Towards Producing Bilingual Lexica from Monolingual Corpora

Jingyi Han, Núria Bel
Universitat Pompeu Fabra
Roc Boronat, 138, 08018, Barcelona
jingy.han@upf.edu, nuria.bel@upf.edu

Abstract

Bilingual lexica are the basis for many cross-lingual natural language processing tasks. Recent works have shown success in learning bilingual dictionary by taking advantages of comparable corpora and a diverse set of signals derived from monolingual corpora. In the present work, we describe an approach to automatically learn bilingual lexica by training a supervised classifier using word embedding-based vectors of only a few hundred translation equivalent word pairs. The word embedding representations of translation pairs were obtained from source and target monolingual corpora, which are not necessarily related. Our classifier is able to predict whether a new word pair is under a translation relation or not. We tested it on two quite distinct language pairs Chinese-Spanish and English-Spanish. The classifiers achieved more than 0.90 precision and recall for both language pairs in different evaluation scenarios. These results show a high potential for this method to be used in bilingual lexica production for language pairs with reduced amount of parallel or comparable corpora, in particular for phrase table expansion in Statistical Machine Translation systems.

Keywords: automatic bilingual lexicon production, lexical resources, bilingual dictionaries.

1. Introduction

Bilingual lexica are the key resource for many cross-lingual processing tasks and they are crucial components of machine translation systems. A large amount of research has focussed in the automatic production of bilingual dictionaries, mostly based on the large parallel or comparable corpora. However, such large bilingual related corpora are not readily available for many language pairs. Considering this data shortage problem, our objective is learning bilingual lexicons out of large monolingual, and not necessarily comparable, corpora in source and target languages.

The first work in this area by Rapp (1995) was based on the hypothesis that translation equivalents in two languages have similar distributional profiles or co-occurrence patterns. Recently, Mikolov et al. (2013a) showed that word embeddings (distributed, dense and real-valued vector representations of words) indeed project word semantics into a vector space from their distributional characteristics. More interestingly, it is claimed that the relationship between vector spaces that represent different language word semantics can be captured by a linear transformation. Therefore, we propose to use word embedding vectors of translation word pairs to train a supervised classifier which can predict whether a new pair of words in two different languages can be under a translation relation, according to the regularities learned from a small training set.

A previous attempt to use supervised classification for inducing bilingual lexica from non-parallel nor comparable corpora is Irvine and Callison-Burch (2013), although their approach is based on using non-distributional information (signals like temporal similarity, topical information, orthographical similarity, among others) to train the model. Moreover, the use of supervised methods has proved to be effective in statistical machine translation (SMT) too, a closely related task (Och & Ney, 2002).

Our approach benefits from a supervised method and from the dense and small vectors produced by word2vec tool (Mikolov et al. 2013b) to build a binary classifier which can find the general relation between translation word pairs. Our classifier achieved good performance on binary classification evaluation. The learning curve shows that with a rather small quantity of training data (about 300 positive and 5-1 negative random examples) it is possible to achieve more than a 90% accuracy for two quite different language pairs: Spanish-English and Chinese-Spanish. The results are especially encouraging because for the Chinese-Spanish language pair there is not many parallel or comparable corpora to exploit. In addition, in order to compare the performance of our classifier with the work of Mikolov et al. (2013a) which approached the task of finding translation pairs as a ranking task, we conducted another experiment on WMT11 data by ranking all test candidates according to the classifier confidence score.

The rest of the paper is structured as follows: section 2 reports the previous works related to our approach; section 3 describes our supervised bilingual lexicon learning method; section 4 sets the experimental framework; section 5 reports our test results; and section 6 shows error analysis; section 7 describes the ranking experiment; and finally, in section 8 conclusion and future work are presented.

