Exploring Context Variation and Lexicon Coverage in Projection-based Approach for Term Translation

Raphaël Rubino
Laboratoire Informatique d'Avignon
339, chemin des Meinajaries
Agroparc BP 1228
84911 Avignon Cedex 9, France
raphael.rubino@univ-avignon.fr

Abstract

Identifying translations in comparable corpora has inspired many studies in bilingual terminology extraction [4, 5]. Projection-based approaches, which are among the most popular ones, rely on a seed bilingual lexicon. Surprisingly, there is no careful analysis of the impact of the size the initial context and coverage of the lexicon. This is precisely the focus of this study. We observe that source context size and lexicon coverage influence robustness in projection-based term translation. In particular, we show that increasing the number of seed words by a factor of three leads to a 20% relative improvement in accuracy.

1 Introduction

Parallel corpora have been widely used in machine translation. For example, sentence and word aligment models proposed by [2], or applications to multilingual terminology extraction [8]. This approach, using parallel corpora, yields good results. But the lack of parallel texts is still an issue. This is particularly true for specific domains. Building parallel corpora is time-consuming and relies heavily on human translators. Even if domain specific parallel corpora exist, terminology extraction often amounts to reverse engineering the work of human translators.

This is a reason why comparable corpora [11] are studied by researchers in multilingual terminology extraction. Some authors have shown that statistical methods can make use of comparable corpora. In particular, [17] paved the way for a family of approaches that assumes that co-occurrences of words which are translations of each other are correlated in comparable corpora. Basically, we can observe that a word and its translation appear in the same lexical environment, which can be used as a context vector [5]. This projection-based translation approach can be applied to terminology extraction. For example in [14], a term extraction program coupled with a lexical alignment program are implemented to manage this task.

Usually, in projection-based term extraction, a bilingual or multilingual lexicon is needed to do the projection from one language to another. This step depends on the lexicon and the context vector. To the best of our knowledge, there are no analyses on the impact of the initial context size or on the coverage of the lexicon in such projection-based methods. This is precisely the focus of this paper. The remainder of this paper is organized as follows : in section two, we present the projection-based approach and describe related work. In section three, we explain the experimental settings. In section four, we present the resources used. Finally the results are presented in section five, followed by a discussion in section six.

2 Projection-based Approach

2.1 Description

In the source language text, the term to be translated is surrounded by a context consisting of other terms. This information helps us build a context vector, with a flexible window around the term [6, 14] for example. Then this context has to be projected in the target language. The target context vector is built thanks to a bilingual lexicon. This lexicon-based step is the basis of projection-based approaches. To retrieve translation candidates, the projected context vector has to be compared with all possible vectors in the target language built directly from corpora.

This comparison can be computed with different similarity measures. Usually, the Cosine, Jaccard or Dice coefficient [9] are used. [18] obtained better results using the city-block metric than using the Cosine, Jaccard coefficient, Euclidean distance and scalar product. [14] have also made studies on the impact of different metrics to extract terms' translation pairs. The figure 1 illustrates this projection-based approach.

2.2 Related Works

Studies on parallel corpora has allowed to identify features like co-occurrence position of a word and its translation [10, 19, 21]. Switching to non-parallel corpora implies that the co-occurrence feature is not directly applicable because there are no direct correspondences between sentences or segments. Another word feature correlating words pairs, called context heterogeneity [5], can be applied to texts in different languages which are not translations of each other. Computed on comparable corpora, the context heterogeneity measure can be used to retrieve domain-

Student Research Workshop, RANLP 2009 - Borovets, Bulgaria, pages 66–70

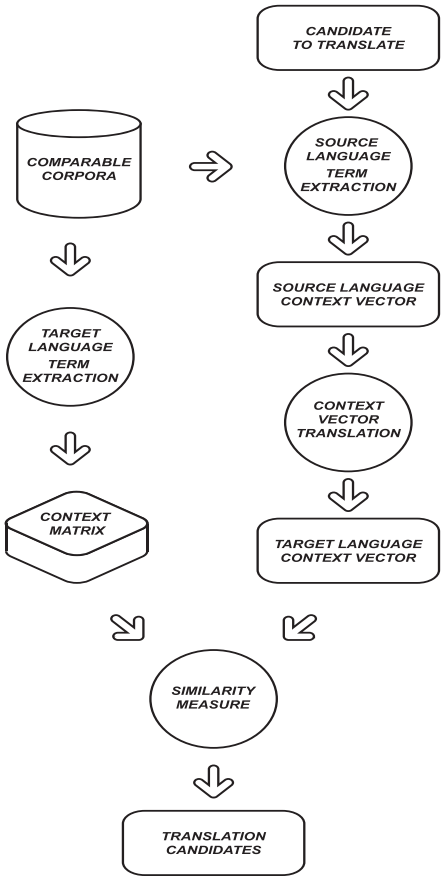

Fig. 1: *General representation of the projection-based approach with comparable corpora.*

specific word pairs in unrelated languages (English and Japanese for example).

