Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora

Amir HAZEM, Emmanuel MORIN

Laboratoire d'Informatique de Nantes-Atlantique (LINA)
Université de Nantes, 44322 Nantes Cedex 3, France
Amir.Hazem@univ-nantes.fr, Emmanuel.Morin@univ-nantes.fr

Abstract

One of the main resources used for the task of bilingual lexicon extraction from comparable corpora is : the bilingual dictionary, which is considered as a bridge between two languages. However, no particular attention has been given to this lexicon, except its coverage, and the fact that it can be issued from the general language, the specialised one, or a mix of both. In this paper, we want to highlight the idea that a better consideration of the bilingual dictionary by studying its entries and filtering the non-useful ones, leads to a better lexicon extraction and thus, reach a higher precision. The experiments are conducted on a medical domain corpora. The French-English specialised corpus 'breast cancer' of 1 million words. We show that the empirical results obtained with our filtering process improve the standard approach traditionally dedicated to this task and are promising for future work.

Keywords: Comparable corpora, Bilingual lexicon extraction, Words filtering

1. Introduction

Bilingual lexicon extraction from comparable corpora has become a source of great interest since the 1990s (Rapp, 1995; Fung, 1998; Fung and Lo, 1998; Peters and Picchi, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Gaussier et al., 2004; Morin et al., 2007, among others), mainly because of the scarcity of parallel corpora, especially for language pairs not involving English. The main work in this domain relies on the simple observation that a word and its translation tend to appear in the same context. The basis of this observation consists in the identification of first-order affinities for each source and target language: '*First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word*' (Grefenstette, 1994a, p. 279). These affinities can be represented by context vectors, and each vector's element represents a word which occurs within the window of the word to be translated. The translation candidates for a word are obtained by comparing the translated source context vector with the target context vectors through a bilingual dictionary.

The main shortcoming of this standard approach is that its performance greatly relies on the coverage of the bilingual dictionary. When the context vectors are well translated, the translation retrieval rate in the target language improves. Although, the coverage of the bilingual dictionary can be extended by using specialised dictionaries or multilingual thesauri (Chiao and Zweigenbaum, 2003; Déjean et al., 2002), translation of context vectors remains the core of the approach. Following this observation, we want to give a particular attention to bilingual dictionary, as it seems to be crucial for the performance of the standard approach.

In order to be less dependent on the coverage of the bilingual dictionary, Déjean and Gaussier (2002) have proposed an extension to the standard approach. The basic intuition of this approach is that words sharing the same meaning will share the same environments. The approach is based on the identification of second-order affinities in the source language: '*Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar*' (Grefenstette, 1994a, p. 280).

The concept of text filtering or filtering in general is used in many domains. Information filtering systems are typically designed to sort through large volumes of information and present the user with sources of information that are likely to satisfy the information requirement. In information retrieval for instance, a list of stop words is used as a pre-processing step of a given task. By Following the same principle, we want to introduce a filtering pre-processing step to the standard approach for the task of terminology extraction from a specialised comparable corpora.

The remainder of this paper is organised as follows. Section 2. presents the standard approach dedicated to word alignment from comparable corpora. Section 3. describes our lexicon filtering method to adapt the bilingual dictionary to the comparable corpora. Section 4. describes the different linguistic resources used in our experiments and evaluates the contribution of our method on the quality of bilingual terminology extraction through different experiments. Section 5. presents our discussion and finally, Section 6. presents our conclusions.

2. Standard Approach

The implementation of the standard approach can be carried out by applying the following four steps (Rapp, 1995; Fung and McKeown, 1997):

Context characterization
All the lexical units in the context of each lexical unit i are collected, and their frequency in a window of n words around i extracted. For each lexical unit i of the source and the target languages, we obtain a context vector $\mathbf{i}$ where each entry, $\mathbf{i}_j$, of the vector is given by a function of the co-occurrences of units j and i. Usually, association measures such as the mutual information (Fano, 1961) or the log-likelihood (Dunning, 1993) are used to define vector entries.