2. State of the Art

Different previous works have shown how to learn bilingual lexicons from non-parallel, but still comparable corpora, i.e. collections of source-target document pairs that are not direct translations but are topically related. Yu and Tsujii (2009) extracted bilingual lexica from comparable corpora by considering the similarity of syntactic dependencies. Matsumoto et al. (2013) generated a dictionary by combining topic modeling and alignment techniques. Ananiadou et al. (2014) extracted bilingual terminology from comparable corpora using compositional and contextual clues. The main limitation of these approaches is that their performance still depends on the availability of large comparable corpora.

Recently, several interesting works treat bilingual lexicon generation as a classification problem. For example, Aker et al. (2013) generated bilingual terminologies from comparable corpora by using a SVM binary classifier with training data derived from EUROVOC thesaurus (Steinberger et al. 2002) and GIZA++ phrase-table-based features plus cognate information features. The performance of their classifier reaches 100% precision for 15 of the 21 language pairs addressed, but recall in these cases remains around a 70% average.

Irvine and Callison-Burch (2013) employed a supervised approach (a linear classifier trained by stochastic gradient descent to minimize squared error) and combined extra-linguistic monolingually-derived signals (contextual, temporal, topical, orthographic, and frequency) as features for the model. For the training, they used 1250 positive examples and the number of negative instances is three times the number of positive. Their results are delivered in the form of ranked lists of English translations for 22 languages achieving very different top-10 accuracy rates: the best results are for Spanish with 85% and the worst for Nepali with 13.6%. Differences are not related, though, with monolingual data available and the learning curve shows that performance is stable after about 300 positive training instances.

The approach presented here is similar to Irvine and Callison-Burch (2013) but it uses only linguistic distributional features to train a SVM classifier. Our method basically trains a classifier using as features the word embedding representations proposed by Mikolov et al. (2013b).

Word embedding vector representation has been shown to afford relevant distributional information in different semantic tasks: word similarity judgments and word analogy detection (Baroni et al. 2014; Levy et al. 2015 among others). Mikolov et al. (2013a) in particular proposed using this distributed representation to automate the process of generating bilingual dictionaries and phrase tables. Their method learns a transformation matrix between vector spaces of two particular languages on the data provided by a 5000 entries seed dictionary. At test time, a new word can be translated by projecting its vector representation from the source language space to the target language space. Once the vector space in the target language is obtained, similar target language vectors (found by cosine similarity assessment) are ranked as possible translations. This transformation matrix is found via optimization with a stochastic gradient descent algorithm. Their results in the form of ranked lists are further refined with a confidence threshold that tries to balance precision and recall, i.e. coverage. Thus, for the pair English-Spanish, with a coverage of 92.5%, precision at top position is 53%. Best precision reported is 78% (better results are obtained when refining with edit distance) but with a coverage of 17%.

In our approach we were inspired by Necsulescu et al. (2015) recent evidence that simple concatenation of word embeddings is effective for finding lexical semantic relations (i.e. hyponymy, hyperonymy, meronymy, attribution and properties) holding in word pairs with supervised methods. Our task was accordingly defined as whether a SVM could learn the translation relation between source and target words.

3. Supervised Bilingual Lexicon Learning with Word Embeddings

The task of our experiment was to train a SVM binary classifier with vectors made of concatenated word embeddings of source and target words. Word embeddings are produced from monolingual, not necessarily domain related, corpora for the two languages involved. In testing mode, new word pairs are classified as being one the translation of the other or not.

For the training and testing set, each translation pair is represented by concatenating the word embedding vector representation of the source word and of its corresponding translation, for positive examples, and of random words for negative ones. Formally, given a translation word pair (x, y), x being a source and y a target word, whose vector features are $v(x) = (x_1, x_2, \dots, x_n)$ and $v(y) = (y_1, y_2, \dots, y_n)$ respectively, then $v(x, y)$ is defined as the concatenation of $v(x)$ and $v(y)$: $v(x, y) = (x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n)$. In this work, we only experimented with unigrams of three word classes: noun, verb and adjective.

