공학석사 학위논문

Bilingual Lexicon Extraction Using a Modified Perceptron Algorithm

변형된 퍼셉트론 알고리즘을 사용한 이중언어 어휘 추출

지도교수 김 재 훈

2014년 8월

한국해양대학교 대학원

컴퓨터공학과

권 홍 석

본 논문을 권홍석의 공학석사 학위논문으로 인준함.

위원장 공학박사 이 장 세 (인)

위 원 공학박사 박 휴 찬 (인)

위 원 공학박사 김 재 훈 (인)

2014년 7월 21일

한국해양대학교 대학원

Contents

List of Figures

List of Tables

변형된 퍼셉트론 알고리즘을 사용한 이중언어 어휘 추출

권홍석

컴퓨터공학과

한국해양대학교 대학원

초록

전산 언어학 분야에서 병렬 말뭉치와 이중언어 어휘는 기계번역과 교차 정보 탐색 등의 분야에서 중요한 자원으로 사용되고 있다. 예를 들어, 병렬 말뭉치는 기계번역 시스템에서 번역 확률들을 추출하는데 사용된다. 이중언어 어휘는 교차 정보 탐색에서 직접적으로 단어 대 단어 번역을 가능하게 한다. 또한 기계번역 시스템에서 번역 프로세스를 도와주는 역할을 하고 있다. 그리고 학습을 위한 병렬 말뭉치와 이중언어 어휘의 용량이 크면 클수록 기계번역 시스템의 성능이 향상된다. 그러나 이러한 이중언어 어휘를 수동으로, 즉 사람의 힘으로 구축하는 것은 많은 비용과 시간과 노동을 필요로 한다. 이러한

이유들 때문에 이중언어 어휘를 추출하는 연구가 많은 연구자들에게 각광받게 되었다.

본 논문에서는 이중언어 어휘를 추출하는 새롭고 효과적인 방법론을 제안한다. 이중언어 어휘 추출에서 가장 많이 다루어지는 벡터 공간 모델을 기반으로 하고, 신경망의 한 종류인 퍼셉트론 알고리즘을 사용하여 이중언어 어휘의 가중치를 반복해서 학습한다. 그리고 반복적으로 학습된 이중언어 어휘의 가중치와 퍼셉트론을 사용하여 최종 이중언어 어휘들을 추출한다.

그 결과, 학습되지 않은 초기의 결과에 비해서 반복 학습된 결과가 평균 3.5%의 정확도 향상을 얻을 수 있었다

KEY WORDS: 이중언어 어휘 추출 (Bilingual lexicon extraction), 비교 말뭉치(Comparable corpora), 퍼셉트론 (Perceptron)

Chapter 1

Introduction

Bilingual lexicons play an important role in many natural language processing tasks, such as statistical machine translation (SMT) (Brown et al., 1993) and cross lingual information retrieval (CLIR) (Grefenstette, 1998), 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 labour-intensive. For these reasons, many researchers have focused on automatic bilingual lexicon extraction. The direct and simple way of automatic bilingual lexicon extraction is to align words in parallel corpora (Wu and Xia, 1994), which contain source texts and their translations. 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 (Fung, 1995; Yu and Tsujii, 2009; Ismail and Manandhar, 2010).

One of the approaches in the bilingual lexicon extraction is the context-based approach using information retrieval (IR) techniques (Rapp, 1995; Fung, 1998; Gaussier et al., 2004; Hazem et al, 2011; Seo et al., 2013). This approach has shown significant performances for high-frequent words, but a large-scale seed dictionary is required to translate context-vectors.

Recently, Chatterjee et al., (2010) and Chu et al., (2014) 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 advantage of the two approaches, in this thesis, we propose an iterative method for bilingual lexicon extraction using a Perceptron algorithm. Besides we modify the Perceptron algorithm for bilingual lexicon extraction in order to automatically generate training examples. The main idea underlying our method is that bilingual translation words in a comparable corpus may be occurred within different contexts, especially for the different domain. Furthermore, while translating source words into target words, the data sparseness problem may be suffered due to a small size of an initial seed dictionary. For these reasons, we generate synonym vectors of both languages from context vectors, and weights of the seed dictionary can be learned and bilingual lexicons are newly generated by the modified single-layer Perceptron.

The successful development of the bilingual lexicon extraction system would provide enormous contribution to the related research community. In this thesis, we developed the novel system to build bilingual lexicons.

The remainder of this thesis is organized as follows. Chapter 2 presents related works for bilingual lexicon extraction that are linguistic resources, vector space model, Perceptron and evaluation metrics. Chapter 3 describes overall system architecture of our method. Chapter 4 discusses a part of our method for building an initial seed dictionary using a context-based approach. Chapter 5 presents the other part of our method to extract bilingual lexicon using an iterative approach. Finally, Chapter 6 discusses our conclusions.

Chapter 2

Literature Review

This chapter presents background about the techniques used in the following chapters. The linguistic resources such as parallel corpora and comparable corpora will be introduced. The previous works on the vector space model will be reviewed since it is a basic idea in this thesis. Moreover, we present a single-layer Perceptron that is used in a weight learning task. Finally, we present evaluation metrics for bilingual lexicon extraction.

2.1 Linguistic resources: The text corpora

In natural language processing, text corpora are essential linguistic resources for data-driven approaches. Generally, the text corpora are huge and structured set of texts. In the bilingual lexicon extraction, the text corpora are used to provide vocabularies and contextual and statistical information. The text corpora are divided into two types: Parallel and Comparable corpora.

Parallel corpora which contain source texts and their translations as target texts are easy to extract bilingual lexicons than the other (comparable corpora discussed later). For example, the European Parliament Proceedings (Europarl) is one of a freely available parallel corpus. It includes 21 European languages: Romanic, Germanic, Slavic, Finni-Ugric, Baltic, Greek and so on. However, the large amount of parallel corpus is hard to collect and barely available to well-known languages with fewer resources or in narrow domains. But no missing translations are in the target texts (Fung, 1998). Therefore, bilingual lexicon extraction task is more simple than comparable corpora.

Therefore, a comparable corpus is a pair of corpora in two different languages which related to certain characteristics such as event, domain, topic, date or subject. Unlike the parallel corpora, the comparable corpora in which translations might not exist in the target texts (Fung, 1998) are more difficult to extract bilingual lexicons.

The types of corpora can be divided into four types according to the comparability of the texts. Skadina et al. (2010) devided the comparable corpora into four types as follows:

- Parallel texts
- Strongly comparable texts
- Weakly comparable texts
- Non-comparable texts

The parallel texts are a pair of texts that are source texts and their accurate translations or approximate translations. The strongly comparable texts are related texts which are reporting the same event or describing the same subject. The third category is weakly comparable texts which contains texts in the same narrow subject, domain and genre, but describing different events and dates. Finally, the non-comparable texts are pairs of texts that are randomly selected from a pair of very large collections of text in two different languages regardless of domain, subject, event or genre.

In this thesis, we use the comparable corpora that were collected from the news articles and the Europarl. The comparable corpora have similar proportions between comparability levels (50% weakly comparable texts, 50% non-comparable texts) and details of corpora statistics are presented in Section 5.2.

2.2 A vector space model

In this chapter, we review related work on bilingual lexicon extraction using vector space model. The vector space model is the widely used model for bilingual lexicon extraction.

1) Basic concepts of vector space model

In vector space model, source and target words are represented as points in vector space. The dimension of the space can be determined according to the number of context words of a target language. A source word is represented by a vector with its contextual words and a target word is represented in the same way. Here, the contextual words are weighted by their degree of association. However, the source vectors and the target vectors cannot be represented into the same space since they are made up of different languages. For these reasons, the source vectors have to be translated in the target language using an initial seed dictionary. In this translation process, the volume of the initial seed dictionary is very important. The larger volume of the initial seed dictionary is more helpful to represent the source vectors accurate into the target vector space. Therefore, the source word can be represented into the target vector space and then the source word can be compared with the target vectors in the target space.

To compare with the target vectors, Cosine similarity (Salton and McGill, 1983) is commonly used as a similarity measure in bilingual lexicon extraction. The closest target word vector to the translated source word vector can be extracted as a translation candidate pair.

2) Association measures

An association value is a degree of relationship between two measured quantities. In bilingual lexicon extraction, the association measure is used to weight a word in which is co-occurs with certain words. It indicates how the word is associated with certain words. There are some examples of the association measure:

- Term Frequency (TF) ($tf_{t,d}$) is the number of times that a term *t* occurs in document *d*. In bilingual lexicon extraction, Fung (1998) used TF by collecting contextual term *i* in the context of *j* and count their occurrence frequency:

  $$tf_{i,j} = freq(i,j) \quad (2.1)$$

  where $freq(i,j)$ is the co-occurrence frequency between i and j.

- Inverse Document Frequency (IDF) is used to estimate the rarity of a term *t* in the whole document collection. If *t* occurs in all documents of the collection, its IDF is zero. In bilingual lexicon extraction, the idf_i is given as follows:

  $$idf_i = log\frac{freq_{max}}{freq(i)} + 1 \quad (2.2)$$

  where $freq_{max}$is the maximum frequency of any word in the corpus and $freq(i)$ is the total number of occurrences of word i in the corpus.

- Term Frequency-Inverse Document Frequency (TF-IDF) is to reflect how important a word is to a document in a corpus, Term Frequency-Inverse Document Frequency is used as an association metric. The TF-IDF of a term is the product of its TF weight and its IDF weight. The TF-IDF is denoted as follows:

  $$tf\text{-}idf = tf_{ij} \cdot idf_i \quad (2.3)$$

Likewise, there are other association measures such as pointwise mutual information (PMI) (Church and Hanks, 1990) and log-likelihood ratio (LLR) (Huelsenbeck et al., 1996) and chi-square (χ^2) (Plackett, 1983).

3) Similarity measures

A similarity measure can represent the similarity between two documents, two queries, or two objects. In bilingual lexicon extraction, the similarity can be obtained by measuring distance between words in vector space. The followings are some of the examples of the similarity measure:

- Inner product

  A basic similarity measure is inner product. Also called dot product or scalar product. The similarity obtained by multiplying the corresponding coordinates of each of two vectors and summing up the products. The similarity value is not bounded. Given two *N* dimensions vector $\vec{x}$ and $\vec{y}$, the inner product given as follows:

$$Inner(\vec{x},\vec{y}) = \sum_{i=0}^{N} x_i y_i \quad (2.4)$$

- Cosine similarity

  According to Li (2013) the Cosine similarity measures the angle between two vectors and is the most popular similarity measure. The cosine similarity performs the inner product of the vectors and then divides the product by their norms (Ismail, 2012). The similarity value is bounded between 0 and 1.Given two *N* dimension vectors $\vec{x}$ and $\vec{y}$, the cosine similarity between them is calculated as follows:

$$Cosine(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{N} x_i \times y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \quad (2.5)$$

4) A Standard context-based approach

Rapp (1995) and Fung (1998) proposed the standard context-based approach using a vector space model. According to Hazem et al., (2011), the implementation of the context-based approach can be summarized as follows:

i. Context characterization

All the words in the context of each word i are collected, and their frequency in a window of n words around i extracted. For each word i of the source and the target languages, we obtain a context vector $\mathbf{i}$ where its entry is $\mathbf{i}_j$ and the $\mathbf{i}_j$ is determined by association measures.

ii. Vector translation

The entry $\mathbf{i}_j$ of context vector $\mathbf{i}$ are translated using an initial seed dictionary. The entry $\mathbf{i}_j$ with no translation in the seed dictionary are discarded.

iii. Target vector matching

For the similarity measure, $\mathrm{sim}(\bar{\mathbf{i}}, \mathrm{t})$ is used to score between the translated source context vector $\bar{\mathbf{i}}$ and the target context vector. The most commonly used similarity measure is the cosine similarity.

iv. Candidate translation

The translation candidates of a source context vector $\mathbf{i}$ are the target context vectors ranked by the similarity score.

