumor, gene, operate, needle...*), which really shows that $CrossEntropy+$ selected contexts related to the specialized vocabulary. However, we find real sentences among the highest ranked (see line 115), but the $CrossEntropy$ seems to favor phrases. As intended, for the least interesting sentences ($CrossEntropy-$), we can't find any relation between the breast cancer domain and the proposed sentences.

Sentences	$CrossEntropy+$	$CrossEntropy-$
1	tumor suppressor gene	marginal worker household industry worker
2	receiver operate characteristic	notable former player
3	single photon emission compute tomography	main worker household industry worker
4	fine needle aspiration	id bass value blue legend bass
...	...	...
115	breast density is positively associated with breast cancer	-

Table 5: Example of sentences on WIKI selected for BC by the $CrossEntropy$ approach.

Finally, Table 6 illustrates the evolution of the candidate translations over three stages of data selection. The words in bold are the correct translations. The first pair *pressure-pression* illustrates the problem of having only a small specialized corpus, where the words do not have enough context to be precisely represented. For the second pair *breast-sein*, we already have two words with a lot of occurrences, so we do not have this problem of bad representation of the words, even if some candidates should not be here (*année, der...*). And the data augmentation column (100%) shows the introduction of polysemy: we can't see *sein* as a correct translation because of the French preposition *au sein de*, but we still have related terms like *poitrine*. This problem is mostly resolved by the data selection column (10%), where *sein* is still in the top candidates (rank 3).

Pair	0%	10%	100%	Pair	0%	10%	100%
	année	cancer	poitrine		demi-vie	**pression**	**pression**
Breast	**sein**	ovaire	cuisse	Pressure	résorber	tension	tension
-	der	**sein**	cou	-	millimètre	compression	poussée
Sein	mammaire	prostate	épaule	Pression	résorption	poussée	aspiration
	cancer	mammaire	utérus		gradient	dilatation	instabilité

Table 6: Examples of the evolution of the candidate translations of two test pairs: *pressure-pression* and *breast-sein* for the $CrossEntropy$ selection on WIKI for BC.

7 Conclusion

In order to improve BLI from specialized comparable corpora, we have studied in this paper the impact of common data selection techniques such as $CrossEntropy$, $Tf\text{-}Idf$ and also applied $BERT$ for the first time as a data selection technique. The experiments on two specialized comparable corpora and two general corpora revealed that selecting a small and adapted amount of out-of-domain data is sufficient to obtain better results (MAP of 87.6% for BC and 80.9% for WE) than a high amount of data (resp. 83.9% and 75.9%) while reducing computation time by a factor 10. The data selection strategy also allowed to reduce the impact of polysemy compared to data augmentation. This strategy shows that bringing new occurrences for both the word to translate and the general vocabulary improves their mutual discrimination. An interesting perspective could be to select for each word to translate the adequate amount of out-of-domain data instead of a common selection for all the terms to translate.

Acknowledgments

This research has received funding from the French National Research Agency under grant ANR-17-CE23-0001 ADDICTE (Distributional analysis in specialized domain) as well as Atlanstic 2020 and the Canadian Institute for Data Valorisation (IVADO).

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, New Orleans, LA, USA.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pages 789–798, Melbourne, Australia.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 355–362, Edinburgh, Scotland, UK.

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.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*, pages 4171–4186, Minneapolis, Minnesota.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2019. Embedding strategies for specialized domains: Application to clinical entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 295–301, Florence, Italy.

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. 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.

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.

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. 2018. Leveraging meta-embeddings for bilingual lexicon extraction from specialized comparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*, pages 937–949, Santa Fe, NM, USA.

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.

Audrey Laroche and Philippe Langlais. 2010. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Bo Li and Éric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 644–652, Beijing, China.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, pages 7203–7219, Online, July. Association for Computational Linguistics.

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

Robert C. Moore and William D. Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 220–224, Uppsala, Sweden.

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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18)*, pages 2227–2237, New Orleans, LA, USA.

Emmanuel Prochasson, Emmanuel Morin, and Kyo Kageura. 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings of the 12th Conference on Machine Translation Summit (MT Summit XII)*, pages 284–291, Ottawa, Canada.

Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff. 2020. Overview of the fourth BUCC shared task: Bilingual dictionary induction from comparable corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora (BUCC'20)*, pages 6–13, Marseille, France.

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.

Anthony Rousseau. 2013. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, (100):73–82.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2013. Overviewing important aspects of the last twenty years of research in comparable corpora. In S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung, editors, *Building and Using Comparable Corpora*, pages 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg.

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, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, pages 479–484, Portland, OR, USA.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Yi Lu, and Junwen Xing. 2014. A Systematic Comparison of Data Selection Criteria for SMT Domain Adaptation. *The Scientific World Journal*, 2014:10.

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.