For the ranking task on WMT11, the experiment scenario is similar to Statistical Machine Translation production of phrase tables, where pairs of words are extracted from all possible combinations of words occurring in a given set of aligned sentences and the probability of a particular word being the translation of others is estimated. With our method, each of the source word representation from the test set was concatenated with all target word (only nouns) representations from the corpus. Then all the concatenated candidates were ranked by the confidence score produced by our classifier.

4. Methodology

In this section, we describe the experimental settings and the results of our supervised learning classifier. The outline of our experiments is: (i) Generation of the right and wrong translation lists. (ii) Obtaining the corresponding word vector representation from monolingual word embedding models. (iii) Concatenation of the vector representations of the source word and its translation equivalent (or random word for negative instances). (iv)Training a Sequential Minimal Optimization (SMO[1]) classifier using the previously generated concatenated representation. (v) Evaluating the classifier by binary classification testing and top-k ranking task.

[1] As implemented in WEKA (Hall et al., 2009)

4.1 Data sets

We conducted our experiments on two quite distinct language pairs Chinese and Spanish (CH-ES), and English and Spanish (EN-ES). The monolingual corpora used were: Chinese Wikipedia Dump corpus [2] (54M words); Spanish Wikipedia corpus [3](150M, 2006 dump); and for English, the BNC[4] (100M). Despite the fact that Spanish and Chinese corpora are Wikipedia dumps, there is no intended topic overlap. Also note that Chinese corpus is much smaller than Spanish one, therefore they cannot be considered neither parallel nor comparable. In addition, the corpora that we used for the ranking task are WMT11[5] text data of English (59M) and Spanish (59M).

Training set and test set are prepared in the same way. For the binary classification experiment, each source word can only be paired with one target word. To obtain a translation list (or positive instances called *right translation*) for training and testing, we randomly extracted a list of words for each PoS (only noun, verb and adjective), from the ES monolingual corpus. To PoS tag the Spanish and English corpora, we used Stanford PoS Tagger[6] (Toutanova et al., 2003). These randomly selected words were translated from source language (ES) to target language (EN and CH) using on-line Google Translator. Since not all the produced translations could be found in the target monolingual corpus, we removed from our datasets those words whose corresponding translation was not in the target corpus because we needed to obtain its word embedding.

To build the no-translation set (called *no translation*), we randomly selected non-related source and target words from the monolingual corpus of each language and randomly combined them. The ratio was 5 negative instances for each positive example.

This dataset was divided into training and testing sets. Final figures of the datasets are provided in Table 1.

	ES-CH				ES-EN			
	Training		Testing		Training		Testing	
	YES	NO	YES	NO	YES	NO	YES	NO
Noun	451	2390	99	449	449	2379	94	469
Adj.	302	1492	71	398	300	1500	99	500
Verb	400	1999	113	599	300	1500	99	500
Total	1153	5881	283	1446	1049	5379	292	1469

Table 1: Translation pair datasets for ES-CH and ES-EN

4.2 Word Embedding

We obtained word embeddings from the monolingual corpora described in 4.1 for the Spanish, English and Chinese words in the *right translation* and *no translation* lists using the Continuous Bag-of-words (CBOW) method as implemented in word2vec[7] tool, because it is faster and more suitable for larger datasets (Mikolov et al., 2013a). To train the CBOW models we used the parameters with window size 8, minimum word frequency 5 and 200 dimensions for both source and target vectors. For the ranking experiment, we used 300 dimensions for all vectors.

5. Classification Experiment

5.1 Evaluation and results

We trained and tested SMO (Platt, 1998) classifiers on ES-EN and ES-CH for three word categories: noun (N), adjective (Adj) and verb (V), and another for the three categories together. The evaluation was double, as we performed a 10 fold cross-validation with the training set and we tested again the model with the held-out test set. The results in terms of precision (P), recall (R) and F1-measure (F1) are presented in Tables 2 and 3. They show average P, R and F1 of both classes (*right translation* and *no translation*) separately for the experiment with all word categories.