[17] concludes that patterns of co-occurrences of words between different languages are correlated in non-parallel corpora. In [18], co-occurrence matrices are built from comparable corpora, and used to compare the projected vectors with all possible vectors from the initial text. In the same study, a small seed word lexicon, which does not cover the test set, is used and expanded during the experiments with the projection-based approach.

Based on these studies, [14] proposed an approach to solve multi-word term translation from non-parallel corpora. They first adapt the single-word term context vector approach proposed by [7] to multi-word term. Then, an implementation of the direct context vector method is proposed and applied to terminology extraction between unrelated languages (French/Japanese). Different metrics are compared to compute similarity between the context vector of the term to translate and the context-matrix built from the initial corpora [14].

In these studies, much attention is paid to similarity metrics between context vectors, built on single or multi-word term co-occurrence values. The projection-based approach described here requires a bilingual seed word lexicon, and it is very surprising that its coverage is not yet carefully studied. Only terms in the lexicon are projected in the target context vector, so this aspect is very important for initial context-vector projection.

3 Experimental Settings

Our system takes as input a list of domain-specific single and multi-word terms to translate. Each of them is a query for document retrieval in a comparable corpus. Terms from these documents compose an initial context vector. Then we use a seed word lexicon to project the context from source to target language. The resulting target language vector is used to retrieve documents from a comparable corpus. The aim of this study is to manage and score document retrievals containing initial term translations, according to initial translation references. The impact of the initial context's size variations, its seed word lexicon coverage and the lexicon size are studied. Experiments were conducted on French to English single and multi-word term translations. The implementation is composed of five parts :

1. Document retrieval (the query is the term to translate)
2. Initial context vector construction from words contained in documents
3. Projection of the vector in the target language using the seed word lexicon
4. Document retrieval with projected vector
5. Oracle scoring on documents : containing or not the term translation

The score is computed on retrieved documents. If a returned document contains the candidate translation, an oracle is set to 1, otherwise it remains at 0. It is possible that the initial context vector can not be built, because the corpus does not contain the term to translate. It is also possible that the projected vector cannot be used to retrieve documents in the target language. The first reason is that the translation reference is not in the corpora. We decide then to compute two oracle scores : on all candidates from the initial term list and on candidates with an initial context covered by the corpora.

The task is to verify whether the target language context is robust enough to retrieve documents containing the initial term translation. It is the first thing we want to measure. We make variations of the initial context and the seed word lexicon size, but also of the seed word lexicon coverage.

4 Resources

4.1 Comparable Corpora

Using the World Wide Web as a non-parallel corpus can solve the problems of accessibility, relevance and quantity of data. Wikipedia is a well known online free collaborative encyclopedia. Many articles are domain specific, and each document represents one concept

only [13]. As in [16], we use Wikipedia[1] as a comparable corpus, for the abundance of the multilingual content freely available. Wikipedia is used in many natural language processing domains, like named entity disambiguation [3], the retrieval of similar sentences across different languages [1], thesaurus extraction [13], semantic relation extraction [20], etc.

Although Wikipedia has a structure that can help identify translations (cross languages links, titles of pages and section, ...), we do not consider this information this study. We want to build the context vectors for terms to translate with Wikipedia articles, which are used like concept-related word lists or semantic networks.

In order to extract the information we need from Wikipedia, we rely in this work on a tool called NLGbAse[2]. This tool provides a search engine with cosine similarity computed between the query and the returned documents. For each document, NLGbAse gives the list of contained words ranked by their *tf.idf* measure computed on the whole Wikipedia corpora.

4.2 Candidates

In this series of experiments, a term list is taken from the MeSH thesaurus [15]. 10 000 single and multiwords terms in French, along with their translations in English, are extracted [12]. None of these terms are covered by the lexicon used for the experiments. Prior to the experiments, two filtering steps are done. The first is made to be sure that Wikipedia can be used to build source language contexts. This means that the term to translate is in the corpus. Then, a second filtering step is done on Wikipedia with the translation proposed in MeSH for every term. After these two steps, 2 000 terms were removed, because none of them are covered by Wikipedia.

