Rare Word Translation Extraction from Aligned Comparable Documents

Emmanuel Prochasson and Pascale Fung
Human Language Technology Center
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{eemmanuel,pascale}@ust.hk

Abstract

We present a first known result of high precision rare word bilingual extraction from comparable corpora, using aligned comparable documents and supervised classification. We incorporate two features, a context-vector similarity and a co-occurrence model between words in aligned documents in a machine learning approach. We test our hypothesis on different pairs of languages and corpora. We obtain very high F-Measure between 80% and 98% for recognizing and extracting correct translations for rare terms (from 1 to 5 occurrences). Moreover, we show that our system can be trained on a pair of languages and test on a different pair of languages, obtaining a F-Measure of 77% for the classification of Chinese-English translations using a training corpus of Spanish-French. Our method is therefore even applicable to low languages without training data.

1 Introduction

Rare words have long been a challenge to translate automatically using statistical methods due to their low occurrences. However, the *Zipf's Law* claims that, for any corpus of natural language text, the frequency of a word w_n (n being its rank in the frequency table) will be roughly twice as high as the frequency of word w_{n+1}. The logical consequence is that in any corpus, there are very few frequent words and many rare words.

We propose a novel approach to extract rare word translations from comparable corpora, relying on two main features.

The first feature is the *context-vector similarity* (Fung, 2000; Chiao and Zweigenbaum, 2002; Laroche and Langlais, 2010): each word is characterized by its context in both source and target corpora, words in translation should have similar context in both languages.

The second feature follows the assumption that specific terms and their translations should appear together often in documents on the same topic, and rarely in non-related documents. This is the general assumption behind early work on bilingual lexicon extraction from parallel documents using sentence boundary as the context window size for co-occurrence computation, we suggest to extend it to aligned comparable documents using document as the context window. This document context is too large for co-occurrence computation of functional words or high frequency content words, but we show through observations and experiments that this window size is appropriate for rare words.

Both these features are unreliable when the number of occurrences of words are low. We suggest however that they are complementary and can be used together in a machine learning approach. Moreover, we suggest that the model trained for one pair of languages can be successfully applied to extract translations from another pair of languages.

This paper is organized as follows. In the next section, we discuss the challenge of rare lexicon extraction, explaining the reasons why classic approaches on comparable corpora fail at dealing with rare words. We then discuss in section 3 the concept of *aligned comparable documents* and how we exploited those documents for bilingual lexicon extraction in section 4. We present our resources and implementation in section 5 then carry out and comment several experiments in section 6.

2 The challenge of rare lexicon extraction

There are few previous works focusing on the extraction of rare word translations, especially from comparable corpora. One of the earliest works is from (Pekar et al., 2006). They emphasized the fact that the context-vector based approach, used for processing comparable corpora, *perform quite unreliably on all but the most frequent words*. In a nutshell[1], this approach proceeds by gathering the context of words in source and target languages inside *context-vectors*, then compares source and target context-vectors using similarity measures. In a monolingual context, such an approach is used to automatically get synonymy relationship between words to build thesaurus (Grefenstette, 1994). In the multilingual case, it is used to extract translations, that is, pairs of words with the same meaning in source and target corpora. It relies on the *Firthien* hypothesis that *you shall know a word by the company it keeps* (Firth, 1957).

To show that the frequency of a word influences its alignment, (Pekar et al., 2006) used six pairs of comparable corpora, ranking translations according to their frequencies. The less frequent words are ranked around 100-160 by their algorithm, while the most frequent ones typically appear at rank 20-40.

We ran a similar experiment using a French-English comparable corpus containing medical documents, all related to the topic of *breast cancer*, all manually classified as *scientific discourse*. The French part contains about 530,000 words while the English part contains about 7.4 millions words. For this experiment though, we sampled the English part to obtain a 530,000-words large corpus, matching the size of the French part.

