Contemporary Engineering Sciences, Vol. 7, 2014, no. 24, 1335 - 1343
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.49165

Iterative Bilingual Lexicon Extraction from Comparable Corpora Using a Modified Perceptron Algorithm

Hong-Seok Kwon, Hyeong-Won Seo, Minah Cheon, Jae-Hoon Kim

Korea Maritime and Ocean University
727 Taejong-ro, Yeongdo-gu, Busan 606-791, South Korea

Abstract

We present a novel iterative approach on bilingual lexicon extraction from comparable corpora. The approach is based on vector space model for word representation and a modified Perceptron algorithm. The approach requires a seed dictionary and a large amount of unlabeled training data. The seed dictionary is generated using the pivot-based approach and the unlabeled training data is dynamically labeled by the modified Perceptron algorithm using a similarity measure during learning process. In this paper, we extract bilingual lexicons by iteratively applying our proposed approach via the modified Perceptron algorithm. The empirical results have shown that our proposed approach significantly improves the accuracy for the top 1 candidate. In the future we will try to apply the multilayered Perceptron algorithm to our iterative approach for effective word representation.

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

1. Introduction

Bilingual dictionaries play an important role in many domains, for example, machine translation (MT), cross lingual information retrieval (CLIR), and so on. Basically, bilingual lexicons can be obtained by manually extracting appropriate translation pairs from texts, but it takes a lot of effort to actually get the appropriate translation pairs. Hence, automatic bilingual lexicon extraction has received considerable attention since the 1990s.

The direct way of automatic bilingual lexicon extraction is to align word pairs from parallel corpora. However, parallel corpora are not available for less-known language pairs and collecting a large amount of the parallel corpora is onerous and restricted to specific domains. So in recent years, extracting bilingual lexicons from comparable corpora has become the more popular approach. The approach that is widely used in bilingual lexicon extraction is the context-based approach using information retrieval (IR) techniques [1, 2]. This approach showed significant performances for high-frequent words, but a large-scale seed dictionary is required to translate context-vectors. Recently, Chattergee and Chu [3, 4] proposed an iterative approach which extracts bilingual lexicons, uses extracted bilingual lexicons as a new seed dictionary, and repeats the extraction procedure until convergence. The iterative approach has shown improvement of the performance in a few epochs.

With taking advantages of the two approaches, in this paper, we propose an iterative method for bilingual lexicon extraction using a Perceptron algorithm. Besides we modify the Perceptron algorithm in order to dynamically assign labels to unlabeled training data during learning process.

2. Related works

2.1. Standard approach

Widely used approach in the bilingual lexicon extraction is a context-based approach (CA) known as the standard approach using information retrieval techniques [1, 2]. Generally, the standard approach builds context vectors for each source and text word, translates the source context vectors in a target language using a bilingual dictionary called a seed dictionary, and compares the translation with the target context vector in order to get their translation candidates. This approach showed significant performances for high-frequent words, but a large-scale seed dictionary is required to translate context-vectors and can affect the performances of the system.

2.2. Pivot-Based Approach

The standard approach uses comparable corpora and a seed dictionary [1, 2]. The performance, however, is dependent on the size and the quality of the seed dictionary and moreover, constructing the large amount of and the high-quality of

the seed dictionary is tedious and expensive. Kwon *et al* [5, 6] proposed the pivot-based approach (PA) using two parallel corpora sharing a pivot language like English with more accurate alignment information instead of comparable corpora. The pivot language represents both of source context vectors and target context vectors which are comparable to each other because they are the same dimension represented in the pivot language. As the result, PA does not need the initial seed dictionary anymore. Besides, PA uses a freely available word aligner, called Anymalign, to construct context vectors. Anymalign showed high accuracy for low-frequent words to extract translation candidates [8]. The PA can be summarized in the following three steps. 1) To build a source context vector and a target context vector for each word in a source language (e.g., KR) and a target language (e.g., ES) using two independent parallel corpora that are KR-EN and EN-ES, respectively. All words in the context vector are weighted by Anymalign [5, 6]. 2) To calculate the similarity between a source context vector and a target context vector. We use the cosine measure for the similarity. 3) To sort the top k target words for a source word based on their similarity scores.