5) Previous works

The direct way of bilingual lexicon extraction is to align words from parallel corpora (Wu and Xia, 1994). Bitext word alignment is an important role for most methods of statistical machine translation. Automatic word alignment is typically done by choosing that alignment which best fits the statistical machine translation model. Brown et al., (1993) proposed IBM statistical machine translation system for the word-level translation model. According to him, the word-level translation model works as follows:

Translation probability:

$$p(\mathbf{e}, \mathrm{a}|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) \quad (2.6)$$

where for a foreign sentence $\mathbf{f} = \left(f_1, \ldots, f_{l_f}\right)$ of length l_f

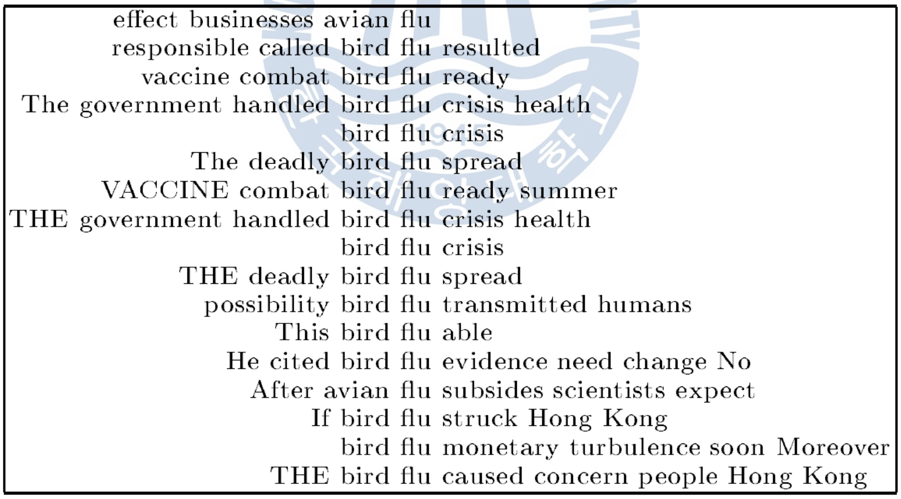
for a English sentence $\mathbf{e} = \left(e_1, \ldots, e_{l_e}\right)$ of length l_e

An alignment of each English word e_j to a foreign word f_i: alignment function $a: j \rightarrow i$

ϵ is a normalize constant

The translation probability is learned by Expectation-Maximization (EM) algorithm. However, this approach requires huge volumes of parallel corpora. Especially, for less-known languages, parallel corpora is not easy to collect and hard to find. Also, they are restricted to specific domains. Therefore, researchers have studied a method of using pivot language as an intermediary language to extract bilingual lexicons (Tanaka and Ummemura, 1994; Wu and Wang, 2007; Tsunakawa et al., 2008; Kwon et al, 2013). The method by Kwon et al., (2013) proposed a simple and effective method to extract bilingual lexicon between two less-known language pairs using a pivot language and word alignment tool using Information Retrieval (IR) technique. This method used two pair of parallel corpora that are source-pivot language pair (L_s-L_p) and pivot-target language pair (L_p-L_t). The pivot language is used for representing both of context vectors of a source language and target language. In our proposed method use this model for generating an initial seed dictionary and the process is described in more detail in the section 3.1. However for some well-known language pairs, the size of the parallel corpora is poor and domain-restricted as well. For these reasons, many researchers have focused on comparable corpora (Fung, 1995; Rapp, 1995; Yu and Tsujii, 2009; Ismail and Manandhar, 2010). Fung (1998) suggested an IR based approach from comparable corpora for bilingual lexicon extraction. It represents a word into context words vector on both source and target language and relies on the simple assumption. The assumption is a word and its translation tends to appear in the similar context. This assumption is based on the first-order affinities: *What other words are likely to be found in the immediate vicinity of a given word* (Grefenstette, 1994). Table 2.1 shows an example of contexts of flu. According to Fung (1998), the IR based approach is implemented as follows:

1. Construct context vectors of all unknown words $\boldsymbol{s}$ in the source language
2. Construct context vectors of all candidate translation words $\boldsymbol{t}$ in the target language
3. For all $\boldsymbol{s}$ and $\boldsymbol{t}$, compute similarity$(\boldsymbol{s},\boldsymbol{t})$.
4. Rank the output according to this similarity score.
5. Choose the N highest ranking $\boldsymbol{t}$ as translation candidate for $\boldsymbol{s}$
6. Choose the M highest ranking $(\boldsymbol{s},\boldsymbol{t})$ as new lexicon entries for the bilingual dictionary



```
        effect businesses avian flu
        responsible called bird flu resulted
           vaccine combat bird flu ready
   The government handled bird flu crisis health
                          bird flu crisis
               The deadly bird flu spread
           VACCINE combat bird flu ready summer
   THE government handled bird flu crisis health
                          bird flu crisis
               THE deadly bird flu spread
              possibility bird flu transmitted humans
                     This bird flu able
                 He cited bird flu evidence need change No
                  After avian flu subsides scientists expect
                       If bird flu struck Hong Kong
                          bird flu monetary turbulence soon Moreover
                      THE bird flu caused concern people Hong Kong
```

Figure 2.1: An example of contexts of *flu* in English newspaper articles

Source: Fung (1998)

To build context vectors of all unknown words, Fung used Term Frequency (TF) and Inverse Document Frequency (IDF). TF is co-occurrence frequency between context word and unknown word. IDF is used to emphasize the significance of common words. According to Fung, IDF accounts for the overall occurrence frequency of context words in the entire corpus:

$$IDF = log\frac{n_{max}}{n_i} + 1$$

where n_{max} is the maximum frequency of any word in the corpus
n_i = the total number of occurrences of word *i* in the corpus

For the similarity measure between source and target context vectors, Fung use the cosine similarity which is the most common similarity measure in the IR community:

$$sim(W_s, W_t) = \frac{\sum_{i=1}^{d}(w_{si} \cdot w_{ti})}{\sqrt{\sum_{i=1}^{d} w_{si}^2 \times \sum_{i=1}^{d} w_{ti}^2}} \quad (2.7)$$

where W_s is a source context vector
W_t is a target context vector
w_{si} is $TF_{si} \times IDF_i$
w_{ti} is $TF_{ti} \times IDF_i$

Furthermore, Fung (1998) used a measure that reflects reliability of the initial seed lexicon known as Confidence Weighting. If a word i_s is the *k*-th candidate for word i_t, then $w_{i_{ds}} = w_{i_{ds}}/k_i$.

The similarity measure then becomes:

$$sim'(W_s, W_t) = \frac{\sum_{i=1}^{d}(w_{si} \cdot w_{ti})/k_i}{\sqrt{\sum_{i=1}^{d} w_{si}^2 \times \sum_{i=1}^{d} w_{ti}^2}} \quad (2.8)$$

Fung evaluated the method on a comparable corpora consisting of various English and Chinese newspaper articles. She tested on English to Chinese translation. The experimental results show that the translation accuracy is 30% at the top 1 and 76% at the top 20.

Another recent method using comparable corpora is that extracts new translation candidates and then uses the translation candidates as a new initial seed dictionary. This method is called iterative approach, which has been presented in Chatterjee et al., (2010) and Chu et al., (2014). A work by Chu et al., (2014) proposed a bilingual lexicon extraction system that uses topical and contextual knowledge in the iterative process. The system consisted of two main methods, namely topic model based method (TMBM) and context based method (CBM). The TMBM measures the similarity of two words on cross-lingual topical distributions, while CBM measures the similarity on contextual distributions across languages. In their study, exploiting both topical and contextual knowledge can make bilingual extraction more reliable and accurate than only using one knowledge source. The summarization of the system is as follows:

- TMBM can extract bilingual lexicons from comparable corpora without any prior knowledge. The extracted lexicons are semantically related and provide useful contextual information in the target language for the source word. Therefore, it is appropriate to use the lexicons extracted by TMBM as an initial seed dictionary, which is an input of CBM.

- The lexicons extracted by CBM can be combined with the seed dictionary to further improve the accuracy.

- The combined lexicons again can be used as a new seed dictionary for CBM. Therefore the accuracy of the lexicons can be improved iteratively.

Chu et al. conducted on Chinese-English and Japanese-English Wikipedia data. The experimental results show that Chu et al.'s method can significantly improve the performance in the first few epochs.

2.3 Neural networks: The single layer Perceptron

Perceptron is a type of artificial neural networks. It was introduced by Rosenblatt (1958) for binary classification. The Perceptron is an online learning algorithm that supervised classification, trains the weight of a linear function so that it makes no error on the training example. It can be proved that if the training example is linearly separable, then the Perceptron learning algorithm will converge to a certain value. A structure of Perceptron is depicted in Figure 2.1. It takes a vector of real-value inputs, calculates a linear combination of these inputs, and then outputs 1 if the result is greater than certain threshold or -1 otherwise. The output o$(x_1, \dots, x_n)$ given input x_1 through x_n is defined as follows:

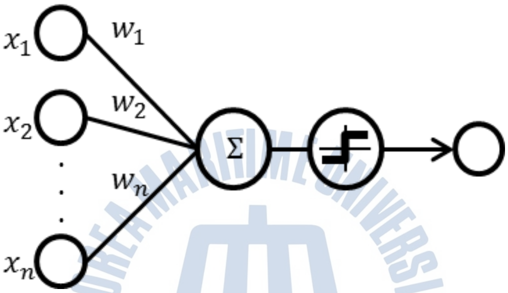


Figure 2.2: The structure of Perceptron

$$\mathrm{o}(x_1, \dots, x_n) = f(x) = \begin{cases} 1, & \text{if} \sum_{i=0}^{n} w_i x_i > 0 \\ -1, & \text{otherwise} \end{cases} \qquad (2.9)$$

where each w_i is a real-value constant weight that reflects the importance of input x_i to the Perceptron output.

The initial weights are assigned randomly. w_i is updated when misclassification occurs in each training example. The weight update is defined as follows:

$$w_i \leftarrow w_i + \Delta w_i \quad (2.10)$$

where Δw_i is $\alpha(t - o)x_i$

t is the desired output of the current training example, o is the output generated by the Perceptron, and α is a learning rate. The learning rate is bounded between 0 and 1. This process is repeated until the training example is classified correctly, or reaches the maximum iteration defined by users.

2.4 Evaluation metrics

To evaluate system performance, in bilingual lexicon extraction, the accuracy, the recall and the MRR (Mean Reciprocal Rank) (Voorhees, 1999) are commonly used as evaluation metrics similar to evaluation in information retrieval.

