Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora

Amir Hazem, Emmanuel Morin

Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora

Amir Hazem[1] **Emmanuel Morin**[1]

[1] LINA - UMR CNRS 6241, Université de Nantes, France
{amir.hazem,emmanuel.morin}@univ-nantes.fr

Abstract

Comparable corpora are the main alternative to the use of parallel corpora to extract bilingual lexicons. Although it is easier to build comparable corpora, specialized comparable corpora are often of modest size in comparison with corpora issued from the general domain. Consequently, the observations of word co-occurrences which are the basis of context-based methods are unreliable. We propose in this article to improve word co-occurrences of specialized comparable corpora and thus context representation by using general-domain data. This idea, which has been already used in machine translation task for more than a decade, is not straightforward for the task of bilingual lexicon extraction from specific-domain comparable corpora. We go against the mainstream of this task where many studies support the idea that adding out-of-domain documents decreases the quality of lexicons. Our empirical evaluation shows the advantages of this approach which induces a significant gain in the accuracy of extracted lexicons.

1 Introduction

Comparable corpora are the main alternative to the use of parallel corpora for the task of bilingual lexicon extraction, particularly in specialized and technical domains for which parallel texts are usually unavailable or difficult to obtain. Although it is easier to build comparable corpora (Talvensaari et al., 2007), specialized comparable corpora are often of modest size (around 1 million words) in comparison with general-domain comparable corpora (up to 100 million words) (Morin and Hazem, 2016). The main reason is related to the difficulty to obtain many specialized documents in a language other than English. For example, a single query on the Elsevier portal[1] of documents containing in their title the term "breast cancer" returns 40,000 documents in English, where the same query returns 1,500 documents in French, 693 in Spanish and only 7 in German.

The historical context-based approach dedicated to the task of bilingual lexicon extraction from comparable corpora, and also known as the *standard approach*, relies on the simple observation that a word and its translation tend to appear in the same lexical contexts (Fung, 1995; Rapp, 1999). In this approach, each word is described by its lexical contexts in both source and target languages, and words in translation relationship should have similar lexical contexts in both languages. To enhance bilingual lexicon induction, recent approaches use more sophisticated techniques such as topic models based on bilingual latent dirichlet allocation (*BiLDA*) (Vulic and Moens, 2013b; Vulic and Moens, 2013a) or bilingual word embeddings based on neural networks (Gouws et al., 2014; Chandar et al., 2014; Vulic and Moens, 2015; Vulic and Moens, 2016) (approaches respectively noted: *Gouws*, *Chandar* and *BWESG+cos*). All these approaches require at least sentence-aligned/document aligned parallel data (*BiLDA*, *Gouws*, *Chandar*) or non-parallel document-aligned data at the topic level (BWESG+cos). Since specialized comparable corpora are of small size, sentence-aligned (document aligned) parallel data are unavailable and non-parallel document-aligned data at the topic level can't be provided since specialized comparable corpora usually deal with one single topic. Based on the recent comparison in (Vulic and Moens, 2015; Vulic and Moens, 2016) where the standard approach (noted in there article as *PPMI+cos*) performed better in most cases while compared to *BiLDA*, *Gouws* and *Chandar*, and due to the unavailability of non parallel

[1] www.sciencedirect.com

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3401–3411, Osaka, Japan, December 11-17 2016.

document aligned data at the topic level, we only deal with the standard approach and show at least that our approach improve drastically bilingual terminology extraction while adding well selected external data.