4.3 Bilingual Seed Word Lexicon

To manage a pivot between different languages, a domain specific lexicon has to be built. We use the data available from Robert H. Vander Stichele's website[3]. We automatically retrieve a French-English lemma collection of technical and popular medical single words. To handle general terms, we choose to extend the 1800 medical words collection with a general lexicon containing 3200 words. Our lexicon finally contains 5000 word pairs.

For the experiments, three seed word lexicons are used. The first is the full one, with general and domain specific words. The second is only general, and the third is only medical.

We choose to automatically build lexicons from a web resource to be as independent as possible from the candidates to translate extracted from the MeSH thesaurus. The lexicon coverage of the candidate list is null, in order to avoid any bias. It means that only words in context vectors are translated.

[1] http://www.wikipedia.org
[2] http://nlgbase.org/
[3] http://users.ugent.be/~rvdstich/eugloss/welcome.html

4.4 Stop-words List

Words used to build initial context-vectors are more or less significant and can even introduce noise. For example, words in the source language which are not directly related to the term to translate can be "ainsi" ("so"), "quand" ("when") or "toujours" ("always"), etc. We decided to filter these words before building context-vectors with a stop-words list, which contains the 1300 common non-content words of the source language.

5 Results

Table 1 presents the results of the oracle described in the experimental settings. We give details about the number of terms to translate and about the number of seed words (corresponding to the size of the projected context). We also study how the number of documents and the number of terms per document vary. We use the full seed word lexicon (general and domain specific).

The maximum number of candidates handled during the experiments does not reach the maximum number of candidates in the initial list. This can be explained by the lack of vocabulary in the seed word lexicon. A null initial context-vector cannot be used to do a projection-based approach, so the candidate to translate is not handled.

docs	terms	cand.(%)	seeds	oracle	limited
1	10	63.25	2.06	0.29	0.58
1	50	89.54	6.37	0.45	0.64
1	100	94.52	11.07	0.49	0.67
1	200	94.86	19.95	0.53	0.72
1	999	94.90	38.69	0.56	0.76
10	1	49.51	1.74	0.21	0.54
10	2	65.62	2.50	0.29	0.57
10	5	82.30	4.58	0.40	0.63
10	10	89.24	7.99	0.48	0.69
10	20	92.24	14.78	0.53	0.75
10	50	94.00	32.45	0.58	0.80
10	100	95.00	58.16	0.61	0.83
10	200	95.00	103.77	0.63	0.85
20	1	57.91	2.28	0.25	0.56
20	5	84.41	6.81	0.44	0.67
20	10	89.92	12.09	0.51	0.72
20	20	92.38	22.29	0.55	0.77
20	40	93.82	40.10	0.59	0.80
20	50	94.00	48.50	0.60	0.81
20	100	95.00	86.09	0.62	0.84
50	1	64.82	3.31	0.30	0.59
50	2	75.52	5.45	0.37	0.63
50	5	85.20	11.06	0.47	0.70
50	10	90.11	19.66	0.52	0.75
50	20	92.44	35.75	0.57	0.79
100	1	67.38	4.35	0.32	0.61
100	10	90.18	26.93	0.53	0.75

Table 1: *Oracle with the full seed word lexicon. Initial documents and terms are used to build the initial context. The last column is an oracle computed on covered candidates only (third column). The "oracle" column is computed on all candidates (8000 term pairs).*

In order to reach an accuracy of 50%, 20 initial documents and 10 words per document are needed. The source context-vector is then sufficient to do the projection in the target language and to retrieve good documents. In theory, this means that an initial context-vector of 200 terms is required. Besides, the average projected context-vector size is about 12.09 seeds. The seed word lexicon coverage of the initial context is, in theory, about 6.04%. In fact, the size of the initial context-vector depends on the number of terms contained in the documents from Wikipedia. The average lexicon coverage of initial contexts in all experiments is around 11% (from 9% to 13% depending on the size of the initial context-vectors).