The accuracy is the fraction of its translation candidates that are correct. The accuracy is given as follows:

$$Acc_k = \frac{1}{N}\sum_{i=1}^{N} max_{1\le j\le k} a_{ij} \quad (2.11)$$

$$\text{where } a_{ij} = \begin{cases} 1, & \text{if } t_{ij} \in A_i \\ 0, & \text{otherwise} \end{cases}$$

where N is the number of evaluation words

A_i is a set of the i-th translation

t_{ij} is the j-th translation candidate for i-th evaluation word

a$_{ij}$ is correct translation candidates

k is evaluation ranking that means accuracy at the top k

The recall is the ratio of the suggested translation candidates that agree with the marked answer to the total number of translations in the evaluation words. The recall is calculated as follows:

$$REC_k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|A_i|}\sum_{j=1}^{k} a_{ij} \quad (2.12)$$

$$\text{where } a_{ij} = \begin{cases} 1, & \text{if } t_{ij} \in A_i \\ 0, & \text{otherwise} \end{cases}$$

The MRR is the average of the reciprocal ranks of translation candidates that are correct translations for a sample of evaluation words. The MRR is given as follows:

$$MRR_k = \frac{1}{N}\sum_{i=1}^{N} max_{1\leq j\leq k} r_{ij} \qquad (2.13)$$

$$\text{where } r_{ij} = \begin{cases} \frac{1}{j}, & \text{if } r_{ij} \in A_i \\ 0, & \text{otherwise} \end{cases}$$

Chapter 3

System Architecture of
Bilingual Lexicon Extraction System

In Chapter 3, we present a novel bilingual lexicon extraction system. The system is based on the vector space model for word representation and the performance becomes better using the Perceptron algorithm. In this chapter, an overall structure of the proposed system will be discussed.

3.1 Required linguistic resources

The proposed system requires three linguistic resources: parallel/comparable corpora and an initial seed dictionary. According to Ismail (2012), the corpora are employed in bilingual lexicon extraction to provide lexical and statistical information as follows:

- List of vocabularies: these include the source and target words
- Contextual information: the co-occurrence frequency of the words surrounding a certain word. The translations for the source word seem to have similar contextual information, i.e., context words co-occurring frequently with the source word should have translations that co-occur frequently with the translations of the source context word.

We collect these parallel/comparable corpora on Korean, Spanish and French from news articles and Europarl parallel corpora. The detailed descriptions about the statistical information of these corpora statistics are discussed in Sub Section 4.2 and 5.2.

The other linguistic resource is the initial seed dictionary. As mentioned before, the initial seed dictionary plays an important role to translate a source context vector to a target context vector. In other words, the source context vector is represented to a target vector space using the initial seed dictionary. Therefore, the bigger volume of the initial seed dictionary is more helpful to represent the source context vector to the target language. In general, the initial seed dictionary is constructed manually by human. In this thesis, we generated the initial seed dictionary automatically using a context-based approach. The context-based approach will be presented in Section 4.

3.2 System architecture

An overall structure of the proposed method is depicted in Figure 3.1. Our methods consist of two procedures: a context-based approach (CBA) and an iterative approach (IA). We first construct source context vectors from Source/Pivot parallel corpus and target context vectors from Pivot/Target parallel corpus respectively. After that we exploit the CBA to construct an initial seed dictionary from the two context vectors. The initial seed dictionary is used to translate a source synonym vector to a target synonym vector and used as weights for a modified Perceptron algorithm in the IA, and then we construct source synonym vectors and target synonym vector from each source and target comparable corpus respectively. Finally, we apply the IA to obtain bilingual lexicons. Details of the CBA and the IA will be described in Section 4.1 and 5.1 respectively.

The proposed method has three advantages:

- Does not require a large size of an initial seed dictionary. The initial seed dictionary is generated automatically by the CBA and can be revised gradually by the modified Perceptron algorithm described in Section 4.1.

- Does not need labels of training examples that are inputs of the modified Perceptron algorithm. The modified Perceptron algorithm dynamically generates the labels of the training examples during epochs.
- System performances can be improved during epochs.

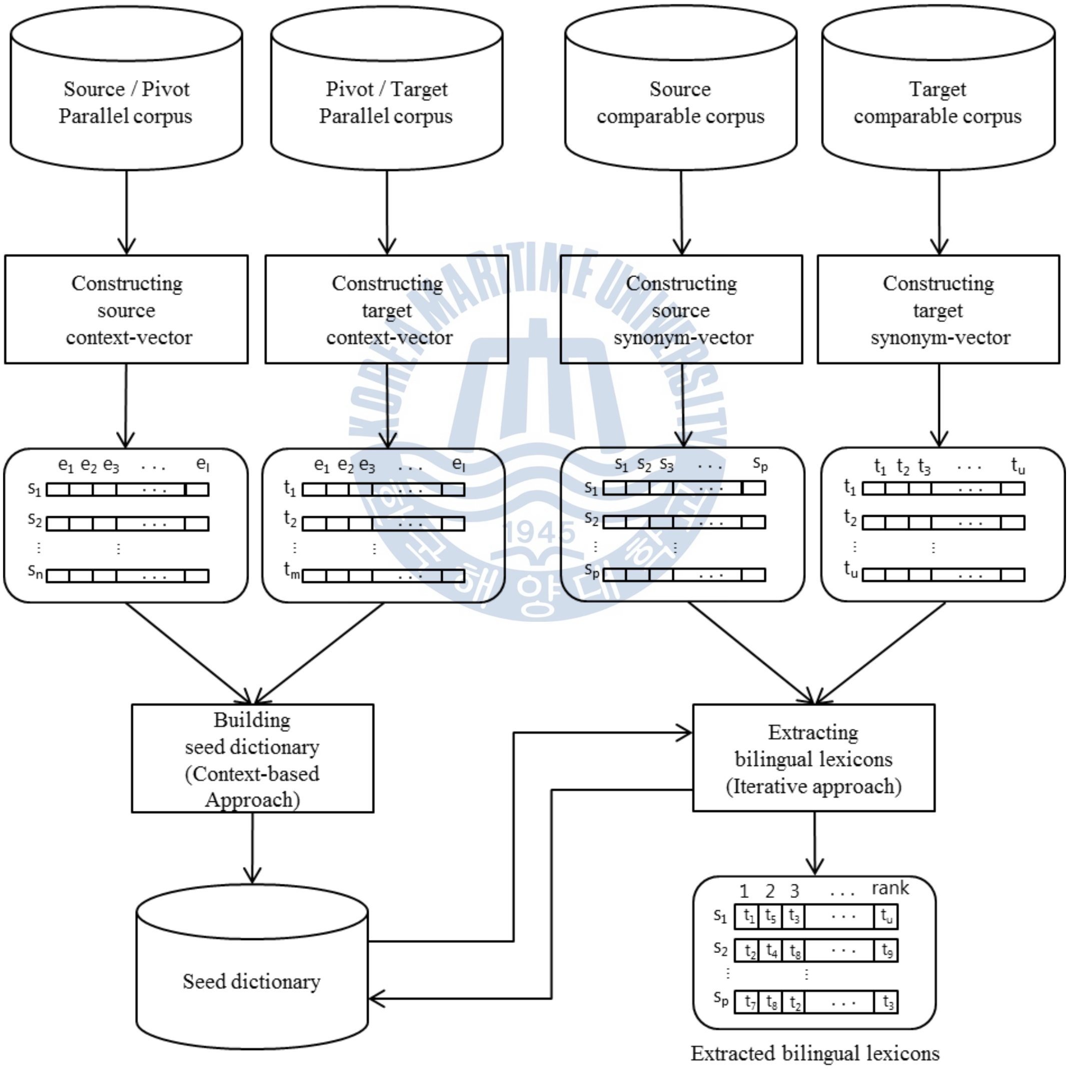


Figure 3.1: An overall structure of the proposed method

Chapter 4

Building a Seed Dictionary

In Chapter 4, we discuss a context-based technique that aimed to construct an initial seed dictionary. The initial seed dictionary is used as inputs for iterative approach and employed to translate source language into target language. To do this, we use the our work which uses a pivot language and IR techniques for calculating similarities between source context vectors and target context vectors represented by the pivot language. The initial seed dictionary is constructed on two different language pairs that are unidirectional Korean-Spanish and Korean-French respectively, and accuracies of the initial seed dictionary based on this approach for the high-frequent words achieved at least 48.5% and up to 88.5% within the top 20 ranking candidates. The low-frequent words achieved at least 50.5% and up to 70% at the top 20 rank.

4.1 Methodology: Context Based Approach (CBA)

The CBA uses parallel corpora with more accurate alignment information instead of comparable corpora. It, however, is difficult to obtain parallel corpora for less-known language pairs. For such reasons, we use a pivot language which is well-known like English. The pivot language is used for representing both of source context vectors and target context vectors. Unlike the previous studies using comparable corpora, therefore, we exploit the two parallel corpora through the pivot language (e.g., English) like Korean-English (KR-EN) and English-Spanish (EN-ES) and use the IR techniques for calculating the similarity between the source context vectors and the target context vectors represented by the pivot language.

In the previous works, source context vectors are required to translate into target language using the initial seed dictionary, but the CBA is not needed anymore. Therefore, any bilingual dictionaries are not expected. Besides, we use a freely available word aligner, called Anymalign, to construct context vectors. Anymalign showed high accuracy for low-frequent words to extract translation candidates (Lardilleux et al., 2011). Figure 4.1 shows an overall structure of the CBA. The CBA can be summarized in the following three steps:

(1) To build source context vector and target source context vector for each word in the source language (e.g., KR) and the 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.

(2) To calculate the similarity between the source context vectors and the target context vectors, we use the cosine measure.

(3) To sort the top k word pairs based on their similarity scores.

Two parallel corpora share a pivot language, English, in our case, and are used to build context vectors because Korean-Spanish bilingual corpora are publicly unavailable. The example of the context vector is depicted in Table 4.1. Anymalign is used to weight all contextual words in the context vectors. The partial output of Anymalign is shown in Table 4.2. We do not use all alignments of Anymalign as contextual word. Instead, we adapt an equation to select informative alignments. The equation is as follows:

$$f(w) = \begin{cases} \dfrac{Pr(s|t) + Pr(t|s)}{2} & \text{if} \quad |Pr(s|t) - Pr(t|s)| \leq \theta_1 \text{ and} \\ & \quad Pr(s|t) \geq \theta_2 \text{ and } Pr(t|s) \geq \theta_2 \\ 0 & \quad \text{otherwise} \end{cases} \tag{4.1}$$

