Extracting Bilingual Lexica from Comparable Corpora Using Self-Organizing Maps

Hyeong-Won Seo **Min-Ah Cheon** **Jae-Hoon Kim**

Department of Computer Engineering, Korea Maritime and Ocean University,
Busan 606-791, Republic of Korea

wonn24@gmail.com minah014@outlook.com jhoon@kmou.ac.kr

Abstract

This paper aims to present a novel method of extracting bilingual lexica from comparable corpora using one of the artificial neural network algorithms, self-organizing maps (SOMs). The proposed method is very useful when a seed dictionary for translating source words into target words is insufficient. Our experiments have shown stunning results when contrasted with one of the other approaches. For future work, we need to fine-tune various parameters to achieve stronger performances. Also we should investigate how to construct good synonym vectors.

1 Introduction

Bilingual lexicon extraction from comparable corpora has been studied by many researchers since the late 1990s (Fung, 1998; Rapp 1999; Chiao & Zweigenbaum, 2002; Ismail & Manandhar, 2010; Hazem & Morin, 2012).

To our knowledge, one of the basic approaches is the context vector-based approach (Rapp, 1995; Fung, 1998) called the standard approach in the literatures, and many other studies have been derived from this approach. Some of these are concerned with similarity score measurement (Fung, 1998; Rapp, 1999; Koehn & Knight, 2002; Prochasson *et al*., 2009), the size of the context window (Daille & Morin, 2005; Prochasson *et al*., 2009), and the size of the seed dictionary (Fung, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002; Koehn & Knight, 2002; Daille & Morin, 2005). The extended approach, one of such approaches, (Déjean *et al*., 2002; Daille & Morin, 2005) has been proposed in order to reduce the load on the seed dictionary. It gathers *k* nearest neighbors to augment the context of the word to be translated. In spite of their efforts, using comparable corpora for extracting such lexica yields quite poor performances unless orthographic features are used. However, such features may bring other costs.

Under the circumstances like this, this paper is motivated to propose an efficient method in which comparable corpora with a minimum of resources are considered for extracting bilingual lexica. The SOM-based approach, we propose in this paper, can yield stronger performances with the same experimental circumstance than earlier studies can do. In order to show this, we compare the proposed method to the standard approach. Of course, it does not meaning our method outperforms for every data. We just show the proposed method is reasonable for this field.

The rest of the paper is structured as follows: Section 2 presents several works closely related to our method. Section 3 describes our method (the SOM-based approach) in more detail. Section 4 shows experimental results with discussions, and Section 5 concludes the paper and presents future research directions.

2 Related Works

2.1 Context-based approach

As has been noted earlier, the standard approach (Rapp, 1995; Fung, 1998) is proposed to extract bilingual lexica from comparable corpora. It uses

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 62–67,
Beijing, China, July 30, 2015.

contextually relevant words in a small-sized window. Selecting similar context vectors between source and target languages is the key feature of the approach. Since the approach uses comparable corpora, a seed dictionary to translate one to another language is required. Additionally, a large scale of corpora as well as sufficient amount of initial seed dictionaries should be prepared for a better performance.

2.2 Self-organizing maps

A self-organizing map (SOM) (Kohonen, 1982; 1995) is one of the artificial neural network models and represents a huge amount of input data in a more illustrative form in a lower-dimensional space. In general, a SOM is an unsupervised and competitive learning network. It has been studied extensively in recent years. For example, SOMs have been studied in pattern recognition (Li *et al*., 2006; Ghorpade *et al*., 2010), signal processing (Wakuya *et al*., 2007), multivariate statistical analysis (Nag *et al*., 2005), data mining (Júnior *et al*., 2013), word categorization (Klami & Lagus, 2006), and clustering (Juntunen *et al*., 2013).

