공학박사 학위논문

언어 자원이 부족한 언어 쌍에 대한 다국어 사전 추출

Multilingual Lexicon Extraction under Resource-Poor Language Pairs

지도교수 김재훈

2015 년 8 월

한국해양대학교 대학원

컴퓨터공학과

서 형 원

본 논문을 서형원의 공학박사 학위논문으로 인준함

위원장 공학박사 박휴찬 인

위 원 공학박사 류길수 인

위 원 공학박사 이장세 인

위 원 공학박사 고영중 인

지도교수 공학박사 김재훈 인

20015 년 6 월 29 일

한국해양대학교 대학원

Multilingual Lexicon Extraction under Resource-Poor Language Pairs

by

Hyeong-Won Seo

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate School of Computer Engineering
Korea Maritime and Ocean University

August 2015

APPROVED BY:

Professor Jae-Hoon Kim
(Advisor)

Professor Hyu-Chan Park
(Chair of Evaluation Committee)

Professor Keel-Soo Rhyu

Professor Jang-Se Lee

Professor Young-Joong Ko

Abstract

In general, bilingual and multilingual lexicons are important resources in many natural language processing fields such as information retrieval and machine translation. Such lexicons are usually extracted from bilingual (e.g., parallel or comparable) corpora with external seed dictionaries. However, few such corpora and bilingual seed dictionaries are publicly available for many language pairs such as Korean–French. It is important that such resources for these language pairs be publicly available or easily accessible when a monolingual resource is considered.

This thesis presents efficient approaches for extracting bilingual single-/multi-word lexicons for resource-poor language pairs such as Korean–French and Korean–Spanish. The goal of this thesis is to present several efficient methods of extracting translated single-/multi-words from bilingual corpora based on a statistical method.

Three approaches for single words and one approach for multi-words are proposed. The first approach is the pivot context-based approach (PCA). The PCA uses a pivot language to connect source and target languages. It builds context vectors from two parallel corpora sharing one pivot language and calculates their similarity scores to choose the best translation equivalents. The approach can reduce the effort required when using a seed dictionary for translation by using parallel corpora rather than comparable corpora. The second approach is the extended pivot context-based approach (EPCA). This approach gathers similar context vectors for each source word to augment its context. The approach assumes that similar vectors can enrich contexts. For example, *young* and *youth* can augment the context of *baby*. In the investigation described here, such similar vectors were collected by similarity measures such as cosine similarity. The third approach for single words uses a competitive neural network algorithm (i.e., self-organizing maps; SOM). The SOM-based approach (SA) uses synonym vectors rather than context vectors to train two different SOMs (i.e., source and target SOMs) in different ways. A source SOM is trained in an unsupervised way, while a target SOM is trained in a supervised way.

The fourth approach is the constituent-based approach (CTA), which deals with multi-word expressions (MWEs). This approach reinforces the PCA for multi-words (PCAM). It extracts bilingual MWEs taking all constituents of the source MWEs into consideration. The PCAM

identifies MWE candidates by pointwise mutual information first and then adds them to input data as single units in order to use the PCA directly.

The experimental results show that the proposed approaches generally perform well for resource-poor language pairs, particularly Korean and French–Spanish. The PCA and SA have demonstrated good performance for such language pairs. The EPCA would not have shown a stronger performance than expected. The CTA performs well even when word contexts are insufficient. Overall, the experimental results show that the CTA significantly outperforms the PCAM.

In the future, homonyms (i.e., homographs such as *lead* or *tear*) should be considered. In particular, the domains of bilingual corpora should be identified. In addition, more parts of speech such as verbs, adjectives, or adverbs could be tested. In this thesis, only nouns are discussed for simplicity. Finally, thorough error analysis should also be conducted.

언어 자원이 부족한 언어 쌍에 대한 다국어 사전 추출

서형원

컴퓨터공학과

한국해양대학교 대학원

초록

일반적으로 다국어 사전은 정보검색, 기계번역과 같은 자연어처리의 연구 분야에서 주요한 자원으로 사용되고 있다. 이와 같은 다국어 사전을 구축하기 위해서는 일반적으로 이중언어 말뭉치(bilingual corpora)와 초기 사전(seed dictionary) 등의 언어 자원이 주로 사용된다. 그러나 초기 사전과 같은 언어 자원은 한 언어 내에서는 쉽게 구할 수 있으나 언어 쌍(예를 들면, 한국어-불어)에 대한 언어 자원은 쉽게 구할 수 없는 실정이다.