CH-ES	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
N	0.947	0.948	0.948	0.933	0.934	0.931
Adj	0.916	0.918	0.917	0.934	0.936	0.932
V	0.955	0.956	0.955	0.957	0.958	0.958
All	0.927	0.928	0.927	0.941	0.942	0.941
YES	0.845	0.796	0.82	0.83	0.809	0.819
NO	0.948	0.962	0.955	0.963	0.967	0.965

Table 2: Test result for Spanish and Chinese

EN-ES	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
N	0.963	0.963	0.963	0.944	0.945	0.944
Adj	0.964	0.964	0.964	0.953	0.952	0.952
V	0.965	0.966	0.965	0.927	0.93	0.928
All	0.922	0.924	0.922	0.921	0.922	0.921
YES	0.804	0.708	0.753	0.782	0.736	0.758
NO	0.944	0.966	0.955	0.948	0.959	0.954

Table 3: Test results for Spanish and English

When observed by classification classes, it is evident that the *no translation* class results are better than the *right translation* class. For ES-EN, we achieved F1 around 0.75 for *right translation* and around 0.95 for *no translation* in both testing scenarios; for ES-CH, results are slightly better with F1 around 0.82 for *right translation*, and 0.96 for *no translation*. We show several examples of translation equivalents, which are correctly classified by our classifier, in Table 4.

ES-CH	ES-EN
amistoso - 友好 (friendly)	económico - economic
antiguo - 古老 (old)	eficiente - efficient
cabeza - 头 (head)	atractivo - attractive
característica - 特征 (characteristic)	actividad - activity
quemar - 烧伤 (burn)	cama - bed
provocar - 导致 (provoke)	idioma- language

Table 4: Examples of translation pairs correctly classified

[2] https://archive.org/details/zhwiki_20100610
[3] http://hdl.handle.net/10230/20047
[4] http://www.natcorp.ox.ac.uk/
[5] http://www.statmt.org/wmt11/training-monolingual.tgz
[6] http://nlp.stanford.edu/software/tagger.shtml
[7] https://code.google.com/archive/p/word2vec/

To explore the relation between the performance of the classifier and the number of training instances, Figure 1 plots the learning curves (F1, P and R) over different number of positive instances from 50 to 450, with negative instances from 250 to 2250, for the language pair Chinese and Spanish. It shows that the ES-CH classifier achieved good results with only 300 positive and 1500 negative training instances. Note that Irvine and Callison-Burch (2013) also have similar learning curve result.

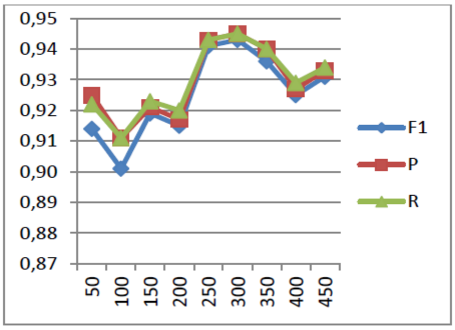

Figure 1: Learning curve over different number of positive instances, up to 450, for ES-CH. The number of negative training instances is five times the number of positive ones.

5.2 Discussion

The evaluation results show that the classifier is able to generalize to a large extent. We carried out an error analysis to assess whether the results could be considered an upper bound limit for the task or there was room for improvements. Error analysis, however, is hindered by the nature of the vectors used: being word embeddings a projection, no special feature selection study can be easily performed (Levy et al. 2014).