Since a SOM tries to keep the topological properties of input data, semantically/geometrically similar inputs are generally mapped around one neuron, usually in the form of a two-dimensional lattice (*i.e*. a map). Significantly, the SOM can be used for clustering the input vectors and finding features inherent to the problem. In this perspective, we can expect that actual similar words have one common winner (winning neuron) or share the same neighbors if input vectors are semantically well-formed.

Based on this characteristic, a main idea of the proposed method is to make two different words that are translations of each other have one common winner. If a new input data has a similar input trained already, the SOMs can extract its translations based on its neighbors. Consequently, neighbors (i.e. semantically similar words) also share similar areas in the feature map.

3 SOM-based approach

The overall structure of the SOM-based approach can be summarized as follows (see Figure 1 for more details):

i. Building synonym vectors: In this paper, the synonym vector indicates a vector that consists of words semantically related to each other. Therefore, synonym vectors should be constructed in a semantic fashion not a co-relational fashion. For example, the vector for *baby* should very similar to the vector for *kid* not just for closely related *toy* or *sitter*. Therefore, building synonym vectors is one of the most important issues in this work. For this, we firstly build context vectors via contextually related words in a fixed-size window. This context vector is weighted by an association measure, such as the PMI or the chi-square. After context vectors are built, similarity scores between the vectors are computed. In this paper, the similarity score, as occurs so often in information retrieval, is computed by cosine similarity.

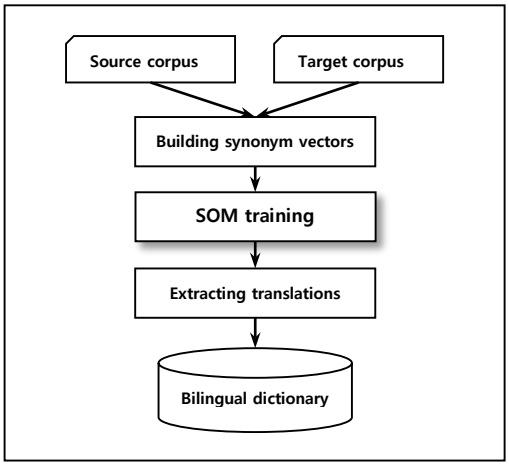

Figure 1. Overall structure of SOM-based approach

Synonyms can be identified based on the scores higher than a reasonable threshold. Synonym vectors are then weighted by the scores. For instance, let kid be a base word to be vectored. In this case, its elements are similarity scores between kid and the most similar *k* words, such as *baby*, *teenager*, and *youth*. Consequently, well-made synonym vectors have a SOM reflects the topological properties of input data and will obtain common winners after the SOMs are trained.

Note that such context vectors are very sensitive to experimental data and parameters such as association/similarity measures, so any kind of vector is welcomed here. We just assume semantically formed synonym vectors are already available before we train SOMs.

ii. SOM training: After the source and target synonym vectors are built, we train two sorts of SOMs in different ways. Figure 2 describes how two SOMs are trained interactively.

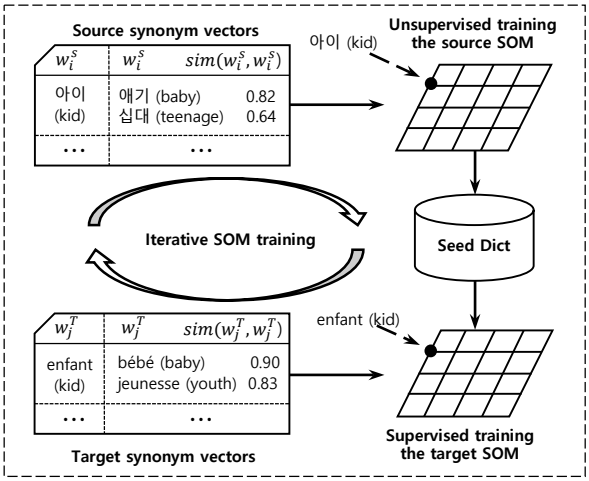

Figure 2. SOM training

Firstly, we train the source SOM in an unsupervised fashion. The general SOM algorithm to train all source words can be summarized as follows:

1) Set an initial weight vector $w(0)$ with small random values [0, 1], and set learning rate $\eta(t)$ with a small positive value $(0 < \eta(t) \leq \eta(t-1) \leq 1)$. The iteration t is for one input data.