Vector transfer
The lexical units of the context vector $\mathbf{i}$ are translated using a bilingual dictionary. Whenever the bilingual dictionary provides several translations for a lexical unit, all the entries are considered but weighted according to their frequency in the target language. Lexical units with no entry in the dictionary are discarded.

Target language vector matching
A similarity measure, $\mathrm{sim}(\bar{\mathbf{i}}, \mathbf{t})$, is used to score each lexical unit, t, in the target language with respect to the translated context vector, $\bar{\mathbf{i}}$. Usual measures of vector similarity include the cosine similarity (Salton and Lesk, 1968) or the weighted jaccard index (WJ) (Grefenstette, 1994b) for instance.

Candidate translation
The candidate translations of a lexical unit are the target lexical units ranked following the similarity score.

3. Adaptive Dictionary

In this section we present our contribution by introducing the filtering process for adapting the dictionary to the bilingual comparable corpora.

The core of the standard approach is the bilingual dictionary, it allows the translation of the context vector of a candidate word and compare it to all the target context vectors to identify the correct translation according to a similarity measure. We present in this section a new method of filtering the entries of the bilingual dictionary. The purpose of this study is to try to answer the following questions : Are all the lexicon entries useful for the characterisation of the words to be translated? Should the size of the dictionary be the main criteria to determine the lexicon quality? To answer these questions, we would like to go furthermore in the study of the entries of the bilingual dictionary. To do so, we present two techniques of filtering the bilingual dictionary.

The first one is merely based on the POS-tagging criteria. Indeed, in the most cases and especially for specialised corpora, words to be translated are often Nouns. So, we would like to know if all the grammatical categories (nouns, verbs and adjectives) are necessary and should be part of context vectors or on the contrary, some of them should be discarded.

The second technique is inspired from the method proposed by (Ahmad et al., 1992). According to (Ahmad et al., 1992) the domain relevance of a term candidate is simply computed as the quotient of its relative frequencies in both the domain specific corpus and the corpus used for comparison. Starting from this idea and in order to estimate the specificity of the lexicon entries, we applied the same principle and adapt it to our bilingual corpora as follow : The relative frequency ($Freq_{rel}(i)$) of a word i in a source corpus is computed by dividing its absolute frequency by the total number of tokens in the source corpus . Analogously, the relative frequency of $T(i)$ which is the translation of i in the target language, is computed by dividing its absolute frequency by the size of the target corpus. Assuming that non relevant lexicon entries are those of a high relative frequency ratio, we filter the lexicon entries according to this criteria. Thus if $Freq_{rel}(i)/Freq_{rel}(T(i)) > \alpha$, the lexicon entry ($i$,$T(i)$) is discarded. The threshold α is fixed empirically.

What we mean by the word ”adaptive” is : instead of taking all the entries of the dictionary to translate a context vector of a word, we only use the words of the lexicon that are more likely to give the best representation of the context vector in the target language. If there is a big difference between the relative frequency of a word and its translation we assume that this lexicon entry and its translation are not relevant due to their high relative frequency ratio and don't follow the same behaviour in both languages.

4. Experiments and Results

4.1. Linguistic Resources

Comparable Corpora
We have selected the documents from the Elsevier website[1] in order to obtain a French-English specialised comparable corpus. The documents were taken from the medical domain within the sub-domain of ‘breast cancer’. We have automatically selected the documents published between 2001 and 2008 where the title or the keywords contain the term ‘cancer du sein’ in French and ‘breast cancer’ in English. We thus collected 130 documents in French and 118 in English and about 530,000 words for each language. The documents comprising the French/English specialised comparable corpus have been normalised through the following linguistic pre-processing steps: tokenisation, part-of-speech tagging, and lemmatisation. Next, the function words were removed and the words occurring less than twice (i.e. hapax) in the French and the English parts were discarded. Finally, the comparable corpus comprised about 7,200 distinct words in French and 8,400 in English.

Dictionary
We used two bilingual lexicons for our experiments. The French-English bilingual dictionary composed of dictionaries that are freely available on the Web (Dicoweb). It contains, after linguistic pre-processing steps 51,600 English single words belonging to the general language. The French-English bilingual dictionary ELRA-M0033[2] dictio-

[1] www.elsevier.com

[2] ELRA dictionary has been done by Sciper in the Technolangue/Euradic project

nary. It contains, after linguistic pre-processing steps, 200 000 English single words belonging to the general language.