We mainly looked at the 54 cases of false negative (FN) produced by the ES-CH classifier on nouns. After manual inspection of FN, we found out that the test set raised different issues. 12 FN were found to correspond to inaccuracies of the test set. They were: (i) One of the words had a wrong, or very unusual, PoS tag therefore the word embedding could only take into account few occurrences. This is the case for pairs such as "católica" ('catholic', normally an adjective but with some occurrences tagged as a noun in the corpus) and "天主教 "('catholicism', a noun in Chinese); (ii) Foreign words were normally misclassified, for example: "number" (an English word in the Spanish corpus) was present in the test-set with the corresponding translation " 号 " ('number'); (iii) Misspellings: some Spanish nouns were misspelled in the corpus. This is the case of "perído" (instead of 'período' –period). The classifier did not recognised that '时期' is indeed a possible translation. The impact of these issues comes from the word embedding representation obtained, which would reflect only a reduced number of occurrences, and therefore, of distributional information.

In line with this reasoning, i.e. that wrong words/pos could not provide good word embeddings, we checked the accuracy achieved with other infrequent words in the corpus. Indeed, 13 word pairs that contained words whose frequency was lower than 100 occurrences, such as "autonómico"- "区域性" ('regional' in a geopolitical sense in Spanish) and "carnívoro"-" 肉食性 "('carnivorous') were also misclassified. We also checked whether among the correctly classified pairs there were similar low frequent words, and indeed it was not the case.

However for some other errors the explanation is less obvious. In 7 cases, we found that, although the translation provided by Google translate could be correct in very particular contexts, there is a semantic difference between the members of the pair: the Chinese word is more general than the Spanish one, or the other way around: "pueblo" (inhabited place or group of people) was paired with "村" (only inhabited place, i.e. village), "reflexivo" ('thoughtful' or 'reflective') was paired with " 反光 " ('light reflective'), or "enlace" ('link' but also 'wedding') with "链接" (only 'link' in Chinese). In these cases, the classifier did not found the pair to hold the translation relation.

6. Ranking Experiment

In order to compare with the research of Mikolov et al. (2013a) that delivered results in terms of a ranked list of possible translations, we used the confidence score (the reliability on the classification decision which ranges from 0 to 1, for a particular instance to belong to a particular class) to fit the ranking task. In this experiment we trained the new classifiers with WMT11 datasets as used by Mikolov et al. (2013a) following the outline of our previous experiments, but applying two different ratios: balanced proportion of positive and negative examples, and 5 negative instances for each positive example. This experiment was conducted on the language pair ES-EN, and only with nouns. The datasets for training and testing are shown in Table 5.

	Training		Testing	
	YES	NO	YES	NO
1:1	990	990	434	832
1:5	990	4950	434	832

Table 5: Translation pair datasets for ES-EN

The classifier was evaluated in two different ways: binary classification accuracy and top-10 ranking task according to its corresponding confidence score. The results for binary classification using two different ratios are provided in Tables 6 and 7.

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.878	0.895	0.886	0.794	0.710	0.750
NO	0.893	0.876	0.884	0.857	0.904	0.880
total	0.885	0.885	0.855	0.835	0.838	0.835

Table 6: Test results for balanced dataset

	10 cross-validation			Held-out test set		
	P	R	F1	P	R	F1
YES	0.825	0.835	0.830	0.963	0.533	0.686
NO	0.967	0.964	0.966	0.802	0.989	0.886
total	0.943	0.943	0.943	0.857	0.833	0.818

Table 7: Test results for the ratio 1:5

Compared to the results of classification experiment on the language pair ES-EN (Table 3), we obtained similar performance on 10-cross validation for the same ratio (1:5). However, for the held-out test set, results of both classes decrease unexpectedly due to some low frequent test data, such as *soy*, *signaling*, *underuse* and *skepticism,* whose frequency are only 6, 10, 5 and 78, respectively. Judging from Table 6 and Table 7, it is obvious that the classifier achieved better result on *right translation* when we gave balanced proportion of positive and negative examples to the training set. On the contrary, when we included 5 negative instances for each positive example, we obtained better results on *no translation* class.