The small size of specialized comparable corpora renders unreliable word co-occurrences which are the basis of the standard approach. In this paper, we propose to improve the reliability of word co-occurrences in specialized comparable corpora by adding general-domain data. This idea has already been successfully employed in machine translation task (Moore and Lewis, 2010; Axelrod et al., 2011; Wang et al., 2014, among others). The approach of using adapted external data, also known as *data selection* is often applied in Statistical Machine Translation (SMT) to improve the quality of the language and translation models, and hence, to increase the performance of SMT systems. If data selection has become a mainstream in SMT, it is still not the case in the task of bilingual lexicon extraction from specialized comparable corpora. The majority of the studies in this area support the principle that the quality of the comparable corpus is more important than its size and consequently, increasing the size of specialized comparable corpora by adding out-of-domain documents decreases the quality of bilingual lexicons (Li and Gaussier, 2010; Delpech et al., 2012). This statement remains true as long as the used data is not adapted to the domain. We propose two data selection techniques based on the combination of a specialized comparable corpus with external resources. Our hypothesis is that word co-occurrences learned from a general-domain corpus for general words (as opposed to the terms of the domain) improve the characterization of the specific vocabulary of the corpus (the terms of the domain). By enriching the general words representation in specialized comparable corpora, we improve their characterization and therefore improve the characterization of the terms of the domain for better discrimination.

The remainder of this article is organized as follows: Section 2 describes the standard approach to bilingual lexicon extraction from comparable corpora. Section 3 presents previous works related to the improvements of the standard approach for specialized comparable corpora. Section 4 describes our strategies to improve the characterization of lexical contexts. Section 5 presents the different textual resources used for our experiments: the specialized and general comparable corpora, the bilingual dictionary and the terminology reference lists. Section 6 evaluates the influence of using lexical contexts built from general comparable corpora on the quality of bilingual terminology extraction. Section 7 presents our conclusions.

2 Standard Approach

Bilingual lexicon extraction from comparable corpora relies on the simple assumption that a word and its translation tend to appear in the same lexical contexts. Based on this assumption, the standard approach can be carried out by applying the following steps:

1. Build for each word w of the source and the target languages a context vector (resp. $\mathbf{s}$ and $\mathbf{t}$ for source and target languages) by identifying the words that appear in a window of n words around w normalized according to the measure of association of each word in the context of w. The association measures studied are Mutual Information (Fano, 1961), Log-likelihood (Dunning, 1993), and the Discounted Odds-Ratio (Evert, 2005).

2. Translate with a bilingual dictionary the context vector of a word to be translated from the source to the target language ($\bar{\mathbf{i}}$ the translated context vector).

3. Compare the translated context vector $\bar{\mathbf{i}}$ to each context vector of the target language $\mathbf{t}$ through a similarity measure and rank the candidate translations according to this measure. The similarity measures employed are Cosine (Salton and Lesk, 1968) and weighted Jaccard (Grefenstette, 1994)

3 Related Work

In the past few years, several contributions have been proposed to improve each step of the standard approach. Prochasson et al. (2009) enhance the representativeness of the context vectors by strengthening the context words that happen to be transliterated and scientific compound words in the target

language. Ismail and Manandhar (2010) also suggest that context vectors should be based on the most important contextually relevant words (in-domain terms), and thus propose a method for filtering the noise of the context vectors. Bouamor et al. (2013) propose an adaption of the standard approach that exploits Wikipedia to improve the context vector representation. From the context vector of a word to be translated, they build a vector of Wikipedia concepts using the ESA inverted index (Explicit Semantic Analysis). This vector of concepts is then translated into the target language. The candidate translations are found by projecting the translated vector of concepts using the ESA direct index onto the context vector of the target language. Prochasson and Fung (2011) propose to use a machine learning approach based both on the context-vector similarity and the co-occurrence features to learn a model for rare words from one pair of languages and this model can be used to find translations from another pair of languages. Hazem and Morin (2013) study different word co-occurrence prediction models in order to make the observed co-occurrence counts in specialized comparable corpus more reliable by re-estimating their probabilities. Morin and Hazem (2016) show the unfounded assumption of the balance in terms of quantity of data of the specialized comparable corpora and that the use of unbalanced corpora significantly improves the results of the standard approach.