	Source(EN)	Target(FR)
Corpus	8492	7230
Dicoweb	2538	2642
ELRA-M0033	3562	3684

Table 1: Number of distinct words of the corpus and the two dictionaries (Dicoweb and ELRA after projection into the corpus).

Evaluation List

In bilingual terminology extraction from specialised comparable corpora, the terminology reference list required to evaluate the performance of the alignment programs are often composed of 100 single-word terms (SWTs) (180 SWTs in (Déjean and Gaussier, 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002), and 100 SWTs in (Daille and Morin, 2005)). To build our reference list, we selected 400 French/English SWTs from the UMLS[3] meta-thesaurus and the *Grand dictionnaire terminologique*[4]. We kept only the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 122 French/English SWTs were extracted.

4.2. Experimental Setup

Three major parameters need to be set for the standard approach, namely the similarity measure, the association measure defining the entry vectors and the size of the window used to build the context vectors. Laroche and Langlais (2010) carried out a complete study about the influence of these parameters on the quality of bilingual alignment. The entries of the context vectors were determined by the log-likelihood (Dunning, 1993), and we used a seven-word window since it approximates syntactic dependencies. As similarity measure, we chose to use the weighted jaccard index (Grefenstette, 1994b):

$$\mathrm{sim}(\mathbf{i},\mathbf{j}) = \frac{\sum_t \min(\mathbf{i}_t, \mathbf{j}_t)}{\sum_t \max(\mathbf{i}_t, \mathbf{j}_t)} \qquad (1)$$

Other combinations of parameters were assessed but the previous parameters turned out to give the best performance. The choice of these parameters is motivated in (Morin et al., 2007).

We note that "Top k" means that the correct translation of a given word is present in the k first candidates of the list returned by the standard approach. We use also the Mean average precision *MAP* (Manning and Schuze, 2008) which represents the quality of the system.

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{Rank_i} \qquad (2)$$

$|Q|$ represents the number of terms to be translated

4.3. Results

To evaluate the performance of the standard approach and the impact of the different filtering techniques, we conduct several experiments on two different bilingual dictionaries. For each dictionary we analyse the impact of the grammatical categories and the relative frequency ratio filtering.

Dicoweb Dictionary

	TOP 1	TOP 5	TOP 10	TOP 15	TOP 20	TOP 50	MAP
BASELINE	**25.40**	**45.08**	**54.09**	**59.01**	**63.93**	**71.31**	**35.08**
N	16.39	30.32	39.34	50.81	51.63	66.39	24.76
ADJ	4.09	13.11	18.85	22.95	27.04	36.06	8.87
V	0.81	5.73	9.01	13.11	16.39	26.22	4.18
V.ADJ	6.55	20.49	31.96	34.42	37.70	45.90	14.29
N.V	18.03	32.78	47.54	50.81	51.63	68.03	26.16
N.ADJ	18.03	34.42	45.90	49.18	58.19	67.21	27.09

Table 2: Accuracy(%) at top k and MAP for Dicoweb according to grammatical categories filtering.

We present in table 2 the results of the standard approach using different combinations of grammatical categories filtering. Table 2 shows that for the Dicoweb dictionary the best results are those using all the grammatical categories. The baseline obtains a $MAP = 35.08\%$ while none of the other combinations reaches better results. We can also notice that nouns are more informative than adjectives and verbs. The filtering process according to nouns only, gives a $MAP = 24.76\%$ while verbs filtering gives a $MAP = 4.18\%$ and adjectives filtering gives a $MAP = 8.87\%$. The filtering process using the combination of nouns and verbs (N.V) and nouns and adjectives (N.Adj) gives almost the same MAP ($MAP = 26.16\%$ for N.V and $MAP = 27.09\%$ for N.Adj).