이런 환경에서, 본 논문은 이렇게 언어 자원을 쉽게 얻을 수 없는 언어 쌍에 대하여 다국어 사전을 구축하는 여러 방법들을 제안한다. 본 논문의 목표는 한국어-불어, 한국어-스페인어와 같은 언어 쌍에 대하여 병렬/비교 말뭉치(parallel/comparable corpora)로부터 다국어 사전을 최대한 쉽고 효율적으로 구축하고자 한다. 이를 위해 본 논문에서는 네 가지 방법을 제안한다. 처음 세가지 방법은 단일단어에 대한 것이고 나머지 한 가지 방법은 다중단어에 관한 것이다. 첫 번째 방법은 PCA(pivot context-based approach)이라고 하며, 중간언어(pivot language)를 이용하여 대상이 되는 두 언어를 연결하는 방법이다. 이 방법은 하나의 중간언어를 공유하는 두 개의 병렬말뭉치로부터 문맥 벡터를 만들고 이들 벡터 간의

유사도를 비교함으로써 대역 단어를 찾는다. 이 방법은 비교 말뭉치 대신에 병렬 말뭉치를 사용하기 때문에 초기 사전과 같은 외부 자원의 사용을 줄일 수 있다는 장점이 있다. 두 번째 방법은 EPCA(extended pivot context-based approach)이라고 하며, 번역하고자 하는 원시 단어와 유사한 문맥 벡터들을 미리 수집하여 번역 단어를 찾고자 하는 일에 사용한다. 즉, 유사한 단어의 문맥이 번역하고자 하는 원시 단어의 문맥을 강화한다는 가정으로부터 출발한다. 예를 들어, '젊은이'와 '아이'가 '아이'의 문맥을 강화하는 데에 쓰인다는 것이다. 세 번째 방법은 SA(SOM-based approach)이라고 하며, 신경망 방법 중에 하나인 자기 조직화 지도(self-organizing map)를 이용한 방법이다. 이 방법은 문맥 벡터 대신에 유사어 벡터를 이용하여 두 개의 서로 다른 SOM 을 각각 다른 방식으로 학습시킨다. 네 번째 방법은 CTA(constituent-based approach)이라고 하며, 단일단어가 아닌 다중단어에 대한 방법이다. 이 방법은 다중단어를 구성하는 각 구성원들도 유사도를 계산하여 그 관계를 함께 고려하는 것이 특징이다. 이를 위해, 먼저 다중단어가 될 후보들을 선정해야 되는데, 이 때 PMI(pointwise mutual information)를 이용하여 먼저 가능한 후보들을 찾고 이전에 언급했던 PCA 를 그대로 이용하여 다중단어에 대한 번역 사전을 구축한다.

실험 결과, 언어 자원이 부족한 환경에서도 본 논문에서 제안하는 방법들은 좋은 성능을 보였다. 특히, PCA 나 SA 는 탁월한 성능을 보였고 EPCA 와 같은 경우는 기대만큼 높은 성능을 보이지는 않았다. 마지막으로 CTA 는 단어들의 문맥이 부족한 경우에 대해서도 높은 성능을 보였다.

향후에는 동음이의어에 대한 문제가 개선되어야 하고, 말뭉치들 간의 영역문제를 해결해야 한다. 또한, 더욱 다양한 품사로의 확장과 좀 더 깊은 오류 분석이 요구된다.

Contents

List of Abbreviations

ACC	Accuracy
CA	Context-based approach
CBM	Context-based method
CHI	Chi-square
CTA	Constituent-based approach
EA	Extended approach
EM	Expectation maximization
EN	English
EPCA	Extended pivot context-based approach
ES	Spanish
FR	French
IA	Iterative approach
KR	Korean
LL	Log-likelihood
LO	Log-odds
LSA	Latent semantic analysis
MRR	Mean reciprocal rank
MWE	Multi-word expression
PA	Pivot-based approach
PCA	Pivot context-based approach
PCAM	Pivot context-based approach for multi-words
PL	Pivot language
PMI	Pointwise mutual information
POS	Part of speech
PRC	Precision
REC	Recall
RR	Rated recall
SL	Source language
SMT	Statistical machine translation

SA	SOM-based approach
SOM	Self-organizing map
TL	Target language
TMBM	Topic model based method

List of Tables

List of Figures

Acknowledgement

Most of all, I wish to record my thanks to God for the unconditional love.

My most sincere thanks go to Prof. Jae-Hoon Kim for his support, guidance, and feedback throughout the duration of this degree. I would also like to thank Prof. Hyu-Chan Park, Prof. Keel-Soo Rhyu, Prof. Jang-Se Lee, and Prof. Young-Joong Ko for their insightful comments and feedback which improved this thesis.