2) For every single input x, find the winning neuron (*i.e.* winner) c which has minimum score based on Euclidean distances between an input and weight vectors $\|x - w_c\| = \min_i \|x - w_i\|$.

3) Update the weight vectors of winning neuron c and its neighbors as follows:
$w_i(t+1) = w_i(t) + \eta(t)h(t)[x(t) - w_i(t)]$, where t denotes time, $x(t)$ is an input vector at t, and $h(t)$ is the neighborhood kernel around the winner c. In this step, we update them in online mode which means one update per one input (*c.f.* an offline mode means one update per all inputs).

4) Repeat the steps 2) and 3) until a certain termination condition like the maximum number of iterations is reached.

After the source SOM is trained in an unsupervised fashion, we train the target SOM in a supervised fashion. In this case, most of steps are the same with the case of the source. Note that we should aware of updating the target weight vectors. Target winners which of words excluded in the seed dictionary are updated naturally as the case of the source. The others which of words included in the seed dictionary are updated by calling related source winners. Therefore, two different words which are translations to each other can be located in the same topological location of two different SOMs. We think that we can teach correct labels to insiders (*i.e.* the target words that included in the seed dictionary) not for outsiders. As mentioned before, if synonym vectors are well-formed as well as two SOMs are well-trained, a source word and its translation will have one common winner. Although a target word is not trained yet, the word can be extracted when its synonym is trained.

iii. Extracting translations: After two SOMs are trained interactively, SOM vectors should be constructed based on each feature map (*i.e.* the source and target). In this case, similarity scores between an input vector and weight vectors become elements of SOM vectors. That is, a length of the SOM is a dimension of the SOM vector.

After the SOM vectors are built, similarity scores between one source SOM vector and all target SOM vectors are calculated by cosine similarity. And then, the top *k* candidates are selected and added to the bilingual lexicon.

4 Experiments

In this paper, we evaluate the proposed method for two language pairs – Korean–French (KR–FR) and Korean–Spanish (KR–ES). Regarding the comparison, we implemented the standard approach mentioned in Section 2.1. Note that the standard approach implemented here is not complete. There are many chances to show better performances by fine-tuning several parameters, such as the size of the context window, and association/similarity measures. However, we can briefly estimate them because both methods are implemented by using same resources. Several parameters are fixed as follows: the context size of the window as 5, and the association measure as a chi-square test, and the similarity measure as a cosine similarity. These measures were empirically chosen from our experimental data.

We used three comparable corpora (Kwon *et al.*, 2014) in Korean, French, and Spanish. Each corpus included around 800k sentences collected from the Web[1]. The Korean corpus consists of news articles and some are derived from different sources (Seo *et al.*, 2006). The others also consists

[1] Korean: http://www.naver.com,
French: http://www.lemonde.fr, and Spanish: http://www.abc.es

of news articles (around 400k sentences), and some are combined with the European parliament proceedings (400k randomly sampled sentences) (Koehn, 2005). The Korean corpus has around 280k word types (180k for French and 185k for Spanish), and the average number of words per sentence is 16.2 (15.9 for French and 16.1 for Spanish). Consequently, the balance of three corpora is well-formed.

We extracted nouns from these corpora for our test sets as well as input data. We considered only nouns to reduce the sizes of the dimensions of either synonym vectors or SOMs. Thus, we finally collected almost 190k Korean noun types (45k for French and 58k for Spanish). The reason why the number of Korean noun types was higher than others was due to Korean characteristics. We should split the Korean words into morpheme units because there are a lot of compound words and omitted morphemes. Furthermore, we collected very finely segmented Korean nouns to eliminate indulgent compound nouns that were possibly missed during a word segmentation task. All collected nouns were considered candidates of both test sets and seed words independently.