On the experiments with 20 initial documents, we can see that increasing the number of seeds by a factor of three leads to a relative improvement in oracle accuracy of 20%. If we look at the results on 1 document and 50 terms per document, compared to the results on 20 documents with 5 terms, for a lower number of seeds (6.37 instead of 6.81) and half of the initial context-vector size (50 words instead of 100), the oracle accuracy of document retrieval is enhanced. It can be explained by a higher seed word lexicon coverage, associated to a *better* initial context-vector. In fact, this context-vector is closer to the candidate to translate. All the words used to build the context are in the first document retrieved by the search engine, that is the document with the best cosine similarity measure.

The tables 2 and 3 contain the oracle and *limited* scores, respectively with the general and the medical seed word lexicons. We can see that a higher recall is obtained with the domain-specific lexicon. The reason is that this seed word lexicon gives a higher initial context coverage. Using the general seed word lexicon, we see that more term-to-translate candidates are handled. This means that it is easier to project contexts, but the recall (documents with good translation retrieval) is lower than with domain specific lexicon.

docs	terms	cand.(%)	seeds	oracle	limited
1	10	28.61	1.36	0.09	0.42
1	50	76.68	2.64	0.21	0.36
1	100	91.67	4.71	0.26	0.37
10	1	19.80	1.21	0.07	0.44
10	10	76.90	3.28	0.26	0.43
10	20	85.94	5.65	0.32	0.48
10	50	91.88	12.69	0.39	0.54
10	100	94.71	25.33	0.43	0.58
10	200	95.00	50.03	0.46	0.62
20	5	69.03	3.02	0.24	0.44
20	20	86.39	8.94	0.35	0.52
20	40	90.45	16.41	0.40	0.57
50	1	41.19	1.76	0.14	0.43
50	10	79.78	8.97	0.33	0.54
50	50	91.96	34.51	0.43	0.60
50	100	94.71	61.71	0.46	0.62

Table 2: *Oracle with the general seed word lexicon.*

On figure 2, the oracle score by seed word is reported. We can see that for an identical number of seeds, the number of documents used to build the initial context-vector has an important impact. Taking less documents but more terms by document increases the oracle score, with the same seed word lexicon. On figure 3, the impact of the type of seed word lexicon is presented. We can see that using only a domain specific seed lexicon, instead of using only a general lexicon, leads to an improvement of the oracle accuracy.

docs	terms	cand.(%)	seeds	oracle	limited
1	10	45.18	1.99	0.19	0.55
1	50	79.52	5.20	0.36	0.58
1	100	86.99	8.22	0.41	0.61
10	1	35.67	1.65	0.15	0.53
10	10	80.36	6.42	0.39	0.63
10	20	88.07	11.02	0.45	0.66
10	50	92.73	22.61	0.52	0.72
10	100	93.53	38.22	0.56	0.76
10	200	94.73	62.21	0.59	0.79
20	5	74.38	5.52	0.35	0.61
20	20	89.73	15.79	0.48	0.69
20	40	92.59	27.23	0.53	0.73
50	1	50.59	2.92	0.22	0.55
50	10	86.02	13.55	0.45	0.67
50	50	93.01	48.2	0.56	0.77
50	100	93.56	77.86	0.59	0.80

