Advanced Science and Technology Letters
Vol.54 (Multimedia 2014), pp.37-40
http://dx.doi.org/10.14257/astl.2014.54.10

A Neural Network Algorithm for Extracting Bilingual Lexica

Hongseok Kwon[1], Hyeongwon Seo[1], Minah Cheon[1] and Jaehoon Kim[1],

[1] Korea Maritime and Ocean University,
Dongsam-Dong, Yeongdo-gu, Busan, South Korea
hong8c@naver.com, wonn24@gmail.com, minah014@outlook.com, jhoon@hhu.ac.kr

Abstract. We propose a neural network algorithm for extracting bilingual lexica, which is based on the well-known Perceptron algorithm. In this paper, we modify the Perceptron algorithm in order to learn with unlabeled training data for bilingual lexicon extraction. To show the feasibility of the algorithm, we also present a novel approach on bilingual lexicon extraction with the proposed neural network algorithm. The empirical results have shown that our proposed method is significantly improving the performances of our model obtained with the modified Perceptron algorithm.

Keywords: Bilingual lexicon extraction, Neural network algorithm, Comparable corpora, Perceptron

1 Introduction

Bilingual lexicons are an important role in many natural language processing tasks, for example, statistical machine translation (SMT) and cross lingual information retrieval (CLIR), and so on. Basically, bilingual lexicons can be obtained by manually extracting appropriate translation pairs for each language, but it is too time-consuming and labor-intensive. For these reasons, many researchers have focused on automatic bilingual lexicon extraction. The direct way of automatic bilingual lexicon extraction is to align words from parallel corpora. However, collecting a large amount of parallel corpora is onerous and restricted to specific domains in some less-known language pairs. For all these reasons, researchers turn to extracting bilingual lexicons from comparable corpora [1].

One of the approaches in the bilingual lexicon extraction is the context-based approach using information retrieval (IR) techniques [2]. This approach showed significant performances for high-frequent words, but a large-scale seed dictionary is required to translate context-vectors.

Recently, Chatterjee [3] proposed an iterative approach, which extracts new translation candidates, uses the candidates as a new seed dictionary, and repeats the procedure until convergence. The iterative approach has shown significant improvement of the accuracy in a few epochs.

With taking advantages of two approaches, in this paper, we propose an iterative context-based method for bilingual lexicon extraction using the Perceptron algorithm.

ISSN: 2287-1233 ASTL

Furthermore we modify the Perceptron algorithm for bilingual lexicon extraction in order to automatically generate training examples.

2 Methodology

In this section, we describe our main work for bilingual lexicon extraction. Our approach consists of two methods: the context-based approach (CBA) [4] for constructing a seed dictionary and the iterative approach (IA) for extracting a bilingual lexicon. We first exploit CBA to construct an initial seed dictionary from two parallel corpora. The seed dictionary is used to translate a source synonym vector written by the source languages into the target language and used as weights for the modified Perceptron in IA, and then apply IA to obtain bilingual lexicons. Details of IA will be described in following Section 3.1. IA requires two linguistic resources: $W(0)$ and comparable corpora. The $W(0)$ is employed as initial weights for the modified Perceptron algorithm as described in Section 2 and conceptually used to translate source synonym vectors into their corresponding target synonym vectors as mentioned before. Comparable corpora are employed for generating synonym vectors in both source and target languages. We use source synonym vectors instead of context vectors as input vectors of the modified Perceptron algorithm because a synonym vector for each word can add new weights into W and as the result we can find translation candidates for new source words. For example, if synonyms of the word 'father' are 'dad', 'daddy', 'papa', and so on and translation candidates for 'daddy' do not exist in an initial seed dictionary, we can find the candidates through learning process in the modified Perceptron algorithm. The implementation of the IA can be carried out by applying the following steps.

- Step (1): To build source synonym vectors $\boldsymbol{S}$ and target synonym vectors $\boldsymbol{T}$. We first build source context vectors and target context vectors in both source language and target language respectively, as in the same way of CBA and the context vectors are represented as words with a fixed window size of ± 2 as the context. The words in a source context vector $\boldsymbol{s}_c$ are weighted by X^2 scores and are selected by the critical value of 3.841 as threshold. In the same way, the words in a context vector $\boldsymbol{t}_c$ are weighted. Next, a source synonym vector $\boldsymbol{s} \in \boldsymbol{S}$ (a target synonym vector $\boldsymbol{t} \in \boldsymbol{T}$) are computed according to similarity scores between source context vectors (target context vectors).
- Step (2):To generate the translated vector $\mathbf{y}$ of a source synonym vector $\boldsymbol{x}$ instead of $\boldsymbol{s}$[1] using the modified Perceptron algorithm as in Equation (1):

$$y_j = \sum_{j=0}^{|\mathbf{y}|} x_i w_{ij}, \qquad (1)$$

where $x_i \in \boldsymbol{x}$ is the i-th source synonym word, $y_j \in \mathrm{y}$ is the j-th translated word in target language, and w_{ij} is a weight between x_i and y_j.

[1] To help readers to understand the notation, we substitute the notation for $\boldsymbol{s}$ with $\boldsymbol{x}$ as the input of the Perceptron.