After the input data was prepared, we built synonym vectors, as mentioned previously. We already introduced the method how to construct synonym vectors. However, this paper doesn't mainly propose the efficient way of representing words semantically in vector spaces. If synonym vectors are built based on context vectors and their similarity scores, the size of the vector dimension would be very huge. It would cause many time-consuming problems. In this paper, we simply use word2vec[2] to build synonym vectors. As far as we know, word2vec cannot yield semantically related vectors as output. However, we used this tool to reduce vector sizes and assume these outputs (*i.e.* vectors) are reasonable as the input data for training SOMs. Some parameters for building synonym vectors can be presented as follows: window size is 5, word vector size is 100, and training iteration is 100.

4.1 Evaluation dictionary

We manually built evaluation dictionaries to evaluate our method because such dictionaries for KR–FR/–ES are publicly unavailable. Each dictionary contains 200 high-frequency nouns. The reason why we picked high-frequency nouns is that these nouns have more chances to have neighbors than low-frequency words. In order to evaluate whether the proposed approach is valid (*i.e.* whether trained SOM can extract new input data that not trained), we need to train words having many neighbors. These 200 source words were selected if actual translations were in their corpora. Thus, the 200th source word did not indicate a 200th high-frequency word. The KR→FR[3] dictionary had total of 288 translations (451 translations in the FR→KR dictionary), and the KR→ES dictionary contained 377 translations (687 translations in the ES→KR dictionary). Additionally, regretfully, there were several duplicated translations for every language. In the case of KR–FR, the Korean words had 447 French translations (420 types) and the French words had 209 Korean translations (189 types). In the case of KR–ES, the Korean words had 456 Spanish translations (369 types) and the Spanish words had 509 Korean translations (421 types). We did not perform any heuristic process to give each source word a unique sense. Instead, we assumed related source words corresponding to a single translation were semantically the same.

4.2 Seed dictionary

The seed dictionaries were also built manually based on the high-frequency nouns as mentioned before. Seed words, however, were not overlapped with evaluation data. We chose 11,910 Korean noun types (8,105 French types and 7,458 Spanish types) out of 94% of the total words in the corpus. As mentioned before, 11,910 Korean noun types out of 190k (total) noun types is an extremely low number. Except 200 of the highest-frequency words (contained in the evaluation dictionary), we finally collected 2,399 Korean seed nouns having their translations in the target corpora for KR→FR, 4,387 Korean seed nouns for KR→ES, 2,138 French seed nouns for FR→KR, and 1,813 Spanish seed nouns for ES→KR, respectively.

5 Results

Unfortunately, we do not have a publicly accepted gold standard or experimental guidelines in these language pairs. By and large, the best performances depended on various experimental settings, such as languages, document domains, and

[2] http://code.google.com/p/word2vec/

[3] The symbol '→' means unidirectional way (*i.e.* source to target only).