Table 3: *Oracle with the domain specific seed word lexicon.*

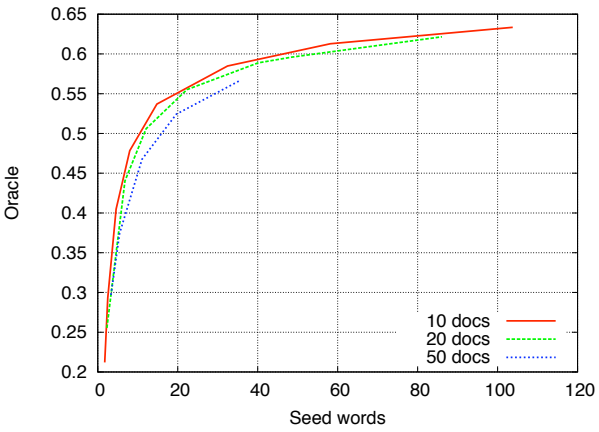

Fig. 2: *Oracle score with seed word variations.*

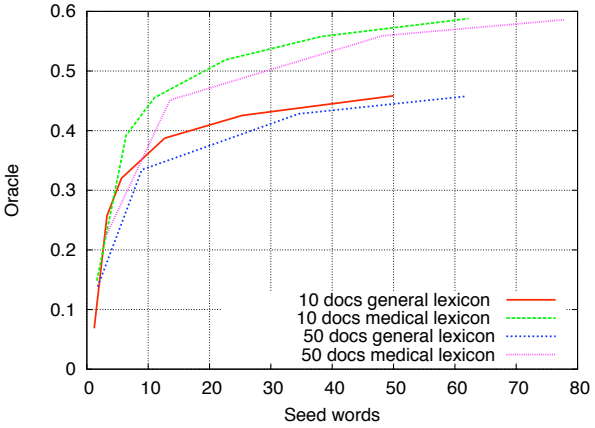

Fig. 3: *Oracle score by lexicon type.*

6 Discussion

In this paper, we describe the impact of the initial context-vector size and its lexicon coverage in projection-based methods for term translation. We also give details about the robustness of this context with variations of the number of documents and term-by-document during its construction. The oracle accuracy can be improved with less documents initially used and a higher seed word lexicon coverage. With an equivalent number of seeds, the smaller the initial context-vector, the higher the oracle accuracy.

We show that for domain specific term translation, a recall score of 85% can be obtained with a domain specific seed word lexicon, completed with a general lexicon. The domain specific lexicon has a better coverage of the initial context while the general lexicon handles more term-to-translate candidates.

This study will help us to continue our works on projection-based domain specific term translations. In particular, using other comparable corpora, like the World Wide Web [7], which can be considered as an higher unrelated non-parallel corpora. We assume that generation of N-Best translation candidates lists, which is a step further in this study, can be improved with robust initial context-vectors.

References

[1] S. Adafre and M. de Rijke. Finding Similar Sentences Across Multiple Languages in Wikipedia. In *Proceedings of the 11th EACL conference*, pages 62–69, 2006.

[2] P. Brown, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, 1990.

[3] S. Cucerzan. Large-scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of EMNLP-CoNLL conference*, pages 708–716, 2007.

[4] B. Daille, É. Gaussier, and J. Langé. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the 15th COLING conference*, volume 1, pages 515–521. ACL, 1994.

[5] P. Fung. Compiling Bilingual Lexicon Entries from a Non-parallel English-Chinese Corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 173–183, 1995.

[6] P. Fung. A Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora. *Lecture Notes in Computer Science*, 1529:1–17, 1998.

[7] P. Fung and L. Yee. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 36th ACL conference*, pages 414–420. ACL, 1998.

[8] É. Gaussier. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In *Proceedings of the 36th ACL conference*, pages 444–450. ACL, 1998.

[9] W. Jones and G. Furnas. Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American society for information science*, 38(6), 1987.

[10] J. Kupiec. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proceedings of the 31st ACL conference*, pages 17–22. ACL, 1993.

[11] J. Laffling. On Constructing a Transfer Dictionary for Man and Machine. *Target*, 4(1):17–31, 1992.

[12] P. Langlais, F. Yvon, and P. Zweigenbaum. Translating medical words by analogy. In *Intelligent Data Analysis in Biomedicine and Pharmacology*, pages 51–56, Washington, DC, USA, 2008.

[13] D. Milne, O. Medelyan, and I. Witten. Mining Domain-specific Thesauri from Wikipedia: A Case Study. In *IEEE/WIC/ACM International Conference on Web Intelligence, 2006. WI 2006*, pages 442–448, 2006.

[14] E. Morin, B. Daille, K. Takeuchi, and K. Kageura. Bilingual Terminology Mining-Using Brain, not Brawn Comparable Corpora. In *Proceedings of the 45th ACL conference*, page 664. ACL, 2007.

[15] S. Nelson and J. Schulman. A Multilingual Vocabulary Project-Managing the Maintenance Environment. *MeSH Section, National Library of Medicine, Bethesda, Maryland*, 2007.

[16] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-based Multilingual Retrieval Model. *Lecture Notes in Computer Science*, 4956:522, 2008.

[17] R. Rapp. Identifying Word Translations in Non-parallel Texts. In *Proceedings of the 33rd ACL conference*, pages 320–322. ACL, 1995.

[18] R. Rapp. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th ACL conference*, pages 519–526. ACL, 1999.

[19] F. Smadja and K. McKeown. Translating Collocations for Use in Bilingual Lexicons. In *Proceedings of the ARPA HLT*, volume 94, 1994.

[20] M. Strube and S. Ponzetto. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proceedings of the AAAI conference*, page 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[21] D. Wu and X. Xia. Learning an English-Chinese lexicon from a Parallel Corpus. In *Proceedings of the 1st AMTA*, pages 206–213. AMTA, 1994.