Using an implementation of the context-vector similarity, we show in figure 1 that frequent words (above 400 occurrences in the corpus) reach a 60% precision whereas rare words (below 15 occurrences) are correctly aligned in only 5% of the time.

These results can be explained by the fact that, for the vector comparison to be efficient, the information they store has to be relevant and discriminatory. If there are not enough occurrences of a word, it is impossible to get a precise description of the *typical* context of this word, and therefore its description is likely to be very different for source and target words in translation.

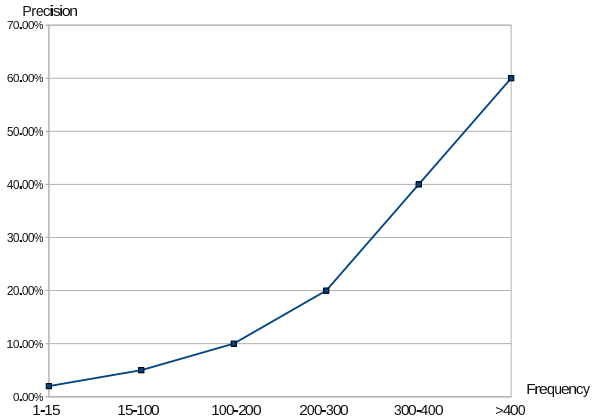

Figure 1: Results for context-vector based translations extraction with respect to word frequency. The vertical axis is the amount of correct translations found for Top_1, and the horizontal axis is the word occurrences in the corpus.

We confirmed this result with another observation on the full English part of the previous corpus, randomly split in 14 samples of the same size. The context-vectors for very frequent words, such as *cancer* (between 3,000 and 4,000 occurrences in each sample) are very similar across the subsets. Less frequent words, such as *abnormality* (between 70 and 16 occurrences in each sample) have very unstable context-vectors, hence a lower similarity across the subsets. This observation actually indicates that it will be difficult to align *abnormality* with itself.

3 Aligned comparable documents

A pair of *aligned comparable documents* is a particular case of comparable corpus: two comparable documents share the same topic and domain; they both relate the same information but are not mutual translations; although they might share parallel chunks (Munteanu and Marcu, 2005) – paragraphs, sentences or phrases – in the general case they were written independently. These comparable documents, when concatenated together in order, form an aligned comparable corpus.

[1]Detailed presentations can be found for example in (Fung, 2000; Chiao and Zweigenbaum, 2002; Laroche and Langlais, 2010).

Examples of such aligned documents can be found, for example in (Munteanu and Marcu, 2005): they aligned comparable documents with close publication dates. (Tao and Zhai, 2005) used an iterative, bootstrapping approach to align comparable documents using examples of already aligned corpora. (Smith et al., 2010) aligned documents from Wikipedia following the interlingual links provided on articles.

We take advantage of this alignment between documents: by looking at *what is common between two aligned documents* and *what is different in other documents*, we obtain more precise information about terms than when using a larger comparable corpus without alignment. This is especially interesting in the case of rare lexicon as the classic context-vector similarity is not discriminatory enough and fails at raising interesting translation for rare words.

4 Rare word translations from aligned comparable documents

4.1 Co-occurrence model

Different approaches have been proposed for bilingual lexicon extraction from parallel corpora, relying on the assumption that a word has one sense, one translation, no missing translation, and that its translation appears in aligned parallel sentences (Fung, 2000). Therefore, translations can be extracted by comparing the distribution of words across the sentences. For example, (Gale and Church, 1991) used a derivative of the χ^2 statistics to evaluate the association between words in aligned region of parallel documents. Such association scores evaluate the strength of the relation between events. In the case of parallel sentences and lexicon extraction, they measure how often two words appear in aligned sentences, and how often one appears without the other. More precisely, they will compare their number of co-occurrences against the expected number of co-occurrences under the null-hypothesis that words are randomly distributed. If they appear together more often than expected, they are considered as associated (Evert, 2008).