Other improvements to the standard approach have been proposed by introducing other paradigms. For instance, Gaussier et al. (2004) propose to apply Canonical Correlation Analysis (CCA) which is a bilingual extension of Latent Semantic Analysis (LSA) whereas Hazem and Morin (2012) propose to use Independent Component Analysis (ICA) which is basically an extension of the Principal Component Analysis (PCA). Vulić et al. (2011) also propose an extension of the Latent Dirichlet Allocation (LDA) taking into account bilinguality and called bilingual LDA (BiLDA), improvements of this latter can be found in (Vulic and Moens, 2013b; Vulic and Moens, 2013a). Gouws et al. (2014) and Chandar et al. (2014) use multilingual word embeddings based on sentence-aligned parallel data and/or translation dictionaries whereas Vulić and Moens. (2015; 2016) learn bilingual word embeddings from non-parallel document aligned data based on skip-gram model. These approaches are beyond the scope of this study because even if they improve the standard approach they are intended for large comparable corpora of general language and/or require parallel aligned data or non parallel aligned documents which are unavailable for specialized corpora. In this paper, we give a particular interest to the massive amount of general domain data that can be found on the web and discuss ways of taking advantage of these resources in order to enrich word context representation and improve the standard approach.

4 Adapted Standard Approach

We propose two adaptations of the standard approach. Based on the assumption that general domain information can benefit the task of bilingual lexicon extraction from specialized corpora, we enhance the standard approach for that purpose by jointly exploiting data from specialized and general domains.

4.1 Global Standard Approach

The first adaptation of the standard approach can be described as basic. It consists to build the context vectors from a comparable corpus composed of the specialized and the general comparable corpora. This adaptation is inspired by the work of Morin et al. (2010) that shows that the discourse categorization (scientific versus popular scientific documents) of the documents in a specialized comparable corpus increases the quality of the extracted French/Japanese lexicon composed of single-word terms despite the data sparsity. For alignment of multi-word terms, the discourse categorisation of documents is not relevant. This work suggests that increasing the size of the specialized comparable corpora by adding popular scientific documents is interesting.

4.2 Selective Standard Approach

In the second adaptation, we first build independently word' context vectors of the two corpora (specialized and general) and then, for each word that belongs to the specialized domain corpus, if it appears in the general domain corpus, we merge its specialized and general context vectors. This allows to filter general domain words that are not part of the specialized corpus and renders the selective standard ap-

proach much less time consuming than the global standard approach. The merging process carried out before the normalization of context vectors of the standard approach (see step 1 - Section 2) is done as follows:

Increasing word co-occurrence counts (Hyp1) if a word w_i co-occurs p times with w in the specialized domain and q times in the general domain, we simply add the two co-occurrence counts so that the merged context vector of w will contain w_i with a co-occurrence count of $p + q$.

Reducing the vector space model sparseness (Hyp2) if a word w_j co-occurs r times with w in the general domain but does not co-occur with w in the specialized domain, we add w_j to the merged context vector of w. In that case, w_j is considered to be as new information that is added to the context vector of w to enrich it.

By enhancing word co-occurrence counts, the context vectors of the words become more reliable. Whereas, by increasing the density of the vector space model, the context vectors of the words become more precise. This twofold strategy enables us to better characterize the words of the specialized comparable corpus without increasing the number of words to characterize. In this way, the candidate translations of a word are always selected from the vocabulary of the specialized comparable corpus.

In the same way that Hazem and Morin (2013), we use a general language corpus to make the observed word co-occurrence counts in a specialized comparable corpus more reliable. Like them, we modify the initial word co-occurrence counts, but unlike them, we introduce new words learned from the general corpus in the vector space model.

5 Data and Resources

In this section, we describe the data and resources we used in our experiments which are conducted on the French/English language pair.

5.1 Comparable Corpora

The specialized comparable corpora were selected in terms of bilingual terminology access of technical domains. For this purpose, comparable corpora gather texts sharing common features such as domain, topic, genre, discourse and period without having a source text-target text relationship which guarantees access to the original vocabulary. For our experiments, we used three French/English specialized comparable corpora:

Breast cancer corpus (BC) is composed of documents collected from the Elsevier website[1]. We have selected the documents published between 2001 and 2008 where the title or the keywords contain *cancer du sein* in French and *breast cancer* in English.

Volcanology corpus (VG) was manually built by gathering documents dedicated to volcanology such as web documents, academic textbooks, popular science books, general newspapers, popular and semi-popular science magazines, travel magazines, and glossaries.