3. Methodology: Iterative Approach

The overall structure of the proposed method is depicted in Fig 1. Our approach consists of two methods: the pivot-based approach (PA) [5, 6] and the iterative approach (IA) [7]. We first exploit PA to construct an initial seed dictionary from two parallel corpora sharing a pivot language like English. Next, we apply the IA to obtain bilingual lexicons using the seed dictionary and the modified Perceptron algorithm. The seed dictionary is used to translate a source synonym vector to a target synonym vector and the association between word pairs in the seed dictionary is used as initial weights for the modified Perceptron. The IA requires two linguistic resources: the initial seed dictionary (W(0)) and comparable corpora for source and target languages. The W(0) is employed as initial weights for the modified Perceptron algorithm and conceptually used to translate source synonym vectors into their corresponding target synonym vectors as mentioned before. Comparable corpora are employed for generating the synonym vectors in both source and target languages. We use synonym vectors instead of context vectors as input vectors of the modified Perceptron algorithm. The reasons are that the synonym vector can resolve one-hot word representation problem and each word in the synonym vector can be 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:

i. To build source synonym vectors (denoted as $\boldsymbol{S}$) and target synonym vectors (denoted as $\boldsymbol{T}$). We first build source context vectors and target context vectors in both source language (denoted as L_s) and

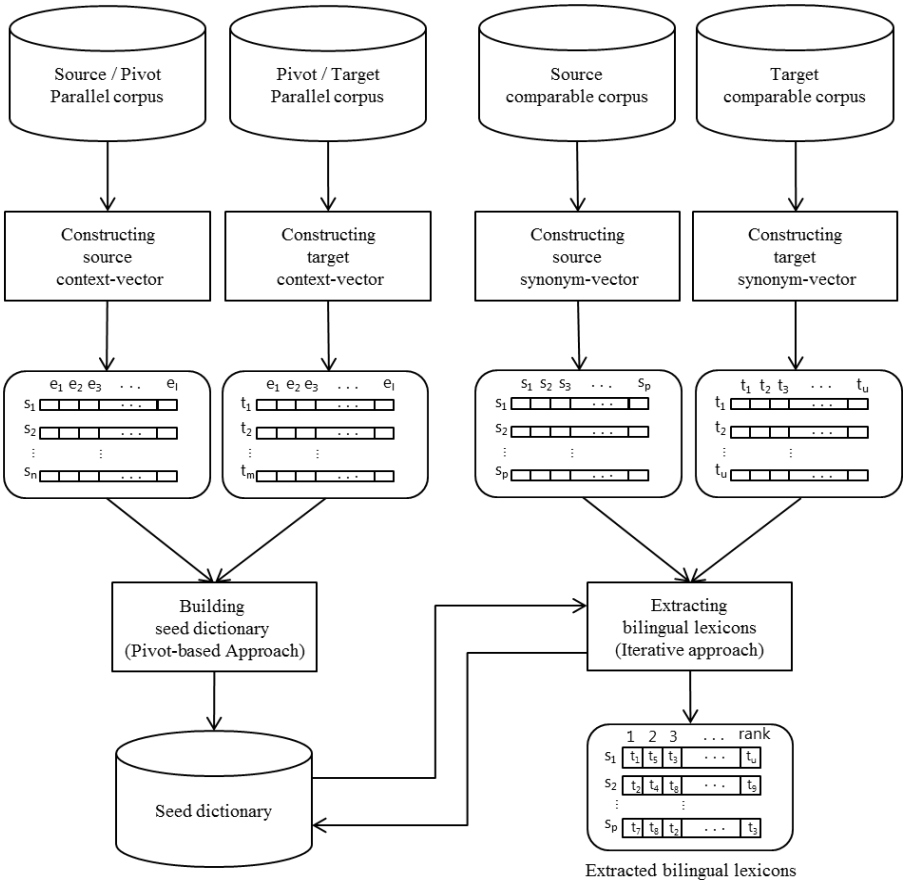

Fig. 1. An overall structure of the proposed method.

target language (denoted as L_t) respectively, as in the same way of the PA and the context vectors are represented as words with a fixed window size of 5 as the context. The words in a source context vector (denoted as 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 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).

ii. To generate the translated vector ($\mathbf{y}$) of a source synonym vector ($\boldsymbol{x}$) instead of $\boldsymbol{s}$ (to help readers to understand notations, we substitute the notation for s with x as the input of the Perceptron) using the modified Perceptron algorithm as in Equation (1):

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

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

iii. To determine the desired synonym vector $\boldsymbol{d}$ of $\boldsymbol{x}$ as follows:

Learning Algorithm

Input: synonym-vectors **S** and **T**, seed dictionary $\mathbf{W}(0)$
for $e = 1, \ldots, E$ **do**
  $\Delta\mathbf{W} = \mathbf{W}(e-1)$
  for $\boldsymbol{x} \in \boldsymbol{S}$ **do**
    $\boldsymbol{y} = 0$
    for $x_i \in \boldsymbol{x}$
      for $y_j \in \boldsymbol{y}$
        $y_j \mathrel{+}= x_i w_{ij}$
      end for
    end for
    $\boldsymbol{d} = \underset{\boldsymbol{t}\in\boldsymbol{T}}{\operatorname{argmax}} sim(\boldsymbol{y}, \boldsymbol{t})$
    $\Delta\mathbf{W} = \alpha(\boldsymbol{d} - \boldsymbol{y})\boldsymbol{x}$
    for $w_{ij} \in \Delta\mathbf{W}$ **do**
      if $w_{ij} < \mathbf{0}$ **then** $w_{ij} = \mathbf{0}$
    end for
    $\mathbf{W}(e) = \mathbf{W}(e\text{-}1) + \Delta\mathbf{W}$
  end for
end for
return W

Fig. 2. A modified Perceptron algorithm for extracting bilingual lexicon

$$\mathbf{d} = \operatorname{argmax}_{\mathbf{t}\in T} \operatorname{sim}(\mathbf{y}, \mathbf{t}) = \operatorname{argmax}_{\mathbf{t}\in T} \cos(\mathbf{y}, \mathbf{t}) \quad (2)$$

where $\cos(\mathbf{y}, \mathbf{t})$ is a cosine similarity of $\mathbf{y}$ and $\mathbf{t}$. As the result, the pair of $(\mathbf{x}, \mathbf{d})$ is one of the training examples of the Perceptron.

iv. To learn W via the Perceptron learning algorithm.

v. To repeat the step (ii) to (iv) until convergence.

vi. To sort the top k word pairs based on Equation (1).

Finally, we can perform the modified Perceptron algorithm. In summary, our modified Perceptron algorithm for updating weights is shown in Fig 2.

4. Experiments and results

In this section, we evaluate our approach for two different language pairs that are Korean-Spanish (KR-ES) and Korean-French (KR-FR), and compare with the standard approach as a baseline. Accuracy@1 (ACC) and Mean Reciprocal Rank (MRR) are used as evaluation metrics. The accuracy@1 means the accuracy of the Top 1.

4.1. Comparable corpora

In this paper, we built two pair of comparable corpora that are KR-ES and KR-FR from the news articles and Europarl [9] corpus. The KR corpus was taken from the news articles on the Web[1] and contains 800,000 sentences. The ES and FR were also collected from the news articles on the Web[2] and from Europarl corpus and have 800,000 sentences each. The average of the words in sentence is 16.2 in KR, 15.9 in ES and 16.1 in FR respectively.

4.2. Data pre-processing

All words were tokenized and lemmatized using the following tools: U-tagger [10] for Korean and Tree-Tagger [11] 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, only content words[3] which occurring more than five were considered when generating context vectors in all languages.

4.3. Building evaluation dictionary

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. Each lexicon is unidirectional, meaning that they list the meanings of words of one language in another. The evaluation dictionary contains 150 high frequent words (denoted by HIGH hereafter) and 150 low frequent words (denoted by LOW hereafter) from comparable corpus respectively.

4.4. Building a seed dictionary

We use the PA to build two initial seed dictionaries that are KR-ES and KR-FR. The quality of the two initial seed dictionaries are evaluated using ACC with 200 high and low frequent evaluation words from parallel corpus respectively. Table 1 shows accuracies of the PA for 200 high and 200 low. We exploit these two sets of initial seed dictionaries as inputs of the IA.

Table 1. Performance of the PA

Language sets	# of entries	Accuracy of 200 high	Accuracy of 200 low	Accuracy of (high+low)/2
KR-ES	30,500	63.5%	41.0%	52.2%
KR-FR	27,700	74.0%	53.5%	63.7%

[1] KR: http://www.donga.com/

[2] ES: http://www.abc.es/
FR: http://www.lemonde.fr/

[3] 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

4.5. Results

We conducted 60 epochs with the learning rate α=0.01 for the KR-ES and KR-FR language pairs. The ACCs and MRRs for the HIGH and LOW words are shown in Table 2. As shown in Table 2, ACCs of the HIGH and LOW are slightly increased during 60 epochs. ACCs of the KR-ES increased from 0.366 to 0.406 and the KR-FR increased from 0.413 to 0.440 for the HIGH. The performances of IA on the KR-ES and the KR-FR are better than the baseline (KR-ES) about 0.24 and 0.28 respectively. For the LOW, ACC of the KR-ES improved from 0.187 to 0.207 and the KR-FR improved from 0.353 to 0.373 respectively. Furthermore, the performance outperforms the baselines in the both language pairs.