We focus in this work on *rare words*, more precisely on specialized terminology. We define them as the set of terms that appear from 1 (hapaxes) to 5 times. We use a strategy similar to the one applied on parallel sentences, but rely on *aligned documents*. Our hypothesis is very similar: words in translation should appear in aligned comparable documents. We used the *Jaccard similarity* (eq. 1) to evaluate the association between words among aligned comparable documents. In the general case, this measure would not give relevant scores due to frequency issue: it produces the same scores for two words that appear always together, and never one without the other, disregarding the fact that they appear 500 times or one time only. Other association scores generally rely on occurrence and co-occurrence counts to tackle this issue (such as the log-likelihood, eq. 2). In our case, the number of co-occurrences will be limited by the number of occurrences of the words, from 1 to 5. Therefore, the Jaccard similarity efficiently reflects what we want to observe.

$$J(w_i, w_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}; A_i = \{d : w_i \in d\} \quad (1)$$

A score of 1 indicates a perfect association (words always appear together, never one without the other), the more one word appears without the other, the lower the score.

4.2 Context-vector similarity

We implemented the context-vector similarity in a way similar to (Morin et al., 2007). In all experiments, we used the same set of parameters, as they yielded the best results on our corpora. We built the context-vectors using nouns only as seed lexicon, with a window size of 20. Source context-vectors are translated in the target language using the resources presented in the next section. We used the log-likelihood (Dunning, 1993, eq. 2) for context-vector normalization (O is the observed number of co-occurrence in the corpus, E is the expected number of co-occurrences under the null hypothesis). We used the Cosine similarity (eq. 3) for context-vector comparisons.

$$ll(w_i, w_j) = 2 \sum_{ij} O_{ij} log \frac{O_{ij}}{E_{ij}} \quad (2)$$

$$Cosine(A, B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B} \quad (3)$$

4.3 Binary classification of rare translations

We suggest to incorporate both the context-vector similarity and the co-occurrence features in a machine learning approach. This approach consists of training a classifier on positive examples of translation pairs, and negative examples of non-translations pairs. The trained model (in our case, a decision tree) is then used to tag an unknown pair of words as either ”Translation” or ”Non-Translation”.

One potential problem for building the training set, as pointed out for example by (Zhao and Ng, 2007) is this: we have a limited number of positive examples, but a very large amount of *non-translation* examples as obviously is the case for rare word translations in any training corpus. Including two many negative examples in the training set would lead the classifier to label every pairs as ”Non-Translation”.

To tackle this problem, (Zhao and Ng, 2007) tuned the imbalance of positive/negative ratio by resampling the positive examples in the training set. We chose to reduce the set of negative examples, and found that a ratio of five negative examples to one positive is optimal in our case. A lower ratio improves precision but reduces recall for the ”Translation” class.

It is also desirable that the classifier focuses on discriminating between confusing pairs of translations. As most of the negative examples have a null co-occurrence score and a null context-vector similarity, they are excluded from the training set. The negative examples are randomly chosen among those that fulfill the following constraints:

- non-null features ;
- ratio of number of occurrences between source/target words higher than 0.2 and lower than 5.

We use the J48 decision tree algorithm, in the *Weka* environment (Hall et al., 2009). Features are computed using the Jaccard similarity (section 3) for the co-occurrence model, and the implementation of the context-vector similarity presented in section 4.2.

4.4 Extension to another pair of languages

The two features used for binary classification are applied identically disregarding the pair of languages involved. Moreover, the co-occurrence model is totally language independent. The context-vector similarity however has been shown to achieve different accuracy depending on the pair of languages involved; but in the case of binary classification of translations, it serves as a side information for the co-occurrence model, which is itself used as a heuristic for alignment of context-vectors since we do not attempt to align pairs of words with null co-occurrences.