- Step (3): To determine the desired synonym vector $\boldsymbol{d}$ of $\boldsymbol{x}$ as follows:
$$\boldsymbol{d} = \text{argmax}_{\boldsymbol{t}\in\boldsymbol{T}}\, \text{sim}(\boldsymbol{y},\boldsymbol{t}) = \text{argmax}_{\boldsymbol{t}\in\boldsymbol{T}}\, \cos(\boldsymbol{y},\boldsymbol{t}), \quad (2)$$
where $\cos(\boldsymbol{y},\boldsymbol{t})$ is a cosine similarity of $\boldsymbol{y}$ and $\boldsymbol{t}$. As the result, the pair of $(\boldsymbol{x},\boldsymbol{d})$ is one of the training examples of the modified Perceptron.
- Step (4): To learn W via the Perceptron learning algorithm.
- Step (5): To repeat the step (2) and (4) until convergence.
- Step (6): To sort the top k word pairs based on Equation (1).

3 Preliminary Experiments

In this section, we evaluate our approach for two different language pairs that are Korean-Spanish (KR-ES) and Korean-French (KR-FR). The accuracy is used as an evaluation metric. In this paper, we built two pair of comparable corpora that are KR-ES and KR-FR from the news articles and Europarl corpus [5]. The KR corpus was taken from the news articles on the Web and contains 800,000 sentences. The ES and FR were also collected from the news articles on the Web and from Europarl corpus and have 800,000 sentences each. All words were tokenized and lemmatized using the following tools: U-tagger [6] for Korean and Tree-Tagger [7] for Spanish and French. All words in Spanish and French were converted to lower case, and those in Korean are morphologically analyzed into morphemes and pos-tagged by U-tagger. Next content words[2] which occurring more than five were considered when generating context vectors in all languages.

We built two sets of evaluation dictionaries (KR-ES and KR-FR) to evaluate the performance of the proposed method manually using the Web dictionary[3]. Each lexicon is unidirectional, meaning that they list the meanings of words of one language in another, and contains 150 high frequent words (denoted by HIGH hereafter).

We built two sets of initial seed dictionaries using the CBA. The sets are used to translate source synonym vectors and to use as initial weights for iterative approach. All entries in the seed dictionaries of language pairs have 20 maximum translation candidates and are all nouns. The number of the entries is about 30,500 in KR-ES and 27,700 in KR-FR respectively.

We conducted 60 epochs with the learning rate $\alpha = 0.01$ for the KR-ES and KR-FR language pairs. The accuracy@1 of the HIGH words is shown in Fig. 3. As shown in Figure 3, the accuracy@1 is slightly increased during 60 epochs. The accuracy of the KR-ES increased from 0.366 to 0.406 and the KR-FR increased from 0.413 to 0.440. In general, our proposed method can significantly improve the accuracies in both languages pairs.

---

2 KR (Sejong tagset): NNG, VV, VA, MAG, SL
ES (Penn Treebank tagset): NC, NMEA, NP, PE, ACRNM, NMON, ADJ, ADV, UMMX, VCLIger, VCLIinf, VCLIfin, VEadj,VEfin, VEger, VEinf, VHadj, VHfin, VHger, VHinf, VLadj, VLfin, VLger, VLinf, VMadj, VMfin, VMger, VMinf, VSadj, VSfin, VSger, VSinf
FR (Penn Treebank tagset): ABR, NOM, ADJ, ADV, INT, VER

3 http://www.dic.naver.com

4 Conclusions and Future works

We have presented an iterative approach for extracting bilingual lexicon extraction from comparable corpora using a modified perceptron algorithm, starting from the context-based approach to construct a seed dictionary as weights that are learned by the modified perceptron algorithm, and continuing with the iterative approach. The basic characteristics of this approach are that it can further improve the accuracy and needs no labels of the training examples for learning weights via the modified Perceptron algorithm. Our experimental results showed that the iterative approach with the modified perceptron helps improve the accuracy.

There are still several future works under consideration. Currently, the proposed method has many parameters to adjust for improving the performance, and was only tested on nouns. In the future, we will adjust parameters to improve the performance. Besides, we will expand to different categories except nouns. Lastly, we will handle multi-word expressions.

Acknowledgements. This work was supported by the IT R&D program of MSIP/KEIT. [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning]

Reference

1. Fung, P.: Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In: 3rd Workshop on Very Large Corpora, pp. 173--183 (1995)
2. Rapp, R.: Identifying word translations in non-parallel texts. In: 33rd Annual Meeting of the Association for Computational Linguistics, pp. 320--322 (1995)
3. Chatterjee, D., Sarkar, S., Mishra, A.: Co-occurrence Graph based iterative bilingual lexicon extraction from comparable corpora. In: 4th International Workshop on Cross Lingual Information Access, pp. 35--42 (2010)
4. Kwon, H., Seo H., Kim, J.: Enhancing performance of bilingual lexicon extraction through refinement of pivot-context vectors. In: Journal of KIISE: Software and Applications, vol. 45, pp. 492--500 (2014)
5. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Machine Translation Summit X, pp 79--86 (2005)
6. Shin, J., Ock, C.: A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary. In: Journal of KIISE: Software and Applications, vol. 39, pp. 415--424 (2012)
7. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing, pp. 44-49 (1994)