I am extremely grateful to the Electronics and Telecommunications Research Institute (ETRI) for the award of a many-year research readerships in 2012-2015 that enabled me to investigate this topic. Moreover, they supported me financially so as to present my work in several conferences and workshops.

I would never have achieved most things in my life, including my Ph.D. research, without the unconditional support and love of my parents and my sister. Also, I would also like to express my deepest gratitude to my friends and colleagues to support me in a variety of ways.

가장 먼저, 살아계신 하나님 아버지께 이 영광을 돌립니다.

그리고, 저를 지금까지 돌봐주시고 격려해주시고 지도해주신 김재훈 교수님께 정말 깊은 감사의 말씀을 전합니다. 또한, 저의 학위논문을 완성하기까지 열정적으로 지도해주시고 도움주신 박휴찬 교수님, 류길수 교수님, 이장세 교수님, 그리고 먼 곳에서도 걸음해주신 고영중 교수님께도 깊은 감사의 말씀을 드립니다.

특히나, ETRI 김영길 박사님과 김창현 박사님 외 여러 관계자분들께도 깊은 감사를 드립니다. 그 분들 덕분에 ‘다국어 사전 구축’이란 연구 주제를 정할 수 있었고, 박사학위 과정 동안에 정말 좋은 환경 속에서 연구할 수 있었습니다.

마지막으로, 저를 아낌없이 사랑해주시고 기도해주시는 우리 부모님과 인애, 가족 모두에게 감사를 드립니다. 또한, 우리 자연어처리 연구실 사람들 모두에게 깊은 감사를 드립니다. 정말 감사합니다.

Chapter 1

Introduction

This chapter describes the notion of multilingual lexicons. Based on this notion, this chapter states the thesis' main subject and the research motivations. Several research objectives are outlined, and the overall organization of the thesis is presented.

1.1 Multilingual Lexicon Extraction

Extraction of bilingual translations of single words from comparable corpora has been studied by many researchers (Tanaka & Iwasaki, 1996; Fung, 1998; Picchi & Peters, 1998; Rapp 1999; Shahzad *et al*., 1999; Déjean *et al*., 2002; Chiao & Zweigenbaum, 2002; Ismail & Manandhar, 2010; Hazem & Morin, 2012). Such extracted lexicons have been used to construct statistical machine translation (SMT) models (Brown *et al*., 1990; Chen, 1993; Fung & Church, 1994; Kay & Roscheisen, 1993; Wu & Xia, 1994) or EM (expectation-maximization)-based models that align words in sentence pairs to construct technical terms (Dagan *et al*., 1993; Dagan & Church, 1994). Some researchers have compiled bilingual lexicons that consist of technical terms using similarity measures from bilingual lexical pairs (Gale & Church, 1991; Kupiec, 1993; Smadja & McKeown, 1996). In addition, other researchers have focused on the alignment of multi-words (Kupiec, 1993; Smadja *et al*., 1996). In most cases, such lexicons have been extracted from comparable corpora even though parallel corpora can provide promising results. However, collecting parallel corpora is time-consuming. Extracting such lexicons from comparable corpora has been studied since the late 1990s (Rapp, 1999; Koehn & Knight, 2002). However, using comparable corpora to extract bilingual lexicons yields poor results when orthographic features are not used. In such cases, large seed dictionaries can be considered to achieve higher accuracy (Koehn & Knight, 2002). Thus, the domains of bilingual corpora should be closely related, or the initial seed dictionaries should be of sufficient size.

Most studies of bilingual lexicon extraction from comparable corpora have used context vectors from two different languages. A context-based approach (CA) was proposed (Rapp, 1995; Fung, 1998), and many other methods have been derived from this approach. However, the CA uses comparable corpora; therefore, the previously mentioned limiting characteristics should be considered. To address the limitations related to the usage of seed dictionaries or orthographic features, many other studies have considered the entry size of the seed dictionary or similarity score measurements (Fung, 1998; Rapp, 1999; Koehn & Knight, 2002; Chiao & Zweigenbaum, 2002; Daille & Morin, 2005; Prochasson *et al*., 2009). Alternatively, some researchers (Chatterjee *et al*. 2010; Chu *et al*., 2014; Kwon *et al*., 2014) have studied methods of extending seed dictionaries by iteratively extracting bilingual lexicons until a reasonable iteration converges.

Nevertheless, the accuracy of bilingual lexicon extraction via comparable corpora is quite poor (Ismail & Manandhar, 2010). Thus, if stronger performance is required, either large-scale bilingual (parallel or comparable) corpora or seed dictionaries with sufficient entries should be prepared. In addition, most previous studies have dealt with resource-rich language pairs such as English to Chinese, Spanish, and German. Accessing or constructing linguistic resources for these language pairs is much easier than it is for Korean → French or Spanish.