For these reasons, we suggest that it is possible to use a decision tree trained on one pair of languages to extract translations from another pair of languages. A similar approach is proposed in (Alfonseca et al., 2008): they present a word decomposition model designed for German language that they successfully applied to other compounding languages. Our approach consists in training a decision tree on a pair of languages and applying this model to the classification of unknown pairs of words in another pair of languages. Such an approach is especially useful for prospecting new translations from less known languages, using a well known language as training.

We used the same algorithms and same features as in the previous sections, but used the data computed from one pair of languages as the training set, and the data computed from another pair of languages as the testing set.

5 Experimental setup

5.1 Corpora

We built several corpora using two different strategies. The first set was built using Wikipedia and the interlingual links available on articles (that points to another version of the same article in another language). We started from the list of all French articles[2] and randomly selected articles that provide a link to Spanish and English versions. We downloaded those, and clean them by removing the wikipedia formatting tags to obtain raw UTF8 texts. Articles were not selected based on their sizes, the

[2] Available on `http://download.wikimedia.org/`.

	[WP] French	[WP] English	[WP] Es	[CLIR] En	[CLIR] Zh
#documents	20,169	20,169	20,169	15,3247	15,3247
#tokens	4,008,284	5,470,661	2,741,789	1,334,071	1,228,330
#unique tokens	120,238	128,831	103,398	30,984	60,015

Table 1: Statistics for all parts of all corpora.

vocabulary used, nor a particular topic. We obtained about 20,000 aligned documents for each language.

A second set was built using an in-house system (unpublished) that seeks for comparable and parallel documents from the web. Starting from a list of Chinese documents (in this case, mostly news articles), we automatically selected English target documents using Cross Language Information Retrieval. About 85% of the paired documents obtained are direct translations (header/footer of web pages apart). However, they will be processed just like aligned comparable documents, that is, we will not take advantage of the structure of the parallel contents to improve accuracy, but will use the exact same approach that we applied for the Wikipedia documents. We gathered about 15,000 pairs of documents employing this method.

All corpora were processed using Tree-Tagger[3] for segmentation and Part-of-Speech tagging. We focused on nouns only and discarded all other tokens. We would record the lemmatized form of tokens when available, otherwise we would record the original form. Table 1 summarizes main statistics for each corpus; [WP] refers to the Wikipedia corpora, [CLIR] to the Chinese-English corpora extracted through cross language information retrieval.

5.2 Dictionaries

We need a bilingual seed lexicon for the context-vector similarity. We used a French-English lexicon obtained from the Web. It contains about 67,000 entries. The Spanish-English and Spanish-French dictionaries were extracted from the linguistic resources of the Apertium project[4]. We obtained approximately 22,500 Spanish-English translations and 12,000 for Spanish-French. Finally, for Chinese-English we used the LDC2002L27 resource from the Linguistic Data Consortium[5] with about 122,000 entries.

5.3 Evaluation lists

To evaluate our approach, we needed evaluation lists of terms for which translations are already known. We used the Medical Subject Headlines, from the UMLS meta-thesaurus[6] which provides a lexicon of specialized, medical terminology, notably in Spanish, English and French. We used the LDC lexicon presented in the previous section for Chinese-English.

From these resources, we selected all the source words that appears from 1 to 5 times in the corpora in order to build the evaluation lists.

5.4 Oracle translations

We looked at the corpora to evaluate how many translation pairs from the evaluation lists can be found across the aligned comparable documents. Those translations are hereafter the *oracle translations*. For French/English, French/Spanish and English/Spanish, about 60% of the translation pairs can be found. For Chinese/English, this ratio reaches 45%. The main reason for this lower result is the inaccuracy of the segmentation tool used to process Chinese. Segmentation tools usually rely on a training corpus and typically fail at handling rare words which, by definition, were unlikely to be found in the training examples. Therefore, some rare Chinese tokens found in our corpus are the results of faulty segmentation, and the translation of those faulty words can not be found in related documents. We encountered the same issue but at a much lower degree for other languages because of spelling mistakes and/or improper Part-of-Speech tagging.