Wind energy corpus (WE) has been released in the TTC project[2]. This corpus has been crawled from the web using *Babouk* crawler (Groc, 2011) based on several keywords such as *vent*, *énergies*, *éolien*, *renouvelable* in French and *wind*, *energy*, *rotor* in English.

In order to evaluate our approach , we explored different types and size of external data. Most of them are parallel corpora often used in multiple evaluation campaigns such as WMT[3]. It is to note that we do not take advantage of the parallel information. Using parallel corpora only insures a good degree of comparability. We briefly describe each corpus:

[2] www.ttc-project.eu
[3] www.statmt.org

Corpus	# content words		# distinct words		Comp.
	FR	EN	FR	EN	
Breast cancer	521,262	525,934	6,630	8,221	79.07
Volcanology	399,828	405,286	9,142	8,623	83.69
Wind energy	313,954	314,551	5,346	6,378	81.61
NC	5.7M	4.7M	23,597	29,489	88.52
EP7	61.8M	55.7M	40,861	46,669	87.90
JRC	70.3M	64.2M	100,004	93,104	85.30
CC	91.3M	81.1M	250,999	259,226	86.13
GW	353.4M	291.8M	299,784	323,280	85.56
UN	421.7M	361.9M	158,647	137,411	84.73

Table 1: Characteristics of the specialized corpora and the external data.

News commentary corpus (NC) is a twelve language parallel corpus of news commentaries provided by the WMT workshop for SMT[4].

Europarl corpus (EP7) is a parallel corpus for SMT extracted from the proceedings of the European Parliament. It contains about 21 languages. We used the French-English version 7 used for the WMT translation task[3]

JRC acquis corpus (JRC) is a collection of legislative European union texts[4]. We used the French-English aligned version at OPUS provided by JRC (Tiedemann, 2012).

Common crawl corpus (CC) is a petabytes of data collected over 7 years of web crawling set of raw web page data and text extracts[5].

Gigaword corpus (GW) is a set of monolingual newswire corpora provided by LDC[6].

United nations corpus (UN) is a six language parallel text of the United Nations originally provided as translation memory (Rafalovitch and Dale, 2009).

The French/English corpora were then normalized through the following linguistic pre-processing steps: tokenization, part-of-speech tagging, and lemmatization. Finally, the function words were removed and the words occurring less than twice in the French and in the English parts were discarded. Table 1 shows the size of the comparable corpora and also indicates the comparability degree in percentages (Comp.) between the French and the English parts of each comparable corpus. The comparability measure (Li and Gaussier, 2010) is based on the expectation of finding the translation for each word in the corpus and gives a good idea about how two corpora are comparable. We can notice that all the comparable corpora have a high degree of comparability.

5.2 Bilingual Dictionary

The bilingual dictionary used in our experiments is the French/English dictionary ELRA-M0033[7]. This resource is a general language dictionary which contains around 244,000 entries.

5.3 Gold Standard

To evaluate the quality of bilingual terminology extraction from comparable corpora, a bilingual terminology reference list that reflects the technical vocabulary of the comparable corpus is required. The

[4] opus.lingfil.uu.se
[5] commoncrawl.org
[6] www.ldc.upenn.edu
[7] www.elra.info

list is usually composed of more or less 100 single words: 95 single words in Chiao and Zweigenbaum (2002), 100 in Morin et al. (2010), 125 and 79 in Bouamor et al. (2013a). We build a reference list for each of the three comparable corpora using specialized glossaries available on the Web. For instance, the list is derived from the UMLS[8] for the breast cancer corpus. Concerning wind energy, the list is provided with the corpus[1]. In order to focus only on the vocabulary characteristic of the specialized corpus we remove technical terms that have a common meaning in the general domain such as *analysis*, *factor*, *method*, *result*, *study*, etc. Without this precaution, these terms would be mechanically better identified in a larger corpus. To discard these terms, we use for French the list of the ScienTexT Project[9] and for English the Academic Keyword List[10]. Each word of the reference lists appears at least 5 times in the specialized comparable corpus. The reference lists are composed of 248 terms for breast cancer, 156 terms for volcanology and 139 terms for wind energy.