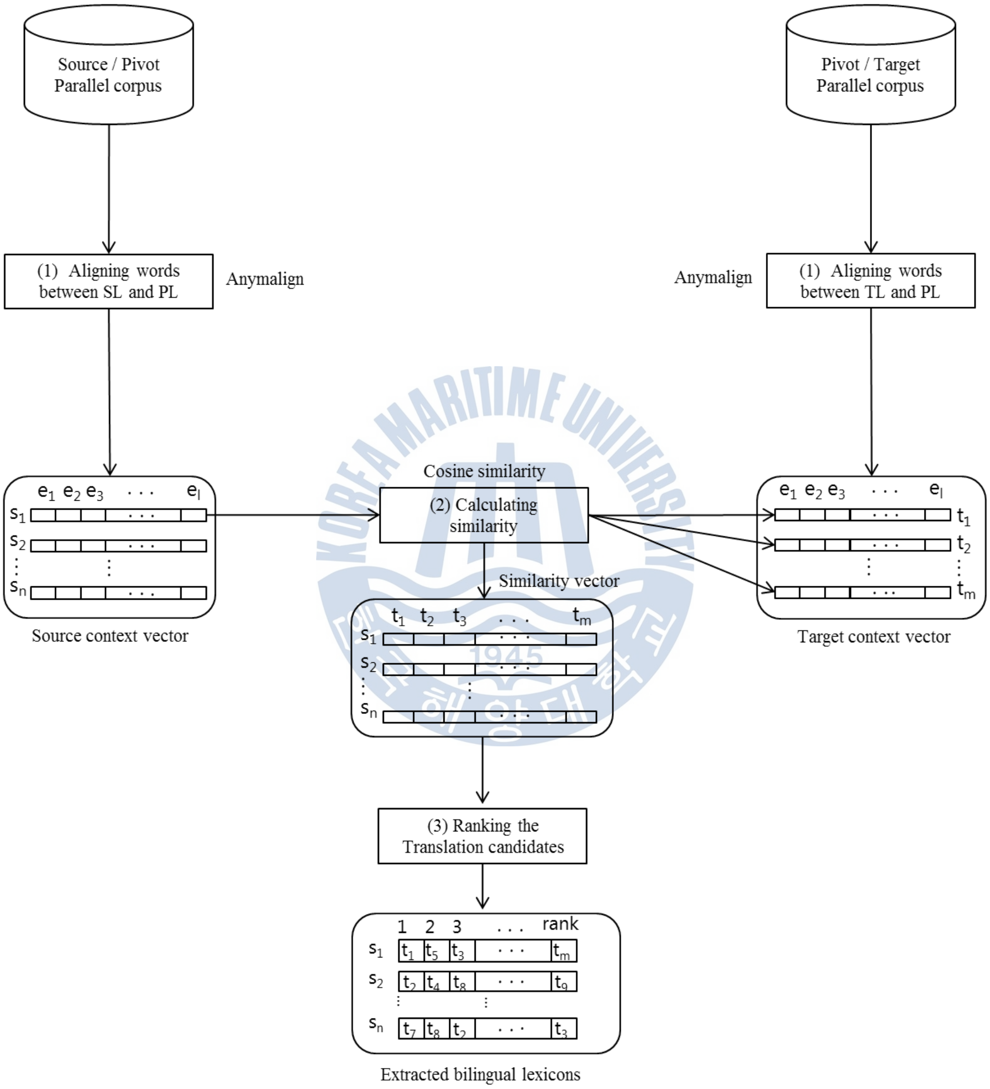


Figure 4.1: An overall structure of the context-based approach (CBA)

Equation 4.1 has two conditions: The first is that the subtraction of the bi-directional translation probabilities $(Pr(s|t), Pr(t|s))$ is below the certain threshold θ_1 and the other is the bi-directional translation probabilities are above the certain threshold θ_2. The alignment satisfying these conditions is an informative alignment. For example, the *Pr*(경찰|scenario) and *Pr*(scenario|경찰) in table 4.2 show a big difference of translation probability. They represent that English word “scenario” may not be a translation of Korean word “경찰”. Besides, the *Pr*(정부|balance) and *Pr*(balance|정부) in Table 4.2 show low values of translation probabilities on both bi-directional translation probabilities. It reflects an intuition that English word “balance” is not a translation of Korean word “정부”. Therefore, if these two conditions are satisfied, the weight of contextual word is determined by averaging the bi-directional translation probabilities. The reason for this is to refine the context vector. This refinement produced better performances (Kwon et al., 2014). We used the θ_1 of 0.005 and θ_2 of 0.5 and 0.003 for each high and low frequent words extraction respectively.

Table 4.1: Examples of the context vector

source word	vector attribute (contextual word)				
교육 (education)	training	school	student	…	boy
	0.837	0.025	0.017	…	0.011
개발 (development)	development	developer	technology	…	company
	0.496	0.061	0.025	…	0.024
국민 (people)	people	public	citizen	…	national
	0.273	0.126	0.081	…	0.036

As mentioned before, in the previous work, a seed dictionary is required to translate context vectors at this time, but we do not carry out them. After context vectors are built once, all source and target context vectors are compared each other to get its similarity by using the cosine measure. Finally, the top k word pairs can be extracted as a result.

Table 4.2: Partial results of Anymalign

Word (KR)	Word (EN)	Lexical weight (KR)	Lexical weight (EN)	**Translation probability p(KR|EN)**	**Translation probability p(EN|KR)**	Absolute frequency
맨체스터 (manchester)	manchester	0.760	0.540	**0.958**	**0.634**	55
정부 (government)	balance	0.175	0.683	**0.007**	**0.008**	84
유통업체 (distributor)	distributor	0.485	0.266	**0.748**	**0.487**	5534
비전 (vision)	long-term	0.088	0.034	**0.008**	**0.003**	73
비전 (vision)	vision	0.377	0.492	**0.810**	**0.866**	3457
경찰 (police)	scenario	0.256	0.335	**0.784**	**0.004**	15

4.2 Experiments and results

In this thesis, we constructed two initial seed dictionaries from two different language pairs that are KR-ES and KR-FR.

4.2.1 Experimental setups

1) Parallel corpora

The statistics of used parallel corpora are described in Table 4.3. We used the KR-EN parallel corpora compiled by Seo et al. (2006) (433,151 sentence pairs), and two sets of sub-corpora (500,000 sentence pairs each) that are randomly selected from ES-EN and FR-EN from the Europarl parallel corpus (Koehn, 2005). The number of words in ES-EN and FR-EN parallel corpora is nearly similar, but the number of KR words (called eojeol in Korean) in KR-EN parallel corpus is lower than that of EN words. In fact, KR words are a little bit different from EN words and others.

Korean words consist of one morpheme or more. Therefore, the number of KR words can be similar to that of EN words if morphemes instead of words are counted.

Table 4.3: Statistics of parallel corpora

	KR-EN		**ES-EN**		**FR-EN**	
Number of sentence pairs	433,151		500,000		500,000	
Average number of words per sentence	**KR**	**EN**	**ES**	**EN**	**FR**	**EN**
	31	19.2	26.4	25.4	29.7	27.1
Number of word types	74,614	93,490	36,403	31,197	21,894	30,196
Domain	News article		Proceedings of the European Parliament			

2) Data pre-processing

All words are tokenized by the following tools: U-tagger[1] (Shin et al., 2012) for Korean, Tree-Tagger[2] (Schmid, 1994) for English, Spanish and French. In case of Korean, Multiword expressions which are composed by more than 4 characters are decomposed by U-tagger. For example, Korean word “인공지능(artificial intelligence)” is decomposed into “인공(intelligence)” and “지능(intelligence)” because the proposed system is targeted towards extracting single words. All words in English, Spanish, and French are converted to lower case, and those in Korean are morphologically analyzed into morphemes and POS-tagged by the U-tagger. After pre-processing, all words have been removed from the corpus, except for nouns. For that reason, when aligning the words, Anymalign do not consider

[1] http://nlplab.ulsan.ac.kr/

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

contextual information, but exploit random sampling and string difference (Lardilleux et al., 2010). Therefore, words that are not noun, can act as noise when aligning the words.

3) Building evaluation dictionary

To evaluate the performance of CBA, we build two sets of bilingual lexicons (KR-ES and KR-FR) 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 200 high frequent words (denoted by HIGH hereafter) and 200 low rare words (denoted by LOW hereafter), respectively. Table 4.4 shows the average number of the translations per source word in each lexicon. The number means the degree of ambiguity and is the same as the number of polysemous words.

Table 4.4: The average number of the translations
per source word in the evaluation dictionaries for CBA

Evaluation dictionary	HIGH	LOW
KR-ES	10.3	5.4
KR-FR	8.4	6.8

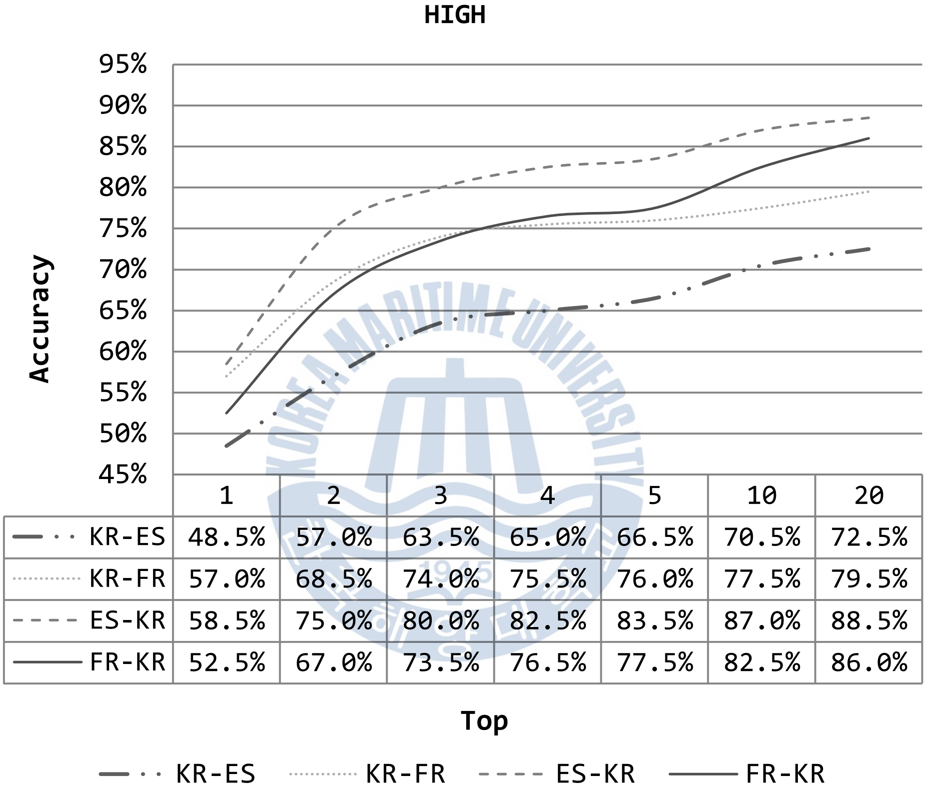
4) Evaluation metrics

We evaluate the quality of translation candidates extracted by the proposed systems. Similar to the evaluation in information retrieval, the accuracy, the recall, and the mean reciprocal rank (MRR) are used as evaluation metrics.

[3] http://dic.naver.com

4.2.2 Experimental results

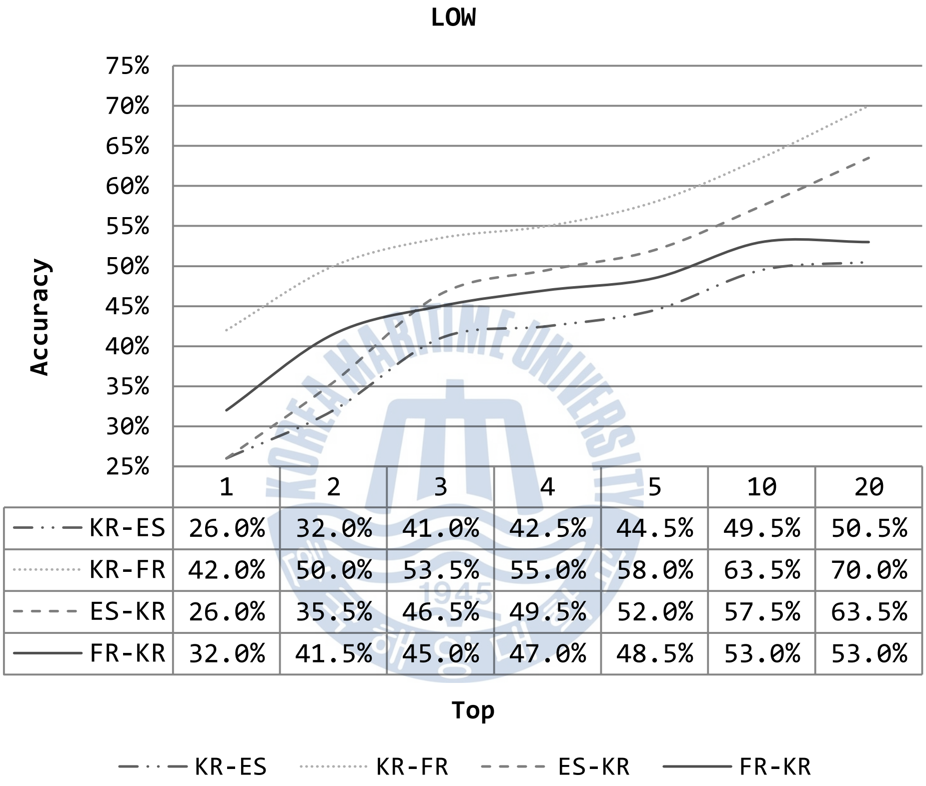
1) Accuracy



	1	2	3	4	5	10	20
KR-ES	48.5%	57.0%	63.5%	65.0%	66.5%	70.5%	72.5%
KR-FR	57.0%	68.5%	74.0%	75.5%	76.0%	77.5%	79.5%
ES-KR	58.5%	75.0%	80.0%	82.5%	83.5%	87.0%	88.5%
FR-KR	52.5%	67.0%	73.5%	76.5%	77.5%	82.5%	86.0%

Figure 4.2: Accuracies of the CBA for HIGH words

The accuracies of the HIGH words are shown in Figure 4.2. As seen in Figure 4.2, the experimental result has demonstrated that the CBA for HIGH words shows the accuracies ranging from 72.5% (KR-ES) to 88.5% (ES-KR) within the top 20.