[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

[4] `http://www.apertium.org`

[5] `http://www.ldc.upenn.edu`

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

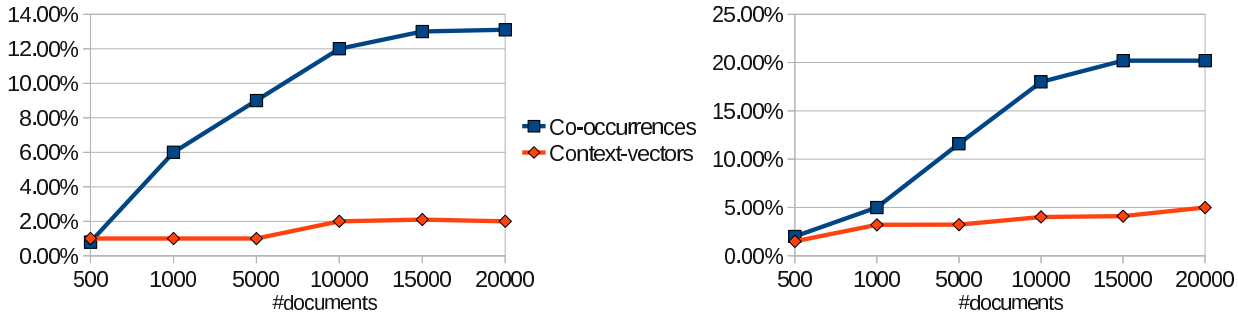

Figure 2: Experiment I: comparison of accuracy obtained for the Top_{10} with the context-vector similarity and the co-occurrence model, for hapaxes (left) and words that appear 2 to 5 times (right).

6 Experiments

We ran three different experiments. Experiment I compares the accuracy of the context-vector similarity and the co-occurrence model. Experiment II uses supervised classification with both features. Experiment III extracts translation from a pair of languages, using a classifier trained on another pair of languages.

6.1 Experiment I: co-occurrence model vs. context-vector similarity

We split the French-English part of the Wikipedia corpus into different samples: the first sample contains 500 pairs of documents. We then aggregated more documents to this initial sample to test different sizes of corpora. We built the sample in order to ensure hapaxes in the whole corpus are hapaxes in all subsets. That is, we ensured the 431 hapaxes in the evaluation lists are represented in the 500 documents subset.

We extracted translations in two different ways:

1. using the co-occurrence model;
2. using the context-vector based approach, with the same evaluation lists.

The accuracy is computed on 1,000 pairs of translations from the set of oracle translations, and measures the amount of correct translations found for the 10 best ranks (Top_{10}) after ranking the candidates according to their score (context-vector similarity or co-occurrence model). The results are presented in figure 2.

We can draw two conclusions out of these results. First, the size of the corpus influences the quality of the bilingual lexicon extraction when using the co-occurrence model. This is especially interesting with hapaxes, for which frequency does not change with the increase of the size of the corpora. The accuracy is improved by adding more information to the corpus, even if this additional information does not cover the pairs of translations we are looking for. The added documents will weaken the association of incorrect translations, without changing the association for rare terms translations. For example, the precision for hapaxes using the co-occurrence model ranges from less than 1% when using only 500 pairs of documents, to about 13% when using all documents. The second conclusion is that the co-occurrence model outperforms the context-vector similarity.

However, both these approaches still perform poorly. In the next experiment, we propose to combine them using supervised classification.

6.2 Experiment II: binary classification of translation

For each corpus or combination of corpora – English-Spanish, English-French, Spanish-French and Chinese-English, we ran three experiments, using the following features for supervised learning of translations:

- the context-vector similarity;
- the co-occurrence model;
- both features together.