	TOP 1	TOP 5	TOP 10	TOP 15	TOP 20	TOP 50	MAP
BASELINE	25.40	45.08	54.09	59.01	**63.93**	71.31	35.08
$RFR \leq 2$	20.49	37.70	47.54	50.00	53.27	62.29	28.57
$RFR \leq 3$	21.31	40.98	51.63	57.37	59.01	70.49	31.31
$RFR \leq 4$	22.95	42.62	56.55	59.83	62.29	71.31	33.17
$RFR \leq 5$	26.22	47.54	**59.83**	62.29	63.11	72.95	**37.18**
$RFR \leq 6$	23.77	46.72	55.73	59.83	62.29	72.95	34.91
$RFR \leq 7$	23.77	46.72	57.37	**63.11**	**63.93**	72.95	34.76
$RFR \leq 8$	25.40	47.54	57.37	61.47	63.11	**73.77**	35.87
$RFR \leq 9$	**27.86**	47.54	56.55	59.83	63.11	72.13	37.16
$RFR \leq 10$	27.04	**48.36**	56.55	60.65	63.11	72.95	36.55

Table 3: Accuracy(%) at top k and MAP for Dicoweb using relative frequency ratio filtering.

We present in table 3 a comparison between the baseline and different relative frequency ratio (RFR) scores. Table 3 shows that using the relative frequency ration filtering improves the results in most of the cases. For $RFR = 5$ we obtain the maximum MAP score with $MAP = 37.18\%$ while for the baseline we obtain a $MAP = 35.08\%$. Other RFR scores give better MAP than the baseline ($RFR = 8, 9, 10$).

[3] http://www.nlm.nih.gov/research/umls

[4] http://www.granddictionnaire.com/

	Top 1	Top 5	Top 10	Top 15	Top 20	Top 50	MAP
Baseline	**33.60**	**55.73**	**64.75**	**69.67**	**72.13**	**81.14**	**43.64**
N	26.22	43.44	50.81	58.19	65.57	75.40	34.70
Adj	11.47	27.04	29.50	32.78	36.06	53.27	18.30
V	1.63	8.19	13.93	18.03	19.67	28.68	5.26
V.Adj	11.47	33.60	36.88	45.90	48.36	57.37	21.38
N.V	26.22	46.72	55.73	62.29	65.57	77.86	35.67
N.Adj	31.14	53.27	62.29	**69.67**	**72.13**	79.50	41.64

Table 4: Accuracy(%) at top k and MAP for ELRA according to grammatical categories filtering.

ELRA Dictionary

We present in table 4 the results of the standard approach using different combinations of grammatical categories filtering on the ELRA dictionary. Table 4 shows that for the ELRA dictionary the best results are those using all the grammatical categories. The baseline obtains a $MAP = 43.64\%$ while none of the other combinations reaches better results. We can also notice that nouns are more informative than adjectives and verbs. The filtering process according to nouns only, gives a $MAP = 34.70\%$ while verbs filtering gives a $MAP = 5.26\%$ and adjectives filtering gives a $MAP = 18.30\%$. The filtering process using the combination of nouns and verbs (N.V) gives a $MAP = 35.67\%$ while the combination of nouns and adjectives gives better results with a $MAP = 41.64\%$.

	Top 1	Top 5	Top 10	Top 15	Top 20	Top 50	MAP
Baseline	33.60	55.73	64.75	69.67	72.13	**81.14**	43.64
$RFR \leq 2$	28.68	48.36	55.73	60.65	63.11	69.67	37.32
$RFR \leq 3$	29.50	50.81	59.01	65.57	71.31	76.22	39.59
$RFR \leq 4$	32.78	53.27	62.29	66.39	69.67	76.22	42.47
$RFR \leq 5$	37.70	54.91	63.93	69.67	72.95	80.32	46.47
$RFR \leq 6$	36.88	56.55	65.57	69.67	73.77	80.32	45.88
$RFR \leq 7$	38.52	58.19	66.39	69.67	**74.59**	80.32	47.35
$RFR \leq 8$	**39.34**	58.19	**68.03**	**70.49**	**74.59**	80.32	**48.17**
$RFR \leq 9$	35.24	**59.83**	67.21	69.67	72.95	80.32	45.85
$RFR \leq 10$	36.06	59.01	67.21	69.67	73.77	80.32	46.16

Table 5: Accuracy(%) at top k and MAP for ELRA using relative frequency ratio filtering.