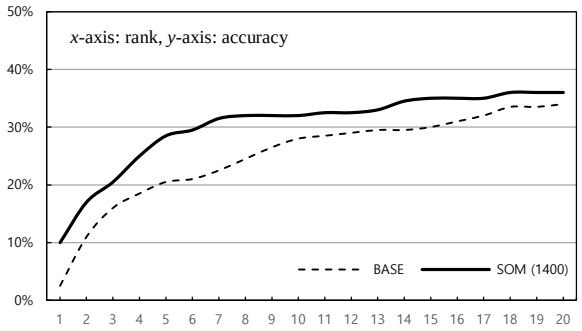

Figure 2. Accuracies of the KR→FR pair

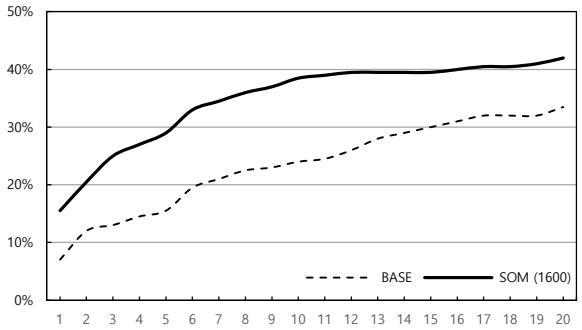

Figure 3. Accuracies of the FR→KR pair

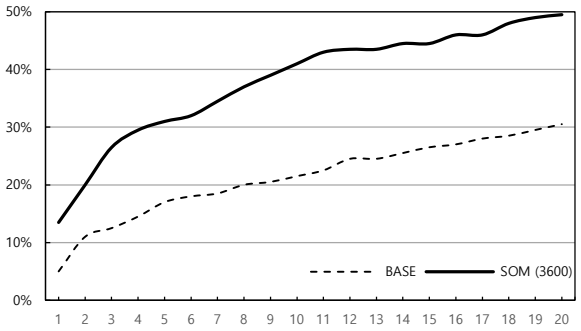

Figure 4. Accuracies of the KR→ES pair

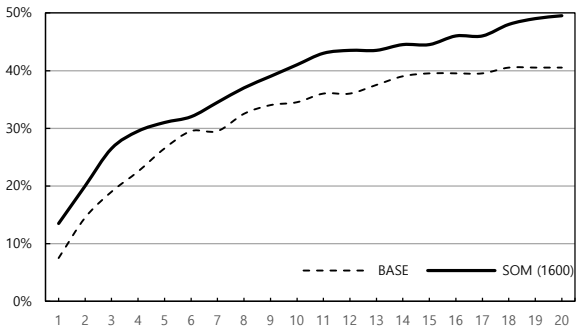

Figure 5. Accuracies of the ES→KR pair

seed dictionaries. Doubtless, the quality of synonym vectors and seed dictionaries including trained SOMs are the most important issues for achieving high performances. Additionally, we ignore evaluations of the quality of synonym vectors in this paper. We only consider accuracies for the top 20 candidates for two sets of language pairs (*i.e*. KR–FR and KR–ES).

For simplified experiments, we fixed several parameters as follows: The dimension of the synonym vector as 100, the size of the Gaussian function as 25 (5×5), the learning rate as 0.1, and the epoch as 2000. These parameters were given based on preceding experiments. In case of a SOM size, all sizes are different for covering most of seed words (one-to-one mapping had shown poor performances due to the fixed and small-sized Gaussian function). We tried to find the best parameters via fine-tuning, but most could be further improved in future research.

The accuracies for two sets of language pairs are described in Figures 2 to 5. In those figures, the BASE means the standard approach, the SOM means the SOM-based approach, the number around brackets means a size of the SOM, *x*-axis indicates ranks, and *y*-axis indicates accuracies. As can be seen, the SOM-based approach outperformed the standard approach over all language settings.

In our experimental results of the KR to FR pair, for example, we extracted *stratégie* (strategy) as the translation of the source word **전략** (jeonryak[4]; strategy, operation) where their neighbors, **작전** [5] (jakjeon; operation, tactic, strategy) and *opération*[6] (operation), are included in the seed dictionary. If new input data (to be tested) have very similar seed words, we can extract correct translations through well-trained SOMs. Although the sizes of SOMs were neither the same nor the best sizes, we can see the proposed approach is quite outstanding compared with the standard approach.

6 Conclusion

This paper proposes a novel method for extracting bilingual lexica from comparable corpora by using SOMs. The method trains two sorts of SOMs, either in an unsupervised fashion and a supervised fashion, respectively. As we can see the experimental results, our method generally outperforms the standard approach under the same experimental conditions (*i.e*. the same seed dictionaries and corpora). Although the given parameters are not the best for both approaches so far, our method shows stunning results.