This thesis deals with bilingual lexicons from bilingual corpora and adapts the methodology to multilingual resources or circumstances. Thus, the thesis provides a comprehensive discussion of multilingual lexicon-extraction methods. For simplicity, the names of bilingual lexicon extractions rather than multilingual lexicons are used in the remainder of the thesis.

1.2 Motivations and Goals

As mentioned previously, extracting bilingual lexicons requires many linguistic resources when comparable corpora are considered. For resource-rich language pairs such as English–* (any language), attempts to collect them are not as significant of undertakings as they are for some other language pairs such as Korean–*. Publicly accepted linguistic resources for resource-poor language pairs such as Korean–French[1] and Korean–Spanish are very rare, whereas monolingual resources are readily available. Even if such resources for resource-poor language pairs are available, they are very small in scale or incomplete. Thus, this thesis focuses on the minimum usage of external/extra linguistic resources.

The primary focus of this thesis is bilingual lexicon extraction specifically when publicly available linguistic resources such as bilingual dictionaries are insufficient. Furthermore, single words and multi-word expressions (MWEs) are discussed. MWE extraction forms a large research field, and MWE lexicons are used for many natural language processing (NLP) domains such as building ontologies (Venkatsubramanyan & Perez-Carballo, 2004) and information retrieval (Doucet & Ahonen-Myka, 2004). The thesis does not focus on bilingual MWE extraction; the primary focus is extracting bilingual single-word or MWE lexicons when only resource-poor language pairs are available.

The main goal of this thesis is to propose effective methods of addressing the limitations of earlier methods of extracting multilingual lexicons from resource-poor language pairs. Several studies that are closely related to the proposed approaches are reviewed. These studies have focused on the extraction of bilingual parallel words, that is, single words or MWEs. Then, several approaches to mitigate the limitations of an approach chosen as the baseline (the standard approach) are proposed. The proposed approaches are based on several assumptions, which can be summarized as follows.

- **Adaptation for resource-poor languages**: There are thousands of languages on this planet and many linguistic resources. Many people speak English as a native or foreign language. Moreover, monolingual resources such as documents in English can be easily found online, and bilingual resources for English are very common. Unfortunately, bilingual resources for specific language pairs such as Korean–French

[1] The symbol "–" indicates bidirectionality; i.e., source to target and target to source.

and Korean–Spanish are very rare. This thesis only considers such resource-poor language pairs.

- **Minimum usage of resources**: This thesis deals with resource-poor languages; therefore, excluding external linguistic resources such as a parser and the scale of a seed dictionary or their extensions is a crucial point.
- **Simplified experiments for efficiency**: This thesis evaluates the effectiveness of many approaches. Thus, reducing the effort and time required to perform experiments is a consideration. Investigating as many words as possible causes inefficient tests, implementations, or evaluations. Thus, the experiments discussed in this thesis focus on nouns for bilingual single-word extraction (resp. noun phrases for multi-word expression extraction).

1.3 Organization

The remainder of this thesis is organized as follows. Chapter 2 presents detailed reviews of many methods that are closely related to the proposed approaches. In particular, several statistical extraction methods for single words and multi-words are reviewed. The CA, the extended approach (EA), and the iterative approach (IA) are used for single words. An earlier approach for multi-words, namely the pivot context-based approach for multi-words (PCAM), which has been presented in several previous studies of MWEs, is reviewed. In addition, self-organizing maps (SOMs) are briefly reviewed. Finally, several evaluation measures are described.

Chapter 3 discusses the pivot context-based approach (PCA). The PCA extracts bilingual lexicons via a pivot language; therefore, using comparable corpora with insufficient overlapped terms or a seed dictionary is unnecessary. Using resource-poor language pairs that share one resource-rich language is the key point in this work. In addition, this chapter presents experimental results with summarized characteristics.

Chapter 4 presents the extended pivot context-based approach (EPCA), which was proposed to improve the PCA. This approach is based on an earlier one, the EA, which is based on the assumption that similar words can reinforce their contexts. The results of several experiments that demonstrate the value of the EA are presented and discussed.

Chapter 5 proposes the SOM-based approach (SA). The SA uses SOMs to improve the CA.

This approach is very useful with small seed dictionaries. The SOMs are used in natural and slightly abnormal ways. This chapter describes how SOMs can be used to extract bilingual lexicons and supports the methodology with reasonable experimental results.

Chapter 6 presents the constituent-based approach (CTA) for extracting MWEs with reference to various earlier studies. Based on in-depth analysis, various associated errors are identified. In addition, ways to improve the CTA are described.