The parameters are discussed in section 4.3. We used all the oracle translations to train the positive

	Precision	Recall	F-Measure	Cl.
English-Spanish				
context-vectors	0.0%	0.0%	0.0%	T
	83.3%	99.9%	90.8%	$\neg T$
co-occ. model	66.2%	44.2%	53.0%	T
	89.5%	95.5%	92.4%	$\neg T$
both	**98.6%**	**88.6%**	**93.4%**	T
	97.8%	99.8%	98.7%	$\neg T$
French-English				
context-vectors	76.5%	10.3%	18.1%	T
	90.9%	99.6%	95.1%	$\neg T$
co-occ. model	85.7%	1.2%	2.4%	T
	90.1%	100%	94.8%	$\neg T$
both	**81.0%**	**80.2%**	**80.6%**	T
	94.9%	98.7%	96.8%	$\neg T$
French-Spanish				
context-vectors	0.0%	0.0%	0.0%	T
	81.0%	100%	89.5%	$\neg T$
co-occ. model	64.2%	46.5%	53.9%	T
	88.2%	93.9%	91.0%	$\neg T$
both	**98.7%**	**94.6%**	**96.7%**	T
	98.8%	99.7%	99.2%	$\neg T$
Chinese-English				
context-vectors	69.6%	13.3%	22.3%	T
	91.0%	93.1%	92.1%	$\neg T$
co-occ. model	73.8%	32.5%	45.1%	T
	85.2%	97.1%	90.8%	$\neg T$
both	**86.7%**	**74.7%**	**80.3%**	T
	96.3%	98.3%	97.3%	$\neg T$

Table 2: Experiment II: results of binary classification for "Translation" and "Non-Translation".

values. Results are presented in table 2, they are computed using a 10-folds cross validation. Class T refers to "Translation", $\neg T$ to "Non-Translation". The evaluation of precision/recall/F-Measure for the class "Translation" are given in equation 4 to 6.

$$precision_T = \frac{|T \cap oracle|}{|T|} \quad (4)$$

$$recall_T = \frac{|T \cap oracle|}{|oracle|} \quad (5)$$

$$FMeasure = 2 \times \frac{precision \times recall}{precision + recall} \quad (6)$$

These results show first that one feature is generally not discriminatory enough to discern correct translation and non-translation pairs. For example with Spanish-English, by using context-vector similarity only, we obtained very high recall/precision for the classification of "Non-Translation", but null precision/recall for the classification of "Translation". In some other cases, we obtained high precision but poor recall with one feature only, which is not a usefully result as well since most of the correct translations are still labeled as "Non-Translation".

However, when using both features, the precision is strongly improved up to 98% (English-Spanish or French-Spanish) with a high recall of about 90% for class T. We also achieved about 86%/75% precision/recall in the case of Chinese-English, even though they are very distant languages. This last result is also very promising since it has been obtained from a fully automatically built corpus. Table 3 shows some examples of correctly labeled "Translation".

The decision trees obtained indicate that, in general, word pairs with very high co-occurrence model scores are translations, and that the context-vector similarity disambiguate candidates with lower co-occurrence model scores. Interestingly, the trained decision trees are very similar between the different pairs of languages, which inspired the next experiment.

6.3 Experiment III: extension to another pair of languages

In the last experiment, we focused on using the knowledge acquired with a given pair of languages to recognize proper translation pairs using a different pair of languages. For this experiment, we used the data from one corpus to train the classifier, and used the data from another combination of languages as the test set. Results are displayed in table 4.