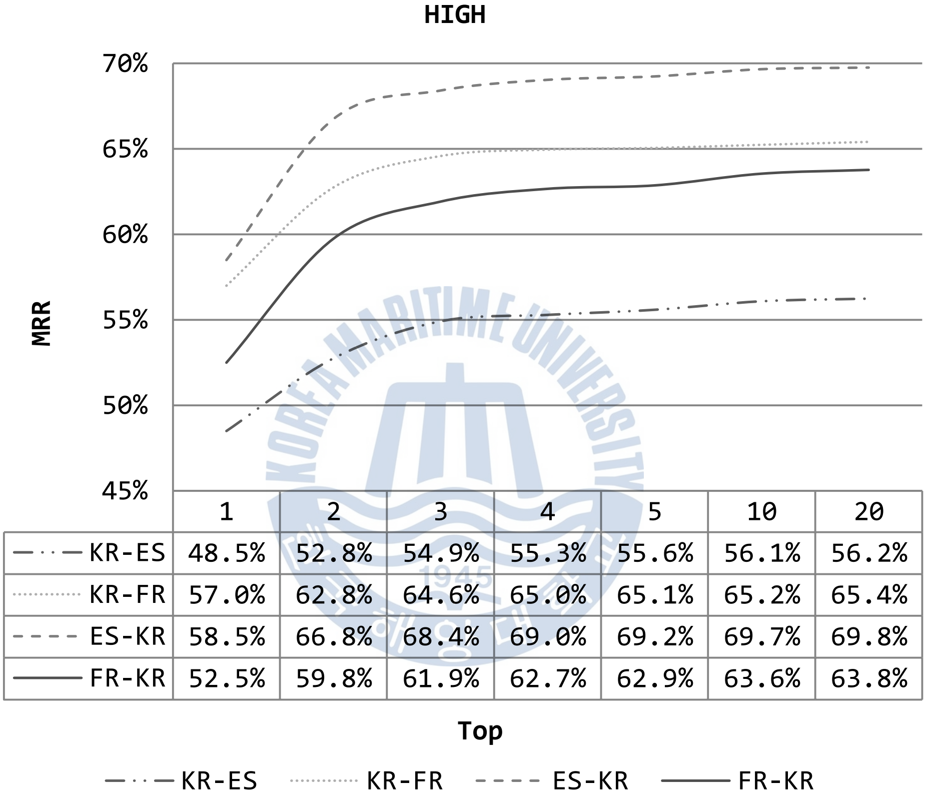


	1	2	3	4	5	10	20
—·— KR-ES	26.0%	32.0%	41.0%	42.5%	44.5%	49.5%	50.5%
········ KR-FR	42.0%	50.0%	53.5%	55.0%	58.0%	63.5%	70.0%
- - - - ES-KR	26.0%	35.5%	46.5%	49.5%	52.0%	57.5%	63.5%
——— FR-KR	32.0%	41.5%	45.0%	47.0%	48.5%	53.0%	53.0%

Figure 4.3: Accuracies of the CBA for LOW words

The accuracies of the LOW words are presented in Figure 4.3. The graph shows that the CBA for LOW words shows the accuracies ranging from 50.5% (KR-ES) to 70.0% (KR-FR) within the top 20.

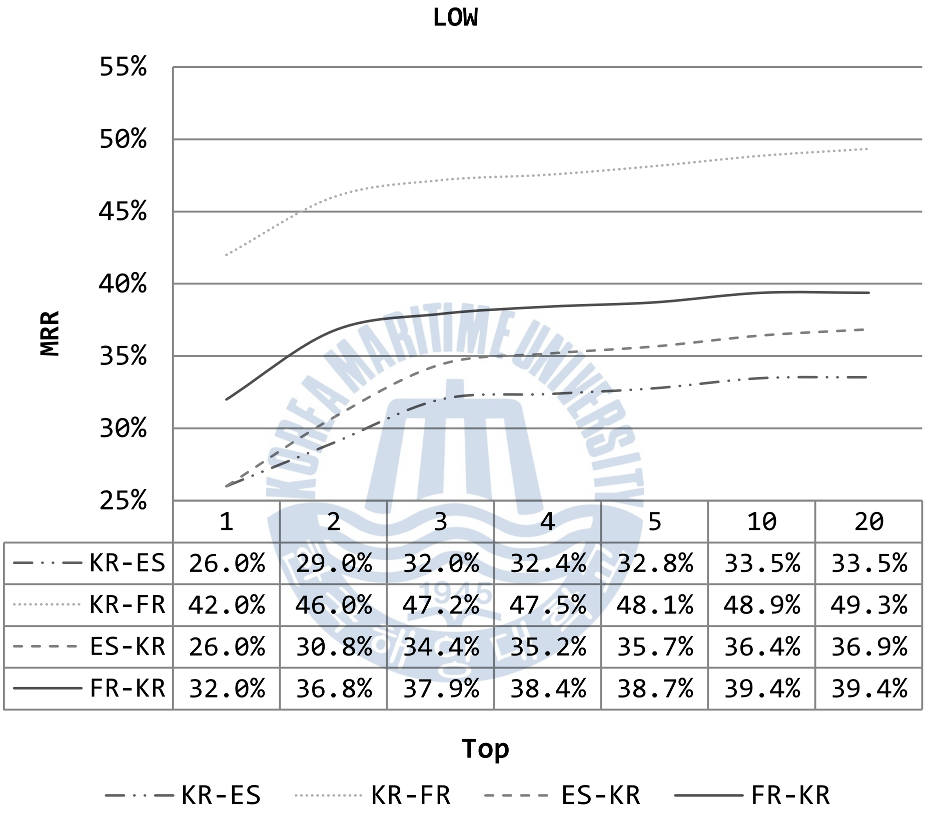
2) MRR



	1	2	3	4	5	10	20
— · — KR-ES	48.5%	52.8%	54.9%	55.3%	55.6%	56.1%	56.2%
········ KR-FR	57.0%	62.8%	64.6%	65.0%	65.1%	65.2%	65.4%
- - - - ES-KR	58.5%	66.8%	68.4%	69.0%	69.2%	69.7%	69.8%
—— FR-KR	52.5%	59.8%	61.9%	62.7%	62.9%	63.6%	63.8%

Figure 4.4: MRRs of the CBA for HIGH words

The MRR results of the CBA are shown in Figure 4.4. As shown in Figure 4.4, the MRR of the HIGH words is rapidly increased until the top 2, after that the MRR is steadily increased. This means that correct translation candidates tend to appear within the top 2.



	1	2	3	4	5	10	20
KR-ES	26.0%	29.0%	32.0%	32.4%	32.8%	33.5%	33.5%
KR-FR	42.0%	46.0%	47.2%	47.5%	48.1%	48.9%	49.3%
ES-KR	26.0%	30.8%	34.4%	35.2%	35.7%	36.4%	36.9%
FR-KR	32.0%	36.8%	37.9%	38.4%	38.7%	39.4%	39.4%

Figure 4.5: MRRs of the CBA for LOW words

In the same experiments of the CBA for LOW words are represented in Figure 4.5. The MRR of the LOW words is rapidly increased near at the top 2 and 3. Therefore, the correct translation candidates tend to be discovered within the top 2 and top 3.

3) Recall

Table 4.5: Recalls of the CBA for HIGH and LOW words at the top 20.

Language pairs	Top 20 Recall	
	HIGH	LOW
KR-ES	17.0%	15.0%
KR-FR	22.5%	18.3%

Lastly, the recalls of the HIGH and LOW words are shown in Table 4.5. As seen in the table, the best recall is 22.5% for the KR-FR for HIGH words. One of reasons for low recall can be why words usually have on sense per corpus in parallel corpus (Fung, 1998). Another reason can be why words do not belong to various domains and our data sets only come from European Parliament proceedings and news articles.

4.3 Discussions

We have presented an IR based approach for extracting bilingual lexicons from parallel corpus via pivot languages. The CBA overcomes some of the problems of previous works that need an initial seed dictionary and use comparable corpora instead of parallel corpora in terms of lack of linguistic resources by using the pivot approach. In this thesis, the CBA is exploited for generating initial seed dictionaries. The two pairs of initial seed dictionary (KR-ES and KR-FR) are used for inputs to the iterative approach.

Chapter 5

Extracting Bilingual Lexicons

In the previous chapter, we described a technique that constructs an initial seed dictionary for inputs of the iterative approach. This chapter discusses a new method that extracts bilingual lexicons using the iterative approach, exploits the initial seed dictionary as the inputs and used as a weight vector for Perceptron learning task. Furthermore, a modified single-layer Perceptron will be introduced in this chapter. A system, which was built based on this technique, has improved the accuracy compared to initial epoch.