Finally, Chapter 7 summarizes and concludes the thesis and presents suggestions for future research.

Chapter 2

Background and Literature Review

This chapter provides background information and reviews several previous studies closely related to the proposed approaches. In particular, this chapter summarizes several statistical extraction methods for single words and MWEs. Note that the context-based approach (CA), a method for single words (Section 2.1.1) is considered the base approach in this thesis. In addition, many previous approaches for extracting MWEs from bilingual corpora are reviewed, and several evaluation measures such as accuracy, precision, and rated recall are described.

2.1 Extraction of Bilingual Translations of Single Words

There have been many previous approaches for extracting bilingual lexicons from bilingual corpora (Tanaka & Iwasaki, 1996; Fung, 1998; Picchi & Peters, 1998; Rapp 1999; Shahzad *et al.*, 1999). Fung (1998) used aligned parallel corpora and comparable corpora to discuss the paradigm change from parallel to comparable corpora. A CA using such comparable corpora has been proposed (Rapp, 1995; Fung, 1998). In addition, approaches that use dependency relationships among words to extract more salient contexts have been proposed (Garera, 2009; Yu & Tsujii, 2009). The dependency-based approach uses external dependency parsers as resources. Collecting or building such parsers for language pairs can be a burden. Therefore, this thesis focuses on the CA (i.e., the base approach using context vectors) and its improvement.

The following sections deal with the CA, EA (Déjean & Gaussier, 2002), and the pivot-based approach (PA).

2.1.1 Context-based approach

The context-based approach (CA) (Rapp, 1995; Fung, 1998) builds context vectors by considering contextually relevant words in small windows. Selecting similar context vectors in the source and target languages is the key idea of this approach, which is based on the assumption that *if two words are mutual translations, then their more frequent collocates are likely to be mutual translations as well* (Déjean *et al*., 2002). It is also based on the identification of first-order affinities for each source and target language. *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word* (Grefenstette, 1994a, p. 279). This approach has been widely studied (Ismail & Manandhar, 2010; Hazem & Morin, 2012). Most earlier studies were closely related in their use of comparable corpora, which are defined as *sets of texts in different languages that are not translations of each other* (Bowker & Pearson, 2002), or of small-scale bilingual seed dictionaries. The use of comparable corpora is generally reasonable because parallel corpora for specific language pairs are not widely available. In addition, collecting or building parallel corpora for all language pairs is almost impossible. However, the use of comparable corpora can lead to poorer performance. However, comparable corpora do not always result in performance worse than that attained with parallel corpora. To achieve higher performance with comparable corpora, larger-scale corpora are required. The structure of the CA is shown in Figure 2.1.

(1) **Building context vectors**: First, two types of context vectors should be built from monolingual corpora. In this case, contexts presented by vectors indicate that some words occur within a fixed window size. At this point, word order is not important for counting co-occurrences. After all word co-occurrences have been counted, association measures such as log-likelihood (LL) (Dunning, 1993), chi-square (CHI) (Manning &