6 Experiments

Table 2 shows the results of the standard approach (noted SA) using only specialized comparable corpora (BC, VG and WE) or using only external data (NC, EP7, JRC, CC, GW and UN). It also shows the two adapted standard approaches (noted GSA and SSA) using the combination of each specialized comparable corpus with each corpus of the external data. The scores are measured in terms of the Mean Average Precision (MAP). We also used the three most exploited association and similarity measure configurations: Mutual Information with Cosine (noted MI-COS), Discounted Odds-Ratio with Cosine (noted OR-COS) and finally, Log-likelihood with weighted Jaccard (noted LL-JAC).

The first column of Table 2 shows the results of the SA for the three specialized comparable corpora. We can see that for each corpus the results differ according to a given measure configuration. Overall, for SA, the best results are obtained using the LL-JAC configuration. The SA for instance obtains a MAP score of 34.6% using BC corpus and a MAP score of 50.4% using VG corpus.

From the second to the seventh column, Table 2 shows the results of the SA using external data only, and our two adapted approaches (GSA and SSA). Column four for instance, shows the results of SA that uses the JRC corpus only. It also shows the results of GSA and SSA that combine the JRC corpus with each specialized corpus. GSA for instance obtains a MAP score of 63.3% and SSA a MAP score of 66.8% while combining the BC corpus with the JRC corpus (MI-COS configuration). Comparatively, and for the same configuration, SA using JRC corpus only, obtains a MAP score of 53.2%.

The first comment concerns the SA where surprisingly, using external data only, almost always improves its performance. This is particularly noticed when using external data of large size such as CC, GW and UN corpora. The good results obtained using these latter corpora can be explained by their characteristics. The Common crawl corpus (CC) for instance which has been crawled from the web, contains many scientific and specialized documents that can improve context representation. In addition, its large size makes co-occurrence counts more reliable. According to Table 3 we can see that more than 90% of the distinct words of the specialized corpora are present in the large general domain corpora.

The second comment concerns GSA and SSA where both always outperform SA for all the configurations. For the BC corpus for instance, we can notice that GSA obtains a MAP score of 81.5% and SSA obtains a MAP score of 83.4% using the GW corpus (LL-JAC configuration) while SA obtains a MAP score of 34.6% using BC and a MAP score of 78.3% using the GW corpus. Using other external data also improves the results of SA using the BC corpus. For instance, SSA obtains a MAP score of 65.9% using JRC, 57.5% using EP7 and 57.8% using NC. This means that adding external data always benefits bilingual lexicon extraction. If both GSA and SSA always improve bilingual lexicon extraction for the three specialized corpora, the results of Table 2 show that SSA outperforms GSA for almost all the configurations. This means that enriching the words that belong to the specialized domain corpus, if they appear in the general domain corpus (by merging context vectors) is more efficient than using a global combination (GSA). In addition, it should be noted that SSA is much faster than GSA.

[8] www.nlm.nih.gov/research/umls

[9] scientext.msh-alpes.fr

[10] www.uclouvain.be/en-372126.html

	BC	NC	EP7	JRC	CC	GW	UN	
SA	25.9	44.9	49.8	53.2	75.8	83.6	57.9	MI-COS
GSA	-	55.8	60.1	63.3	80.7	85.0	66.7	
SSA	-	**57.8**	**60.9**	**66.8**	**81.6**	**85.6**	**67.1**	
SA	27.0	45.3	48.5	52.0	75.5	81.1	55.7	OR-COS
GSA	-	**58.9**	58.3	61.7	80.2	83.2	58.9	
SSA	-	**58.9**	**60.8**	**66.6**	**82.3**	**85.5**	**67.2**	
SA	34.6	45.4	45.4	49.3	72.8	78.3	50.7	LL-JAC
GSA	-	57.4	56.3	63.0	77.2	81.5	62.0	
SSA	-	**57.8**	**57.5**	**65.9**	**78.7**	**83.4**	**65.5**	