5.1 Methodology: Iterative Approach (IA)

In this Section, we describe our main work for bilingual lexicon extraction. Figure 5.1 describes an overall structure of the IA which requires two linguistic resources: the initial seed dictionary ($W(0)$) and comparable corpora. 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 because a synonym vector for each word can be add new weights into W and as the result we

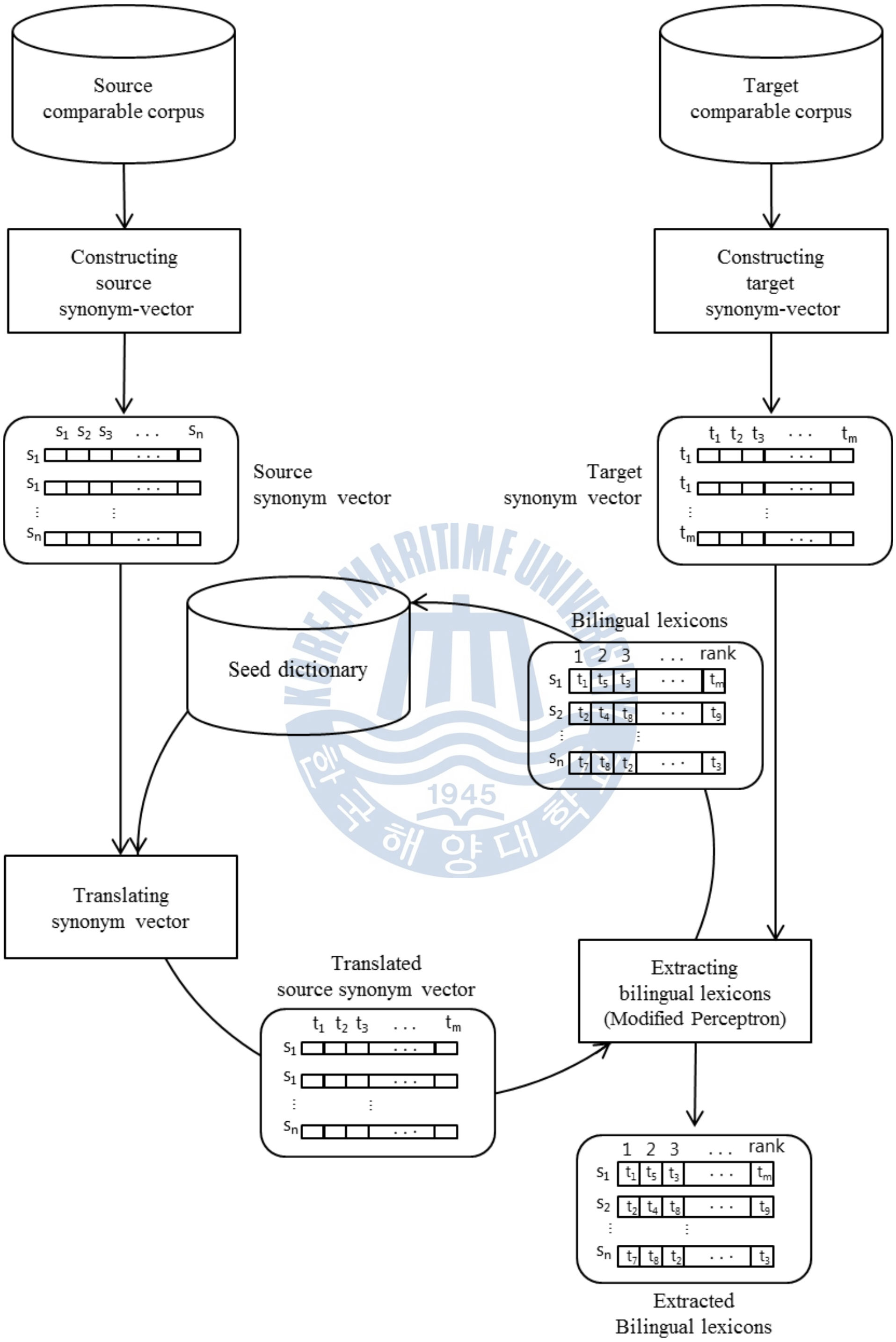

Source comparable corpus
Target comparable corpus
Constructing source synonym-vector
Constructing target synonym-vector
Source synonym vector
Target synonym vector
Seed dictionary
Bilingual lexicons
rank
Translating synonym vector
Translated source synonym vector
Extracting bilingual lexicons (Modified Perceptron)
Extracted Bilingual lexicons

Figure 5.1: An overall structure of the iterative approach (IA)

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:

(1) 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 target language (denoted as L_t) respectively, as in the same way of the 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 (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).

(2) To generate the translated vector ($\boldsymbol{y}$) of a source synonym vector ($\boldsymbol{x}$) instead of $\boldsymbol{s}$[4] using the modified Perceptron algorithm as in Equation (5.1):

$$y_j = \sum_{j=0}^{|y|} x_i w_{ij} \quad (5.1)$$

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

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

(3) To determine the desired synonym vector $\boldsymbol{d}$ of $\boldsymbol{x}$ as follows:

$$\boldsymbol{d} = argmax_{\boldsymbol{t}\in \boldsymbol{T}}\, sim(\boldsymbol{y},\boldsymbol{t}) = argmax_{\boldsymbol{t}\in \boldsymbol{T}}\, cos(\boldsymbol{y},\boldsymbol{t}) \quad (5.2)$$

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 Perceptron.

(4) To learn W via the modified Perceptron learning algorithm.

(5) To repeat the step (2) to (4) until convergence.

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

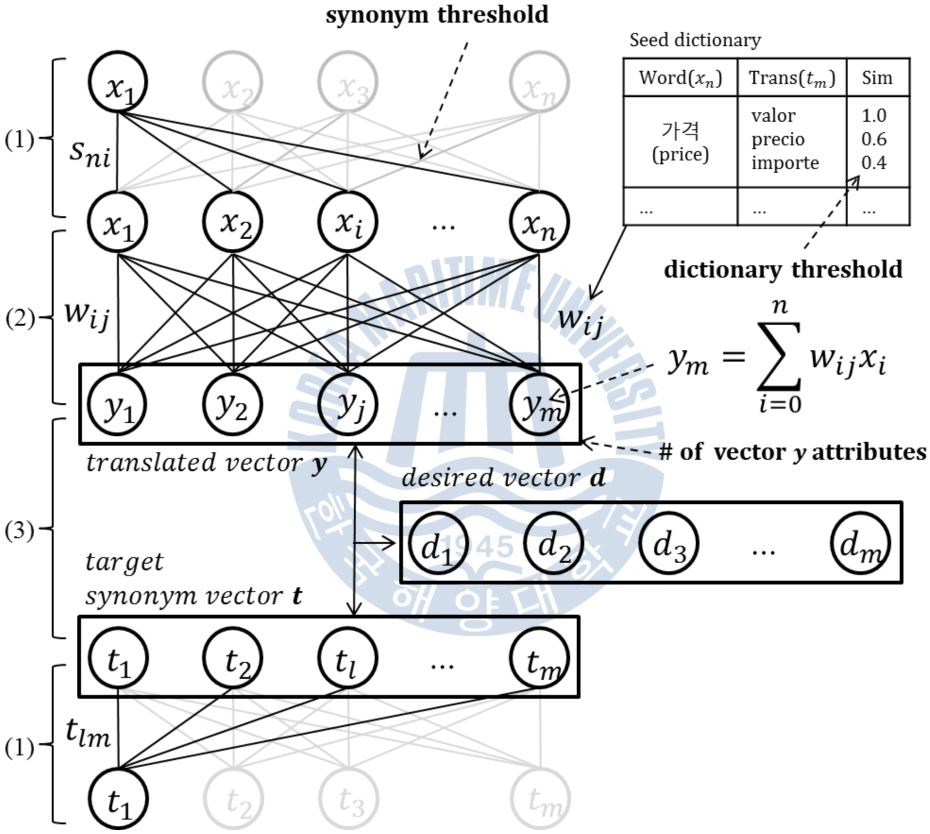
Figure 5.2 shows a detailed description of Step (1) to (3) for the IA implementation. As seen in Figure 5.2, Step (1) describes a source synonym vector $(\boldsymbol{x})$ and a target synonym vector $(\boldsymbol{t})$. Each $x_n\,(t_m)$ has $n\,(m)$ number of synonym words but denoted here $x_i\,(t_l)$ instead of $x_n\,(t_m)$ to designate a degree of similarity $(s_{ni}\,(t_{lm}))$ between two words. We discarded $s_{ni}\,(t_{lm})$ which is less than a synonym threshold.

Step (2) presents a translation procedure. For the same reason, the w_{ij} is discarded by the dictionary threshold.

Step (3) describes the labeling procedure of the training example. As mentioned before, we do not need the labeled training examples of the modified Perceptron made by manually, instead we dynamically compute the desired vectors for the training examples as in the step (3). The translated vector $\boldsymbol{y}$ of $\boldsymbol{x}$ is labeled based on similarity score which is calculated by the cosine similarity between $\boldsymbol{y}$ and target synonym vector. Unlike the previous steps, we restricted the size of the translated vector $\boldsymbol{y}$ to reduce computational cost for calculating similarity.

Finally we can perform the modified Perceptron learning algorithm. In summary, our modified Perceptron algorithm for updating weights is shown in Figure 5.2. Generally Perceptron algorithm can have negative weights but the modified

Perceptron algorithm has non-negative weights. This is because negative weight means they are not involved in translation. Therefore, we set negative weights to 0.



Word(x_n)	Trans(t_m)	Sim
가격 (price)	valor precio importe	1.0 0.6 0.4
…	…	…

Figure 5.2: A detailed description of Step (1) to (3) steps for the IA implementation

Learning Algorithm

```
Input: synonym-vectors $\mathbf{S}$ and $\mathbf{T}$, seed dictionary $\mathbf{W}(0)$
for $e = 1, \dots, E$ 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
        $\boldsymbol{d} = \operatorname*{argmax}_{\boldsymbol{t} \in \boldsymbol{T}} sim(\boldsymbol{y}, \boldsymbol{t})$
        $\Delta\mathbf{W} = \alpha(\boldsymbol{d} - \boldsymbol{y})\boldsymbol{x}$
        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
```

Figure 5.3: A modified Perceptron algorithm for updating weights

5.2 Experiments and results

In this thesis, we evaluate our approach for two different language pairs that are Korean-Spanish (KR-ES) and Korean-French (KR-FR) and compare with context-based approach as a baseline. For evaluation metrics, the accuracy, the mean reciprocal rank (MRR) and the recall are used as evaluation metrics.

5.2.1 Experimental setups

1) Comparable corpora

We built two pairs of comparable corpora that are KR-ES and KR-FR from the news articles and Europarl corpus (Koehn, 2005). 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. The average of the words in sentence is 16.2 in KR, 15.9 in ES, and 16.1 in FR, respectively. The corpora statistics are shown at Table 5.1.

Table 5.1: Statistics of comparable corpora

	KR	ES	FR
Number of sentence pairs	800,000	800,000	800,000
Average number of words per sentence	16.2	15.9	16.1
Number of distinct nouns	16,000	5,900	4,900
Domain	International news (51%)	International news (32%)	International news (61%)
	Not categorized news (49%)	Europarl corpus (68%)	Europarl corpus (39%)

2) Data pre-processing

All words were tokenized and lemmatized using the same tools as in Sub Section 4.2: U-tagger for Korean and Tree-Tagger for Spanish and French. All nouns 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[5] which occurring more than five were considered when generating context vectors in all languages. Finally, the comparable corpora comprised about 16,000 distinct nouns in Korean, 5,900 in Spanish and 4,900 in French each.

3) Building evaluation dictionary

We built two evaluation dictionaries (KR-ES and KR-FR) to evaluate the performance of the proposed method manually using the Web dictionary[6]. 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). Table 5.2 shows the average number of the translations per source word in each lexicon. The number means the degree of ambiguity

Table 5.2: The average number of the translations
per source word in the evaluation dictionaries for IA

Evaluation dictionary	HIGH	LOW
KR-ES	9.1	5.3
KR-FR	8.8	7.0

[5] 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

[6] http://dic.naver.com

4) Evaluation metrics

We evaluate the quality of translation candidates extracted by the IA. Similar to the evaluation in the CBA, the accuracy@1, the recall, and the mean reciprocal rank (MRR) are used as evaluation metrics. Accuracy@1 means the accuracy of the top 1.