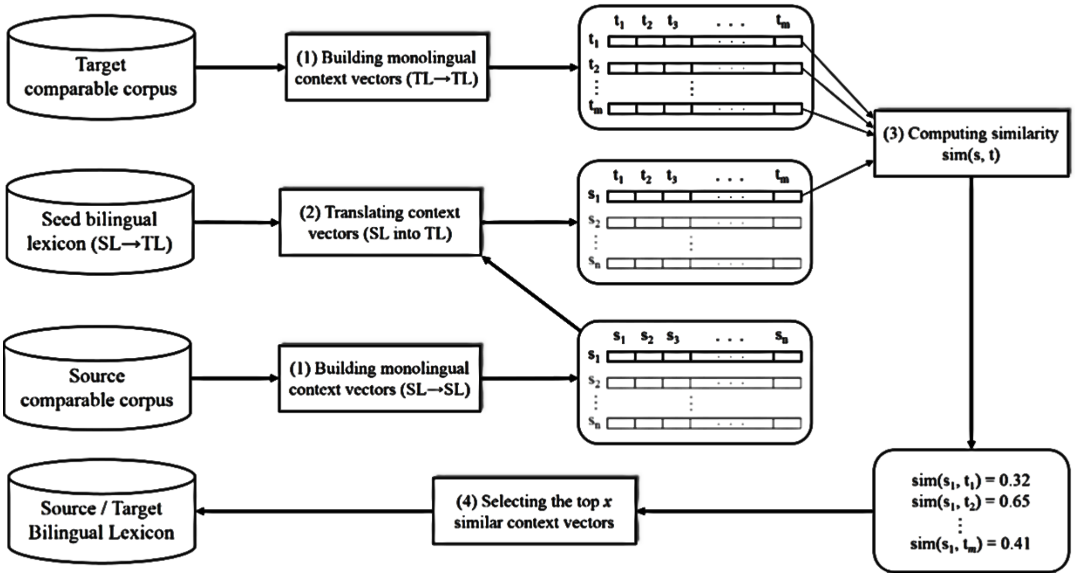

Target comparable corpus
(1) Building monolingual context vectors (TL→TL)
t_1 t_2 t_3 . . . t_m
(3) Computing similarity sim(s, t)
Seed bilingual lexicon (SL→TL)
(2) Translating context vectors (SL into TL)
s_1 s_2 s_n
Source comparable corpus
(1) Building monolingual context vectors (SL→SL)
s_1 s_2 s_3 . . . s_n
Source / Target Bilingual Lexicon
(4) Selecting the top *x* similar context vectors
sim(s_1, t_1) = 0.32
sim(s_1, t_2) = 0.65
sim(s_1, t_m) = 0.41

Figure 2.1: Overall structure of CA

Schütze, 1999), and pointwise mutual information (PMI) (Fano, 1961) are computed. Based on these values, context vectors are built for both the source and target languages.

(2) **Translating context vectors**: In this step, source context vectors should be translated into the target language based on a seed dictionary. In addition, all entries belonging to words not found in the target part of the seed dictionary are eliminated. Thus, only target words found in the seed dictionary (SL→TL)[2] are presented in the vector space. Both context vectors, i.e., those of the source and target, are comparable because of the translation.

(3) **Computing similarity scores**: After all source and target words have been presented using the same vector space dimensions, each source context vector is compared with all of the target context vectors using a vector distance measure (Manning & Schütze, 1999) such as cosine similarity or weighted Jaccard indexes (Grefenstette, 1994b). This thesis assumes that two words that share similar context words in different languages are likely translations.

(4) **Selecting similar context vectors**: After all similarity scores have been computed, the scores are sorted in descending order. Several target context vectors with the highest scores are selected for a single source word. Steps (2) and (3) are repeated for all source

[2] The SL (resp. TL) means ‘source language’ (resp. ‘target language’), and the symbol ‘→’ indicates unidirectionality, i.e., source to target.

words.

As can be seen, the method is quite simple; however, despite this simplicity, it has demonstrated good results with single-word terms from large corpora of several million words. Fung (1998) obtained 76% precision for the top 20 candidates from English and Chinese news articles. Rapp (1999) improved the precision to 89% for the top 10 candidates from English and German news articles. These experimental results indicate that the algorithm is very adaptive to various experimental circumstances and language pairs. The important thing is the coverage of the seed dictionary. Examples of studies that employed the CA with seed dictionaries as well as the features that affected performance are listed as follows:

- Size of the context window: Three sentences (Daille & Morin, 2005), 25 words (Prochasson *et al*., 2009)
- Entry size of seed dictionary: 1k (Koehn & Knight, 2002), 16k (Rapp, 1999), around 2k (Fung, 1998; Chiao & Zweigenbaum, 2002; Daille & Morin, 2005)
- Similarity score measure: city-block measure (Rapp, 1999), cosine distance measure (Fung, 1998; Chiao & Zweigenbaum, 2002; Daille & Morin, 2005; Prochasson *et al*., 2009), Dice or Jaccard indexes (Chiao & Zweigenbaum, 2002; Daille & Morin, 2005)

Comparable corpora and seed dictionaries are essential resources for the CA. It is easier to construct comparable corpora than parallel corpora for specific language pairs, which is advantageous when extracting bilingual lexicons in resource-poor language pairs such as Korean–French. However, the accuracy of bilingual lexicon extraction via comparable corpora is quite poor (Ismail & Manandhar, 2010). Moreover, extraction with comparable corpora requires an additional external linguistic resource, particularly, a seed dictionary. A seed dictionary requires approximately 10k to 20k entries to achieve higher accuracy (Fung, 1995; Rapp, 1999). Thus, if higher performance is required, larger-scale corpora and sufficient bilingual seed entries should be available.

In the next section, the EA, which attempts to improve the performance of the CA, will be presented.

2.1.2 Extended approach

As mentioned previously, the CA relies heavily on the coverage of the seed dictionary. Many approaches to reduce the load of the seed dictionary have been proposed. Chiao and Zweigenbaum (2002) and Déjean *et al*. (2002) focused on extending the entries in a seed dictionary through specialized dictionaries or multilingual thesauri. Alternatively, Déjean and Gaussier (2002) proposed the extended approach (EA), which focuses on enrichment of the context words to be translated. The EA assumes that synonyms share the same environments. This assumption is based on the identification of second-order affinities in the source language: *Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar* (Grefenstette, 1994a, p. 280). Figure 2.2 illustrates the overall structure of the EA in more detail.