We present in table 5 a comparison between the baseline and different relative frequency ratio (RFR) scores. Table 5 shows that using the relative frequency ration filtering improves the results in most of the cases. For $RFR = 8$ we obtain the maximum MAP score with $MAP = 48.17\%$ while for the baseline we obtain a $MAP = 43.64\%$. Other RFR scores give better MAP than the baseline ($RFR = 5, 6, 7, 8, 9, 10$).

5. Discussion

The main interest of our work is to show that by a simple filtering process based on relative frequency ratio, we can improve the accuracy of the standard approach. In general, systems consist on several modules when put together they build the entire system. In this paper we propose to add the filtering module to the alignment system. Our filtering is simple but seems to be necessary to improve the accuracy regarding the results obtained in our experiments. Our idea is based on the simple assumption that a word and its translation should have a small relative frequency ratio, if not, this could bring more noise then useful information for bilingual alignment. We also experienced the filtering process according to grammatical categories. The results have shown that nouns are more informative than verbs and adjectives. On the other hand, the filtering process according to grammatical categories did not improve the baseline's results. It is certainly needed to consider the word not only according to its grammatical category but also to some other fine grade word characteristics such as the type of adjectives (for instance comparative, superlative, etc.) or the type of nouns (for instance, common noun or proper noun, etc.). More experiments are certainly needed.

If we take a look at ELRA dictionary, we can notice that for a given word we can find many possible translations, for different context. ELRA is a general lexicon, may be too general! So, is it appropriate to use a general lexicon for aligning domain specific words? Or, should we use domain specific dictionaries? For the second question it is not always easy to find domain specific lexicons, and using general lexicon leads us to consider translations which are not always in context and thus not relevant. For these reasons, we introduce the idea of adapting the dictionary to the corpus. From a given dictionary and according to the bilingual corpus, we extract only words that are more likely to be relevant to the alignment process. One of the drawbacks of the relative frequency ratio (RFR) is maybe the parameter of RFR, in our experiments we fixed it empirically. We hope in future work to find a way to fix it in a way to maximise the system performance.

We focus in this paper on a medical domain corpora (Breast cancer) as it is our main center of interest, it would be interesting however to see the results of our filtering process in a general domain corpus.

6. Conclusion

We have presented a new lexicon filtering technique for the problem of bilingual lexical extraction from comparable corpora based on relative frequency ratio (RFR) criteria. Pos-tagging filtering have shown no improvements in the performance of the system. The use of relative frequency ratio on the contrary, have shown a better performance than using all the entries of the dictionary. We believe that our model is simple and sound. Regarding the empirical results of our proposition, performance of our filtering technique on our dataset was better than the baseline. further research are certainly needed to confirm these first results but our findings lend support for the hypothesis that an adaptive bilingual lexicon is an appropriate way to improve the accuracy for the task of bilingual lexicon extraction from comparable corpora.

7. Acknowledgement

The research leading to these results has received funding from the French National Research Agency under grant ANR-08-CORD-013.

8. References

K. Ahmad, A. Davies, H. Fulford, and M. Rogers. 1992. What is a term? the semiautomaticextraction of terms from text. *Paper presented at a conference held at the University of Vienna, Institut fr bersetzer- und Dolmetscherausbildung, TranslationStudies - An Interdiscipline*.

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.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. IOS Press.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

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

Pascale Fung and Yuen Yee Lo. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, pages 414–420.

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 ParallelCorpora to Non-parallel Corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Éric 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.

Gregory Grefenstette. 1994a. Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, pages 279–290, Amsterdam, The Netherlands.

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

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.

Raghavan; Manning, Christopher D.; Prabhakar and Hinrich Schuze. 2008. Introduction to information retrieval. Cambridge University Press.

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.

Carol Peters and Eugenio Picchi. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Gregory Grefenstette, editor, *Cross-language information retrieval*, chapter 7, pages 81–90. Kluwer Academic Publishers.

Reinhard Rapp. 1995. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, 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.

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.