5.2.1 Experimental results

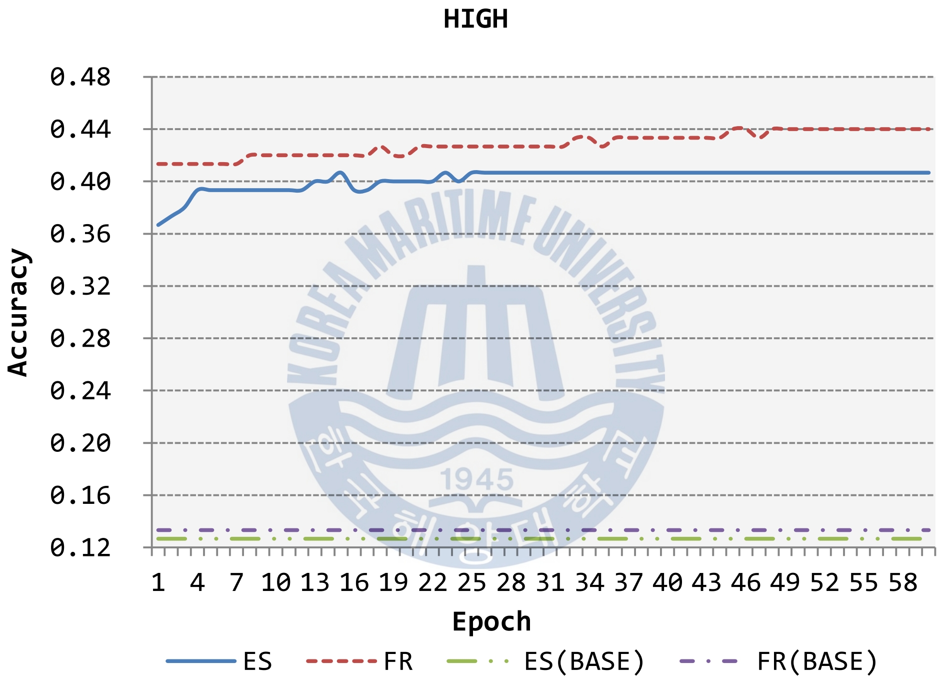


Figure 5.4: Accuracy@1 of the IA for HIGH words

We conducted 60 epochs with the learning rate α=0.01 for the KR-ES and KR-FR language pairs. The accuracy@1 of the HIGH words is shown in Figure 5.4. As shown in Figure 5.4, the accuracy@1 is slightly increased during 60 epochs. The accuracy@1 of the KR-ES increased from 0.366 to 0.406 and the KR-FR increased from 0.413 to 0.440 respectively. The performance of the IA for the KR-ES and the KR-FR is better than the baseline about 0.24 (KR-ES) and 0.28 (KR-FR).

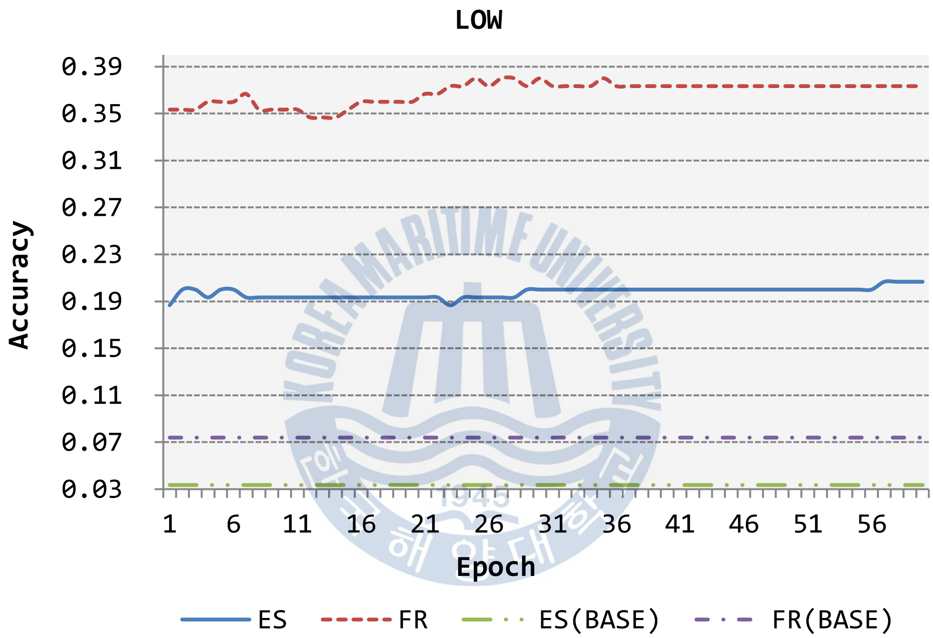


Figure 5.5: Accuracy@1 of the IA for LOW words

The accuracy@1 of the LOW words is shown in Figure 5.5. As seen in Figure 5.4, the accuracy@1 is improved during 60 epochs. The accuracy@1 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.

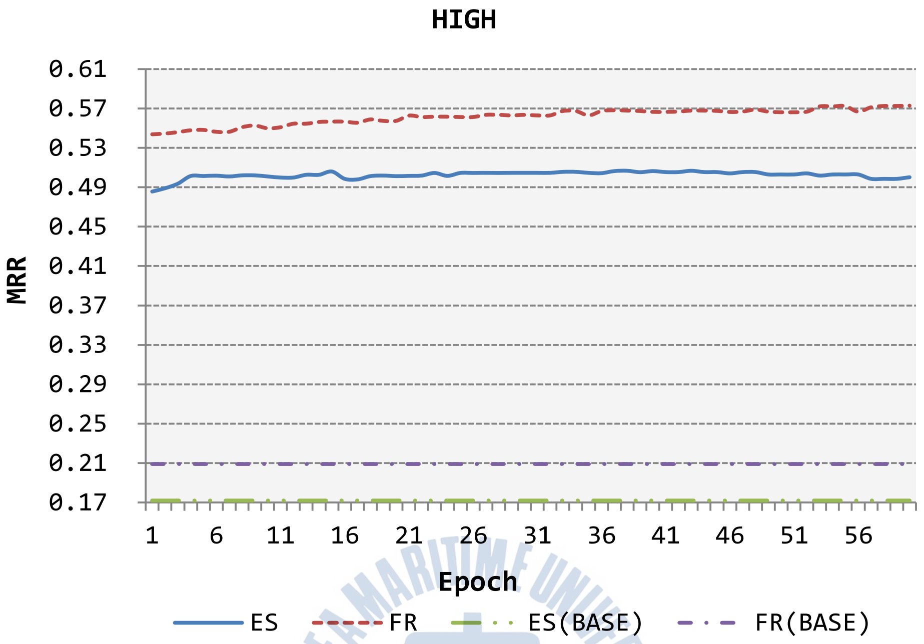


Figure 5.6: MRR of the IA for HIGH words at the top 5

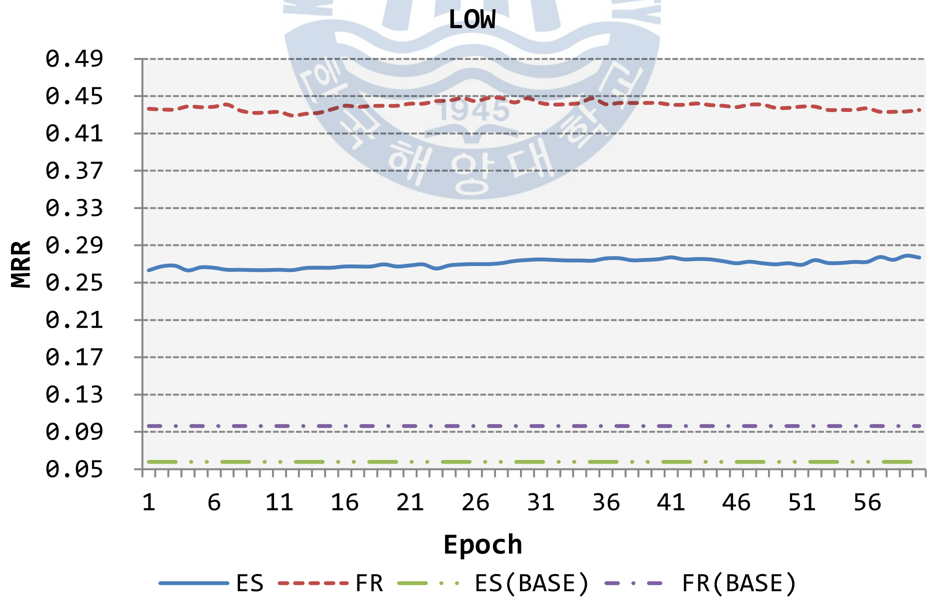


Figure 5.7: MRR of the IA for LOW words at the top 5

The MRR of the HIGH and LOW words at the top 5 are shown in Figure 5.6 and 5.7 respectively. As seen in Figures 5.6 and 5.7 for the HIGH, the MRR is increased about 0.014 (KR-ES) and 0.029 (KR-FR). 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 our experiment, generated synonym vector was noisy except itself. Therefore, the system performances at top 2 and more high ranks were decreased during epochs.