(a) Breast cancer corpus

	VG	NC	EP7	JRC	CC	GW	UN	
SA	22.7	47.9	50.0	51.7	77.5	75.0	62.7	MI-COS
GSA	-	55.1	58.1	61.3	**78.3**	**78.7**	68.6	
SSA	-	**57.5**	**60.7**	**64.4**	78.1	76.0	**68.7**	
SA	37.9	49.7	50.2	49.3	75.6	73.9	59.4	OR-COS
GSA	-	61.6	60.4	59.0	77.2	76.3	67.5	
SSA	-	**62.2**	**61.3**	**62.3**	**78.5**	**78.8**	**68.5**	
SA	50.4	48.4	45.8	45.1	71.2	68.7	50.9	LL-JAC
GSA	-	63.0	60.6	58.3	**73.3**	70.6	59.3	
SSA	-	**64.0**	**62.4**	**58.9**	73.2	**72.8**	**61.2**	

(b) Volcanology corpus

	WE	NC	EP7	JRC	CC	GW	UN	
SA	15.6	41.0	51.0	63.4	72.1	67.4	60.4	MI-COS
GSA	-	47.3	**54.6**	65.3	73.2	69.1	64.1	
SSA	-	**50.5**	53.2	**67.8**	**74.9**	**70.8**	**66.9**	
SA	19.4	45.4	50.0	60.8	71.3	68.1	58.3	OR-COS
GSA	-	52.3	51.8	64.2	72.3	70.3	60.8	
SSA	-	**52.8**	**53.9**	**66.8**	**74.8**	**72.5**	**63.7**	
SA	28.0	43.6	45.1	60.0	65.0	62.5	48.6	LL-JAC
GSA	-	42.9	46.0	59.7	64.7	63.0	50.8	
SSA	-	**43.8**	**48.7**	**61.6**	**66.2**	**65.7**	**53.6**	

(c) Wind energy corpus

Table 2: Results (MAP %) of the *Standard Approach* (SA), the *Global Standard Approach* (GSA) and the *Selective Standard Approach* (SSA) for the breast cancer corpus (BC), the volconalogy corpus (VG) and the wind energy corpus (WE) using the news commentary corpus (NC), the Europarl corpus (EP7), the JRC acquis corpus (JRC), the common crawl corpus (CC), the Gigaword corpus (GW) and the united nation corpus (UN) (the improvements indicate a significance at the 0.001 level using the Student t-test).

		BC+NC	BC+EP7	BC+JRC	BC+CC	BC+GW	BC+UN
# Hyp1(∩)	FR	3,939	4,366	4,789	5,502	**5,907**	5,142
	EN	4,315	4,668	5,451	6,303	**7,103**	5,701
# Hyp2(∪)	FR	721	1,931	1,067	3,211	**4,503**	3,330
	EN	746	1,767	1,013	2,833	**4,952**	3,215
		VG+NC	VG+EP7	VG+JRC	VG+CC	VG+GW	VG+UN
# Hyp1(∩)	FR	6,472	7,184	6,808	**8,426**	8,330	7,901
	EN	6,190	6,581	6,214	7,825	**7,864**	7,142
# Hyp2(∪)	FR	556	1,480	861	2,872	**3,910**	2,700
	EN	614	1,436	904	2,829	**4,866**	2,827
		WE+NC	WE+EP7	WE+JRC	WE+CC	WE+GW	WE+UN
# Hyp1(∩)	FR	3,804	4,136	4,535	4,909	**4,944**	4,770
	EN	4,246	4,582	5,071	5,546	**5,767**	5,331
# Hyp2(∪)	FR	790	2,135	1,204	3,842	**5,531**	3,663
	EN	784	1,901	1,174	3,422	**6,350**	3,715

Table 3: Number of distinct context vectors that have been augmented (enriched).