The MRR of the KR-ES is increased about 0.015 from the first epochs 0.485 and the KR-FR is increased about 0.029 for the HIGH. For the LOW, the MRR is increased about 0.014 on KR-ES and decreased about 0.001 on KR-FR. The reason for decreasing MRR is that the IA is largely dependent on the synonym vectors. If the synonym vectors would be inaccurate, the modified perceptron algorithm might be learned incorrectly. It means the system cannot be found correct translation candidates. In the same manner, the performance for the LOW outperforms the baselines in the both language pairs.

Table 2. Accuracies of the IA for the HIGH and LOW.

Language sets		**Baseline**		**Proposed method**			
				Initial epoch		**60 Epoch**	
		ACC	MRR	ACC	MRR	ACC	MRR
KR-ES	HIGH	0.126	0.172	0.366	0.485	0.406	0.500
	LOW	0.033	0.057	0.186	0.263	0.206	0.276
KR-FR	HIGH	0.133	0.209	0.413	0.543	0.440	0.572
	LOW	0.073	0.096	0.353	0.436	0.373	0.435

4.6. Discussions

In this paper, we propose an iterative approach, which is based on the pivot-based approach [5, 6]. Our proposed method has three advantages. First, we do not require a large size of an initial seed dictionary. The initial seed dictionary is generated by pivot-based approach described in Section 2.1. Second, we do not need labels of training examples. A modified Perceptron algorithm dynamically generates labels of the training examples during epochs. Third, accuracy can be increased during epochs.

5. Conclusions and future work

In this paper, we have described an iterative approach for bilingual lexicon extraction from comparable corpora using a modified Perceptron algorithm, starting from the pivot-based approach to build an initial seed dictionary as weights

that are learned by the modified Perceptron algorithm, and continuing with the iterative approach. Regarding the empirical results of our proposition, the iterative approach can further improve the accuracy and shows good performance without labels of the training examples.
In the future works, we will adjust parameters to improve the performance and expand to different categories except nouns and try to apply the multilayered Perceptron algorithm to our iterative approach for effective word representation. Lastly, we will handle multi-word expressions.

Acknowledgement. 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]. This work is based on my master thesis from Korea Maritime and Ocean University.

References

[1] P, Fung, Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus, *Proc. of 3rd Int. Workshop on Very Large Corpora*, (1995), pp. 173-183.

[2] R, Rapp, Identifying word translations in non-parallel texts, *Proc. of 33rd Annual Meeting of the Association for Computational Linguistics*, (1995), pp. 320-322.

[3] D. Chatterjee, S. Sarkar and Mishra, A, Co-occurrence graph based iterative bilingual lexicon extraction from comparable corpora, *Proc. of 4th Int. Workshop on Cross Lingual Information Access*, (2010), pp. 35-42.

[4] C. Chu, T. Nakazawa and S. Kurohashi, Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge, *Proc. of 15th Int. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing'14)*, (2014), pp. 296-309.

[5] H. Kwon, H. Seo and J. Kim, Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tools, *Proc. of 6th Int. Workshop on Building and Using Comparable Corpora (BUCC'13)*, (2013), pp. 11-15.

[6] H. Kwon, H. Seo and J. Kim, Enhancing performance of bilingual lexicon extraction through refinement of pivot-context vectors, *Journal of KIISE: Software and Applications*, (2014), Vol. 45, pp. 492-500.

[7] H. Kwon, Bilingual lexicon extraction using a modified perceptron algorithm. Master thesis, Department of Computer Engineering, Korea Maritime and Ocean University (2014).

[8] A, Lardilleux, Y. Lepage and F. Yvon, The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach, *Journal of Advanced Intelligence*, (2011), Vol. 3, No. 3, pp. 189-217.

[9] P. Koehn, Europarl: A parallel corpus for statistical machine translation, *Proc. of Machine Translation Summit X*, (2005), pp. 79-86.

[10] J. Shin and C. Ock, A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary, *Journal of KIISE: Software and Applications*, (2012), Vol. 39, pp. 415-424

[11] H. Schmid, Probabilistic part-of-speech tagging using decision trees, *Int. Conf. on New Methods in Language Processing*, (1994), pp. 44-49.

Received: August 17, 2014