These last results are of great interest because they show that translation pairs can be correctly classified even with a classifier trained on another pair of languages. This is very promising because it allows one to prospect new languages using knowledge acquired on a known pairs of languages. As an example, we reached a 77% F-Measure for Chinese-English alignment using a classifier trained on Spanish-French features. This not only confirms the precision/recall of our approach in general, but also shows that the model obtained by training tends

	Tested with			
Trained with	Sp-En	Sp-Fr	Fr-En	Zh-En
Sp-En	98.6/88.8/93.5	98.7/94.9/96.8	91.5/48.3/63.2	99.3/63.0/77.1
Sp-Fr	89.5/77.9/83.9	90.4/82.9/86.5	75.4/53.5/62.6	98.7/63.3/77.1
Fr-En	89.5/77.9/83.9	90.4/82.9/86.5	85.2/80.0/82.6	81.0/87.6/84.2
Zh-En	96.6/89.2/92.7	97.7/94.9/96.3	81.1/50.9/62.5	97.4/65.1/78.1

Table 4: Experiment III: Precision/Recall/F-Measure for label "Translation", obtained for all training/testing set combinations.

English	French
myometrium	myomètre
lysergide	lysergide
hyoscyamus	jusquiame
lysichiton	lysichiton
brassicaceae	brassicacées
yarrow	achillée
spikemoss	sélaginelle
leiomyoma	fibromyome
ryegrass	ivraie
English	**Spanish**
spirometry	espirometría
lolium	lolium
omentum	epiplón
pilocarpine	pilocarpina
chickenpox	varicela
bruxism	bruxismo
psittaciformes	psittaciformes
commodification	mercantilización
talus	astrágalo
English	**Chinese**
hooliganism	流氓
kindergarten	幼儿园
oyster	牡蛎
fascism	法西斯主义
taxonomy	分类学
mongolian	蒙古人
subpoena	传票
rupee	卢比
archbishop	大主教
serfdom	农奴
typhoid	伤寒

Table 3: Experiment II and III: examples of rare word translations found by our algorithm. Note that even though some words such as "*kindergarten*" are not rare in general, they occur with very low frequency in the test corpus.

to be very stable and accurate across different pairs of languages and different corpora.

7 Conclusion

We presented a new approach for extracting translations of rare words among aligned comparable documents. To the best of our knowledge, this is one of the first high accuracy extraction of rare lexicon from non-parallel documents. We obtained a F-Measure ranging from about 80% (French-English, Chinese-English) to 97% (French-Spanish). We also obtained good results for extracting lexicon for a pair of languages, using a decision tree trained with the data computed on another pair of languages. We yielded a 77% F-Measure for the extraction of Chinese-English lexicon, using Spanish-French for training the model.

On top of these promising results, our approach presents several other advantages. First, we showed that it works well on automatically built corpora which require minimal human intervention. Aligned comparable documents can easily be collected and are available in large volumes. Moreover, the proposed machine learning method incorporating both context-vector and co-occurrence model has shown to give good results on pairs of languages that are very different from each other, such as Chinese-English. It is also applicable across different training and testing language pairs, making it possible for us to find rare word translations even for languages without training data. The co-occurrence model is completely language independent and have been shown to give good results on various pairs of languages, including Chinese-English.

Acknowledgments

The authors would like to thank Emmanuel Morin (LINA CNRS 6241) for providing us the comparable corpus used for the experiment in section 2 and the anonymous reviewer for their valuable comments.

References

Enrique Alfonseca, Slaven Bilac, and Stefan Pharies. 2008. Decompounding query keywords from compounding languages. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 253–256.

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.

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

Stefan Evert. 2008. Corpora and collocations. In A. Ludeling and M. Kyto, editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin.

John Firth. 1957. *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman.

Pascale Fung. 2000. A statistical view on bilingual lexicon extraction–from parallel corpora to non-parallel corpora. In Jean Véronis, editor, *Parallel Text Processing*, page 428. Kluwer Academic Publishers.

William A. Gale and Kenneth W. Church. 1991. Identifying word correspondence in parallel texts. In *Proceedings of the workshop on Speech and Natural Language*, HLT'91, pages 152–157, Morristown, NJ, USA. Association for Computational Linguistics.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, Beijing, China, Aug.

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.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 403–411.

Tao Tao and ChengXiang Zhai. 2005. Mining comparable bilingual text corpora for cross-language information integration. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 691–696, New York, NY, USA. ACM.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.