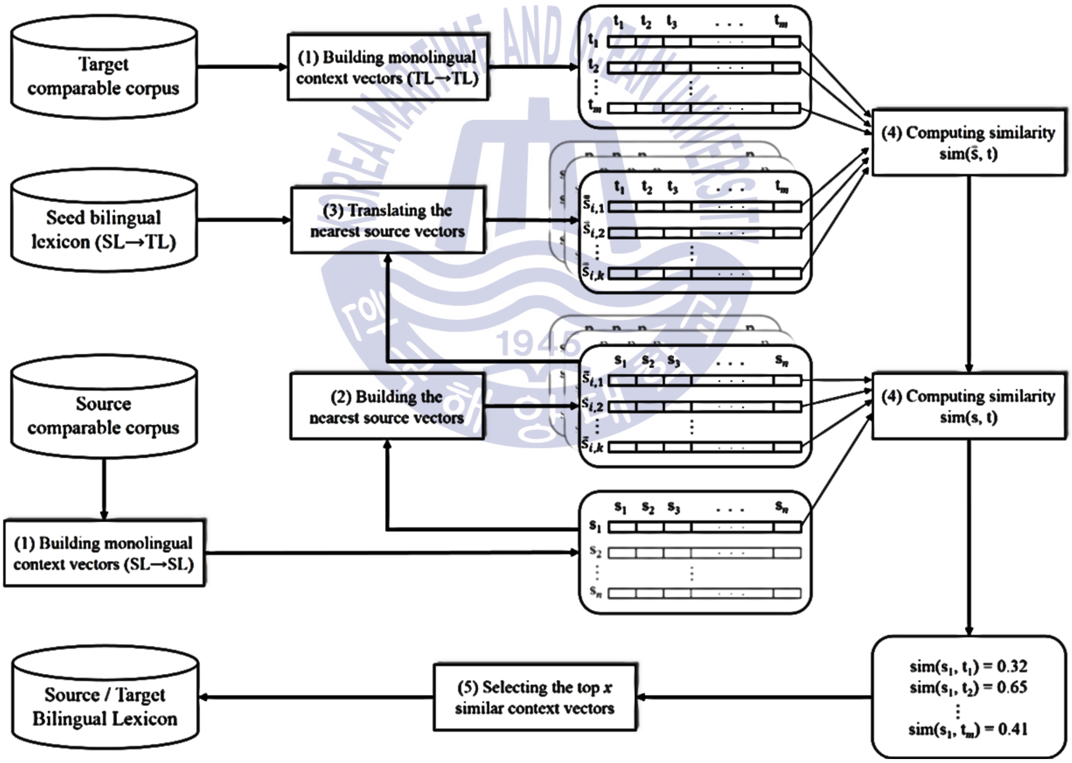

Target comparable corpus
(1) Building monolingual context vectors (TL→TL)
t_1 t_2 t_3 . . . t_m
t_1
t_2
t_m
(4) Computing similarity sim($\bar{s}$, t)
Seed bilingual lexicon (SL→TL)
(3) Translating the nearest source vectors
t_1 t_2 t_3 . . . t_m
$\bar{s}_{i,1}$
$\bar{s}_{i,2}$
$\bar{s}_{i,k}$
Source comparable corpus
(2) Building the nearest source vectors
s_1 s_2 s_3 . . . s_n
$\bar{s}_{i,1}$
$\bar{s}_{i,2}$
$\bar{s}_{i,k}$
(4) Computing similarity sim(s, t)
(1) Building monolingual context vectors (SL→SL)
s_1 s_2 s_3 . . . s_n
s_1
s_2
s_n
Source / Target Bilingual Lexicon
(5) Selecting the top x similar context vectors
sim(s_1, t_1) = 0.32
sim(s_1, t_2) = 0.65
sim(s_1, t_m) = 0.41

Figure 2.2: Overall structure of EA

(1) **Building context vectors**: This step is very similar to the first step of the CA. All source words (resp. target words) are presented using vector-space dimensions. At this point,

each element from the vector space is a co-occurred word within a fixed window size. First, word co-occurrences are counted, following which an association measure such as PMI is computed to define the vector entries.