SSA translation candidates are those of the specialized domain only (around 6,600 candidates for the BC corpus) and GSA translation candidates are those of the specialized domain plus those of the general domain (around 250,000 candidates for CC corpus - see Table 1) which render the computation of vector similarity much more time consuming. Overall, we can see that the results differ according to the configuration measures used. If for SA, the best results are always obtained using LL-JAC, this is not the case for GSA and SSA. For the BC corpus for instance, SSA obtains the highest MAP score of 85.6% using GW and the MI-COS configuration while for the VG corpus combined with GW, we can see that the best MAP score of 78.8% is obtained by SSA using the OR-COS configuration. These differences are mainly due to the measure properties. If the MI measure shows poor results on small corpora, it is mainly because it overestimates low counts and underestimates high counts. This disadvantage is smoothed when using more data. The differences between MI and OR measures are too low to conclude which is the most appropriate one to use as we obtain more or less equivalent results for the used corpora.

Table 3 shows the number of distinct words of each specialized corpus that have been enriched using each general-domain corpus. Hyp1 corresponds to the first hypothesis of SSA in which we assume that only the context vectors of the specialized corpora should be enriched. So the Hyp1 column shows the number of distinct words that appear in both the specialized and the general domain corpora. For instance, Hyp1 of the BC corpus and the NC corpus noted BC+NC, shows that there are 4,315 words in common for their English parts and 3,939 in common for their French parts. One can notice that a high amount of specialized context vectors are enriched thanks to general-domain corpus. Hyp2 corresponds to the mean of the number of new words that have been added to each context vector of the specialized domain words. For instance, Hyp2 for BC+NC shows that in average we add 746 new English words and 721 new French words for each context vector of the BC corpus. Here also we can see that many new words are added to the specialized context vectors. The experimental results previously shown in Table 2 confirm the usefulness of Hyp1 and Hyp2.

7 Conclusion

We have shown in this article how the problem of adding external data could be achieved for improving bilingual lexicon extraction from specialized comparable corpora. We have proposed two approaches that use external data in an adapted way to preserve the original vocabulary. Even if our selective standard approach goes against the mainstream which states that adding out-of-domain data decreases the quality of bilingual lexicons, we never denature the initial specialized comparable corpus. The results obtained

by the selective standard approach show significant improvements for alignment of single-word terms while using any of the external data and confirm the usefulness of exploiting as much data as we have to better characterize context vector representation and thus bilingual lexicon extraction.

Acknowledgments

The research leading to these results has received funding from the French National Research Agency under grant ANR-12-CORD-0020 (CRISTAL project) and the French Projet OCEAN (Outil de Concordance bilinguE libre pour l'Aide la traductioN) of the General Delegation for the French Language and in languages of France.

References

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

Dhouha Bouamor, Adrian Popescu, Nasredine Semmar, and Pierre Zweigenbaum. 2013a. Building Specialized Bilingual Lexicons Using Large Scale Background Knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 479–489, Seattle, WA, USA.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013b. 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)*, 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 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 24th International Conference on Computational Linguistics (COLING'12)*, pages 745–762, Mumbai, India.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

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.

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.

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.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

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'12)*, pages 126–133, Istanbul, Turkey.

Amir Hazem and Emmanuel Morin. 2013. Word co-occurrence counts prediction for bilingual terminology extraction from comparable corpora. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP'13)*, pages 1392–1400, Nagoya, Japan.

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.

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.

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. 2016. Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction. *Natural Language Engineering*, 22(4):575–601.

Emmanuel Morin, Batrice Daille, Koichi Takeuchi, and Kyo Kageura. 2010. Brains, not brawn: The use of "smart" comparable corpora in bilingual terminology mining. *ACM Transactions on Speech and Language Processing*, 7(1):1–23.

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.

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.

Alexandre Rafalovitch and Robert Dale. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the 12th Conference on Machine Translation Summit (MT Summit XII)*, Ottawa, Canada.

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.

Gerard Salton and Michael E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.

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 Vulic and Marie-Francine Moens. 2013a. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*, pages 106–116, Atlanta, GA, USA.

Ivan Vulic and Marie-Francine Moens. 2013b. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1613–1624, Seattle, WA, USA.

Ivan Vulic 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 Vulic and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *J. Artif. Intell. Res. (JAIR)*, 55:953–994.

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.