For the ranking task, we tried to compare our method with the results of the transformation matrix proposed by Mikolov et al., (2013a). To do so, we created a new test set. Each member of the source language test set was paired with all the target language words as possible translation pairs. We used only nouns, both source and target in order to reduce the computational load, resulting in 24,706 pairs for each source noun. After using the classifier, specially trained with the WMT 2011 corpora, the classification confidence score was used to rank all translation pairs classified as *right translation* expecting to find the right word pair ranked in top positions. However, many word pairs obtained the same confidence score making it impossible to properly set up the ranking list. To better understand the results of our ranking experiment, we give some examples of our test result in Table 8. Note that the reported ranking position is with respect to the candidate translations for each source word.

Translation pairs	Ranking Position	Confidence Score
sugar_azúcar	6	0.978
shipyard_astillero	163	1
square_plaza	197	0.922
tribune_tribuna	362	1
sphere_esfera	512	1
sir_señor	694	0.99

Table 8: Examples of ranking experiment

In Table 8, all the examples have been correctly classified and obtained quite high confidence score. However, their position in the ranking list cannot be directly related to their confidence score. For instance, the word pair *sugar_azúcar* obtained higher position compared with other examples, but its confidence score is lower than most of the others. In the case of *shipyard_astillero, tribune_tribuna* and *sphere_esfera,* although they all achieved the score 1, their ranking position are quite different.

7. Conclusions and future work

In this paper, we have proposed a novel method to learn bilingual lexicon from monolingual corpora by training a supervised classifier. On average, we obtained quite good results on binary classification evaluation. However, we could not compare the results in the ranking task as proposed by Mikolov et al., (2013a) when relying only on the classifier confidence score. It is noticeable, that a number of particular target words are getting high confidence score as possible translation of many different source words, what could be a consequence of the 'hubness problem' as reported by Dinu et al. (2015). To further investigate it is part of our future work.

Despite the fact that the confidence score supplied by the classifier is insufficient to tackle the ranking task, judging from its outstanding performance of the classification experiment, we expect it to be very useful for being applied, for instance, to expand phrase tables of SMT systems when no parallel or comparable corpora is available. This is also our future work, in which after producing a bilingual lexicon with the word pairs classified as *right translation*, we will use results for phase table expansion and evaluate its impact on the translation performance.

8. Acknowledgements

Han Jingyi is supported by PhD grant FI-DGR from the Generalitat de Catalunya.

9. References

Aker, A., Paramita, M. and Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the Conference of the Association for Computational Linguistics*.

Baroni, M., Dinu, G. and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Conference of the Association for Computational Linguistics*.

Dinu, G., Lazaridou, A., Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the International Conference on Learning Representations*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.-H. (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Irvine, A. and Callison-Burch, C. (2013). Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics*.

Kontonatsios, G., Korkontzelos, I., Tsujii, J and Ananiadou, S. (2014). Combining String and Context Similarity for Bilingual Term Alignment from Comparable Corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Levy, O. and Goldberg, Y. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of CoNLL*.

Liu, Xd., Kevin, D. and Matsumoto, Y. (2013). Topic Models + Word Alignment = A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*.

Mikolov, T., Le, Q.-V. and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. In *Proceedings HLT-NAACL*.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013b). Efficient estimation of word representations in vector space. In *arXiv* preprint arXiv:1301.3781.

Necsulescu, S., Mendes, S., Jurgens, D., Navigli, R. and Bel, N. (2015) Reading Between the Lines: Overcoming Data Sparsity for Accurate Classification of Lexical Relationships. In *Proceedings of StarSem.*

Och, F.-J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics.*

Platt, J.-C. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in Kernel Methods: Support Vector Learning.*

Rapp, R. (1995). Identifying word translations in nonparallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics.*

Steinberger, R., Pouliquen, B. and Hagman, J. (2002). Cross-lingual Document Similarity Calculation using the Multilingual Thesaurus EUROVOC. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics.*

Toutanova, K., Klein, D., Manning, C. and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics.*

Yu, K. and Tsujii, J. (2009). Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics.*