For future work, we can tune parameter factors such as the size of SOMs, the Gaussian function, and the epoch. Moreover, various parts-of-speech

[4] The Korean gloss is presented before a semicolon in brackets.
[5] The similarity score between **작전** and **전략** is 0.88.
[6] The similarity score between opération and stratégie is 0.82.

could be considered, as we only considered nouns in this work. In addition, a deep analysis of errors is required.

Acknowledgements

This work was supported by the ICT R&D program of MSIP/IITP. [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]

References

Y.-C. Chiao and P. Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational Linguistics*, Coling'02, pp. 1208–1212.

B. Daille and E. Morin. 2005. French-English terminology extraction from comparable corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, IJCNLP'05, pp. 707–718.

H. Déjean, É. Gaussier, and F. Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, 1: 1-7.

P. Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In *Proceedings of the 3rd conference of the Association for Machine Translation in the Americas*, Amta'98, pp. 1–16.

S. Ghorpade, J. Ghorpade, and S. Mantri. 2010. Pattern Recognition Using Neural Networks. *International Journal of Computer Science & Information Technology*, IJCSIT, 2(6): 92-98.

A. Hazem and E. Morin. 2012. Qalign: A new method for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'12, volume 2 of Lecture Notes in Computer Science, pp. 83–96.

A. Ismail and S. Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Coling'10, pp. 481–489.

E. Júnior, G. Breda, E. Marques, and L. Mendes. 2013. Knowledge discovery: Data mining by self-organizing maps. *Web Information Systems and Technologies*, Lecture Notes in Business Information Processing, 140: 185–200.

P. Juntunen, M. Liukkonen, M. Lehtola, and H. Yrjö. 2013. Cluster analysis by self-organizing maps: An application to the modelling of water quality in a treatment process. *Applied Soft Computing*, 13(7): 3191–3196.

M. Klami and K. Lagus. 2006. Unsupervised word categorization using self-organizing maps and automatically extracted morphs. *Intelligent Data Engineering and Automated Learning*, IDEAL 2006, volume 4224 of Lecture Notes in Computer Science, pp. 912–919.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora, In *Proceedings of the Association for computational linguistic on unsupervised lexical acquisition*, pp. 9–16.

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceeding of 10th Conference on Machine Translation Summit*, 79–86.

T. Kohonen. 1982. Self-organized formation of topologically correct feature maps. In *Biological Cybernetics*, 43(1): 59–69.

T. Kohonen. 1995. Self-organizing maps. Springer, volume 30.

H. Kwon, H.-W. Seo, M.-A. Cheon, and J.-H. Kim. 2014. Iterative bilingual lexicon extraction from comparable corpora using a modified perceptron algorithm. *Journal of Contemporary Engineering Sciences*, 7(24): 1335–1343.

C. Li, H. Zhang, J. Wang, and R. Zhao. 2006. A new pattern recognition model based on heuristic SOM network and rough set theory. In *Vehicular Electronics and Safety 2006*, ICVES 2006, IEEE International Conference on, pp. 45–48.

A. K. Nag, A. Mitra, and S. Mitra. 2005. Multiple outlier detection in multivariate data using self-organizing maps title. *Computational Statistics*, 20(2): 245–264.

E. Prochasson, E. Morin, and K. Kageura. 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings of the 12th Conference on Machine Translation Summit (MT Summit XII)*, pp. 284–291.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '33, pp. 320–322.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL '99, pp. 519–526.

H.-W. Seo, H.-C. Kim, H.-Y. Cho, J.-H. Kim, and S.-I. Yang. 2006. Automatically constructing English–Korean parallel corpus from Web documents (in Korean). In *Proceeding of 26th Conference on Korea Information Processing Society fall conference*, KIPS, 13(2): 161–164.

H. Wakuya, H. Harada, and K. Shida. 2007. An architecture of self-organizing map for temporal signal processing and its application to a braille recognition task (in Japanese). *Systems and Computers in Japan*, 38(3):62–71. Translated from Denshi Joho Tsushin Gakkai Ronbunshi, J87-D-II (3): 884–892.