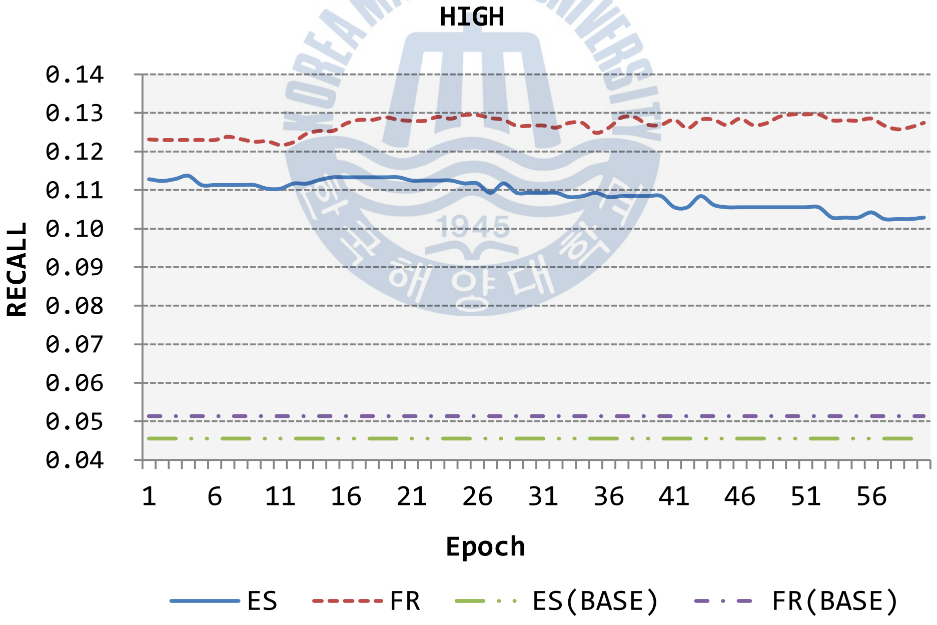


Figure 5.8: Recall of the IA for HIGH words at the top 5

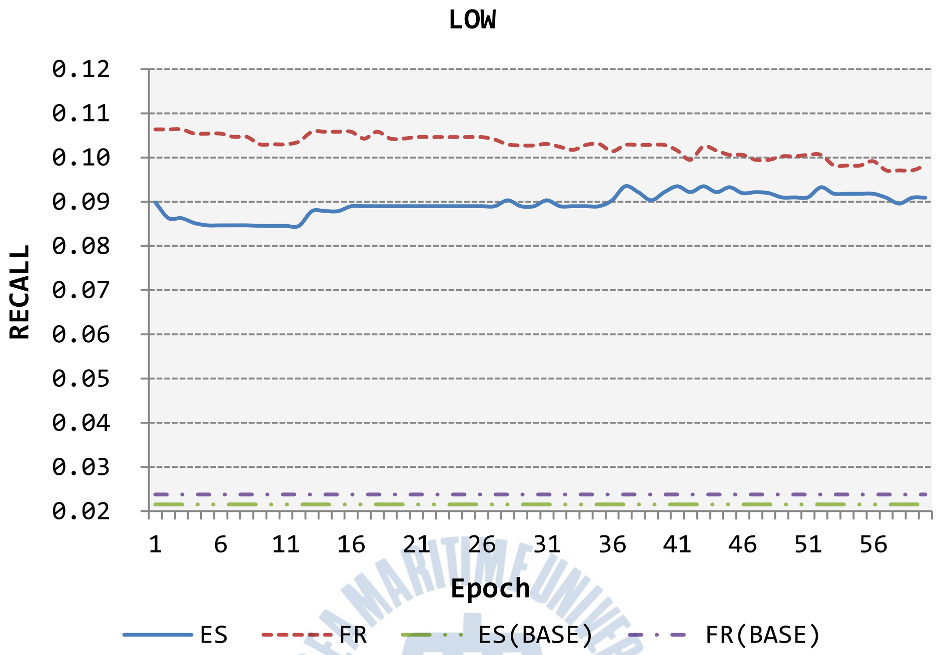


Figure 5.9: Recall of the IA for LOW words at the top 5

The recalls of the HIGH and LOW words at the top 5 are depicted in Figure 5.7 and 5.8 respectively. The recalls of KR-FR on both HIGH and LOW slightly improved but the recalls of KR-ES on HIGH words slightly decreased. The reason why the recall was decreased is the same as the reason described in the MRR. Furthermore, the performance outperforms the baselines in the both language pairs.

5.3 Discussions

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 gets better performance using the Perceptron algorithm. The approach requires a seed dictionary and a large amount of unlabeled training data. In this chapter, the initial seed dictionary is generated using the CBA and unlabeled training data is dynamically labelled by a modified Perceptron algorithm using a similarity measure during learning process. We extract bilingual lexicons using proposed iterative approach via the modified Perceptron algorithm. The empirical results show that our proposed method is significantly improving the performances of our model obtained with a modified Perceptron algorithm.

Now there are several questions to be answered in the experiments.

1) How many epochs are required?

In the experiment, we conducted 60 epochs. The accuracy gradually increased during epochs, and after that the accuracy becomes stable. The reasons for this are a characteristic of the Perceptron algorithm. Rosenblatt proved that if the inputs presented from more than two classes are separable then the Perceptron convergence procedure converges between those classes in finite time. The second reason is that there is a limitation of performance. After several epochs, the performance nearly reaches that limitation, making it hard to be further improved, thus the performance becomes convergence. The conclusion is that the iteration number at which the performance becomes convergence depends on the particular experimental settings such as synonym threshold, dictionary threshold, number of translated vector attribute and learning rate of the modified Perceptron algorithm.

2) How does the proposed method perform on different language pairs?

In our experiments, in Figure 5.4 and 5.5, we can see that the performance on two different language pairs of Korean-Spanish and Korean-French significantly improved. It indicates that the proposed method is language independent.

3) How does the synonym and dictionary threshold affect the performance?

We conducted the experiments using various threshold values between 0.05 and 0.5. There are some relations when the threshold on both synonym and dictionary is changing from 0.05 to 0.5 is presented in Table 5.3.

Higher threshold leads a vector to more reliable, it means that the synonym vector becomes more accurate and the initial dictionary may have more accurate translation candidates. It leads to better accuracy. Moreover, the number of vector attributes and the dimension of the vector are decreased. It reduces time complexity of the Perceptron algorithm. However, the threshold is set too high, it causes a problem that lose some information which are correct synonym in the synonym vectors or translation candidates in the initial seed dictionary thus the recall can be decreased. Otherwise, lower threshold takes more information so that the recalls can be increased but the accuracy decreased. Therefore, we set the synonym threshold to 0.2 and the dictionary threshold to 0.1 respectively.

Table 5.3: The variations of the relationship when the threshold is changing

Synonym and dictionary threshold	**Reliability of the information**	**# of vector attributes**	**Dimension of the vector**	**Recall**	**Accuracy**
θ(↑)	↑	↓	↓	↓	↑
θ(↓)	↓	↑	↑	↑	↓

4) How does the number of translated vector attributes affect the performance?

In Figure 5.2, we restricted the maximum size of the translated vector attributes. The reason why we restrict the size is to reduce computational cost for calculating similarity. The number of translated vector attributes affects same results described in the previous paraphrase. More attributes increases the percentage of words where the correct translation is contained within the top N, it also leads to more noisy and more time consuming. Therefore, we set a small number of attributes such as 50 is appropriate for our proposed method.

5) What kind of errors are occurred?

In this thesis, we have two problems that affect performance are inaccurate representation of synonym vectors and a semantic distinction of word. As seen in Figure 5.8 and 5.9, the recalls of HIGH on the KR-ES and the recalls of LOW on the KR-FR were decreased during epochs. The reason is that our system represented word into the vector by their synonyms and extracts the translation candidates from the most similar target synonym vector. Therefore, the system performance is very dependent to synonym vectors. However, the synonyms extraction is a difficult task to achieve and evaluate. Table 5.4 shows the partial example of the synonym vectors on the KR. The synonym vector of the word is noisy except oneself. Therefore, the performance of rank 2 and later was decreased during epochs.

The second problem is the semantic distinction. When we build the context vectors, we do not consider the meaning of the words. For example, the Korean word “가격(price)” has two meanings “가격(price)” and “가격(hit)” that are used different meanings but they are considered when generating context vector of the Korean word “가격(price)”. It makes the context vector noisy and inaccurate. We leave it as future work for this thesis.

Table 5.4: The partial examples of the Korean synonym vectors

Word	Synonym								
학교 (school)	**학교 (school)**	**1.00**	신학교 (seminary)	0.81	각급 (each class)	0.41	학생 (student)	0.38	...
가격 (price)	**가격 (price)**	**1.00**	하락 (fall)	0.47	동급 (same level)	0.42	인하 (reduction)	0.37	…
고객 (client)	**고객 (client)**	**1.00**	증서 (certificate)	0.89	예탁 (deposition)	0.22	만족도 (satisfaction)	0.11	…
경기 (economy)	**경기 (economy)**	**1.00**	장기화 (long period)	0.76	여파 (aftereffect)	0.67	내수 (demand)	0.37	…

Chapter 6

Conclusions and Future Works

This chapter summarizes all the findings, conclusions and implications based on the work that has been conducted. In addition, future works are presented.

This thesis presents a novel way of extracting bilingual lexicon extraction from comparable corpora based on the idea of information retrieval technique. The proposed method consists of two approaches: context-based approach (CBA) and iterative approach (IA). The CBA uses parallel corpora, pivot language and word alignment tool. The word alignment tool is used to construct context vectors. The pivot language is exploited for representing both of context vectors of a source language and a target language thus the initial seed dictionary is not required to translate a source vector to target language. The experiments are conducted for two language pairs of Korean-Spanish and Korean-French. The experimental results showed for the high-frequent words achieved at least 48.5% and up to 88.5% within the top 20 ranking candidates. The low-frequent words achieved at least 50.5% and up to 70% at the top 20 rank. These two pairs of constructed initial seed dictionary (KR-ES and KR-FR) are used for inputs to the CBA.

The main idea of the proposed method in this thesis is the IA. The IA extracts bilingual lexicons from comparable corpora and exploits 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 answers of the training examples for learning weights via the modified Perceptron algorithm. Our experimental results showed that the IA 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.

Bibliography

Brown, P., Pietra, V., Pietra, S., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263-311.

Chatterjee, D., Sarkar, S., and Mishra, A. (2010). Co-occurrence graph based iterative bilingual lexicon extraction from comparable corpora. In *Proceedings of the 4th International Workshop on Cross Lingual Information Access*, pages 35-42.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014). Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In *Proceedings of the 15th Conference on Intelligent Text Processing and Computational Linguistics (CICLing'14)*, pages 296-309.

Church, K., and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22-29.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora (VLC'95)*, pages 173-183.

Fung, P. (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)*, pages 1-16.

Grefenstette, G. (1994). Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers.

Grefenstette, G. (1998). The problem of cross-language information retrieval. Springer US.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Dejean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, pages 526-533.

Hazem, A., and Morin, E. (2012). Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 288-292.

Huelsenbeck, J. P., Hillis, D. M., and Nielsen, R. (1996). A likelihood-ratio test of monophyly. *Systematic Biology*, 45(4):546-558.

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

Ismail, A. (2012). *Minimally supervised techniques for bilingual lexicon extraction*. PhD thesis, University of York.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79-86.

Kwon, H., Seo H., and Kim, J. (2013). Bilingual lexicon extraction via pivot language and word alignment tools. In *Proceeding of the 6th International Workshop on Building and Using Comparable Corpora (BUCC'13)*, pages 11-15.

Kwon, H., Seo H., and Kim, J. (2014). Enhancing performance of bilingual lexicon extraction through refinement of pivot-context vectors, *Journal of KIISE: Software and Applications*, 41(7):492-500.

Lardilleux, A., Lepage, Y., and Yvon, F. (2011). The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189-217.

Plackett, R. L. (1983). Karl Pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, pages 59-72.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320-322.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

Salton. G, and McGill. M. J (1983). Introduction to Modern Information Retrieval, New York: McGraw-Hill.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44-49.

Seo, H., Kwon H., and Kim, J. (2013). Context-based lexicon extraction via a pivot language. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING'13)*, pages 11-15.

Shin, J. and Ock, C. (2012). A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary, *Journal of KIISE: Software and Applications*, 39(5):415-424.

Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D., and Gornostay, T. (2012). Analysis and evaluation of comparable corpora for under-resourced areas of machine translation. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC'12,)*, pages 17-19.

Tanaka, K. and Umemura. (1994) Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling'94)*, pages 297-303.

Tsunakawa, T., Okazaki, N., and Tsujii, J. (2008). Building bilingual lexicons using lexical translation probabilities via pivot language In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 18-22.

Yu, K and Tsujii, J. (2009). Bilingual dictionary extraction from wikipedia. In *Machine Translation Summit XII*, pages 379-386.

Wu, D., and Xia, X. (1994). Learning an English-Chinese lexicon from a parallel corpus. In Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA'94), pages 206-213.

Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 856-863.

Voorhees, E. (1999). The TREC-8 question answering track report. In *Proceedings of the text retrieval conference (TREC-8)*, pages 77-82.

감사의 글

무엇보다도 본 연구를 수행하는데 귀중한 지식과 지원을 해주신 김재훈 지도교수님께 먼저 감사의 말씀을 드립니다. 그리고 바쁘신 와중에도 논문지도에 많은 도움을 주신 박휴찬 교수님과 이장세 교수님께도 감사드립니다. 영문 감수 및 수정에 도움을 주신 이성민 박사님께도 감사드립니다. 또한 자연어 처리 연구실에서 함께 연구하고 도움을 주신 서형원 박사님, 황보천님, 전길호님, 최명길님, 남유림님, 천민아님, 이정태님과 황훈규 박사님, 강군호 조교님, 김경언 조교님, 김헌기님, 김효승님께도 감사의 말씀을 드립니다. 마지막으로 대학원과정을 무사히 마칠 수 있게 보살펴 주시고 지원해 주신 사랑하는 가족들에게 고마운 마음을 전합니다.