(2) **Building nearest context vectors**: For a source vector $\vec{s}_i$ (i denotes a source word index), its k nearest context vectors $\left(\vec{\bar{s}}_{i,1}, \vec{\bar{s}}_{i,2}, \dots, \vec{\bar{s}}_{i,k}\right)$ are collected in this step. As described previously, synonyms sharing the same environments can enrich the context vectors of the source word to be translated. In other words, enrichment of the context of source word s_i by selecting its synonyms $\left(\bar{s}_{i,1}, \bar{s}_{i,2}, \dots, \bar{s}_{i,k}\right)$ (i.e., the closest k words) can help determine the correct translation of s_i without extending the seed dictionary based on an external bilingual dictionary or a multilingual thesaurus. Synonyms are generally selected based on similarity scores among source context vectors. These similarity scores will be used again in step (4).

(3) **Translating nearest context vectors**: All nearest context vectors $\vec{\bar{s}}_i$ are translated via the seed dictionary (SL→TL). If a source entry has several translations in the seed dictionary, only the most frequent translation in the target corpus is considered. Every target source entry not found in the seed dictionary is eliminated.

(4) **Computing similarity scores**: In this step, the similarity scores between source word s_i and target words t are computed. To measure the similarity score $\mathrm{sim}\left(s_i, t_j\right)$, two types of similarity scores, $\mathrm{sim}(\mathrm{s}_i, \bar{s}_i)$ and $\mathrm{sim}\left(\bar{\bar{s}}_i, t_j\right)$, should be computed first. Here j denotes a target word index, $\bar{s}_i = \left(\bar{s}_{i,1}, \bar{s}_{i,2}, \dots, \bar{s}_{i,k}\right)$ denotes the words nearest to the source word s_i, and $\bar{\bar{s}}_i$ denotes the words nearest to $\bar{s}_i$ translated into the target language. Note that similarity scores $\mathrm{sim}(\mathrm{s}_i, \bar{s}_i)$ have already been computed to obtain the nearest context vectors $\vec{\bar{s}}_i$. Therefore, the similarity score $\mathrm{sim}\left(\mathrm{s}_i, t_j\right)$ is defined by Equation (2.1).

$$\mathrm{sim}\left(s_i, t_j\right) = \sum_{e=1}^{k} \mathrm{sim}\left(\mathrm{s}_i, \bar{s}_{i,e}\right) \times \mathrm{sim}\left(\bar{\bar{s}}_{i,e}, \mathrm{t}_j\right) \qquad (2.1)$$

As can be seen, the k nearest context vectors and their translated vectors strengthen common contexts among source word s_i and target word t_j. If the source word has as many nearest words as possible, a similarity score between two words can be greater than that computed directly.

(5) **Selecting similar context vectors**: After all similarity scores have been computed, these scores are sorted in descending order. Several target context vectors with the highest scores are selected for a single source word. Steps (2) and (4) are repeated for all source words.

Déjean and Gaussier (2002) discussed the problem of selecting such nearest units. Selection of the best translation depends on data. Thus, it is usually defined empirically. Alternatively, another scoring method was proposed by Daille and Morin (2005). They collected the leader vectors of the *x* nearest words and then calculated similarity scores between the collected leader vectors and target context vectors. This thesis simply refers to the EA and uses the key feature in the EPCA, which is described in Chapter 4.

2.1.3 Pivot-based approach

As mentioned previously, the CA (Rapp, 1995; Fung, 1998) uses comparable corpora to extract context vectors. Since comparable corpora do not provide clues or traces about contexts, a bilingual seed dictionary, which is used to translate the source vector entries into the target language, is a very crucial resource. As expected, the applicability and performance of this approach depend on the size/coverage of the seed dictionary (Fung 1995; Rapp, 1999). Therefore, some researchers have studied the extension of seed dictionaries (Koehn & Knight, 2002; Koehn *et al*., 2003; Tsunakawa *et al*., 2008). However, extending a seed dictionary is not a fundamental solution. As Fung (1995) and Rapp (1999) reported, a seed dictionary requires approximately 10k to 20k entries to achieve higher accuracy. In general, the accuracy of bilingual lexicon extraction via comparable corpora is quite poor (Ismail & Manandhar, 2010); however, this low accuracy does not mean that bilingual lexicon extraction via comparable corpora is a useless approach. If parallel corpora are considered to be the input used for bilingual lexicon extraction, corpora of sufficient scale should be available. In addition, such corpora are difficult to collect for all language pairs.

To address these problems, some studies (Tanaka & Umemura, 1994; Bond *et al*., 2001; Paik *et al*., 2001; Shirai & Yamamoto, 2001; Schafer & Yarowsky, 2002; Goh *et al*., 2005) have focused on pivot languages (i.e., pivot-based approach). The key idea of the pivot-based approach

(PA) is to construct a bilingual lexicon between the source and target languages by merging two different bilingual lexicons that share one pivot language (i.e., source–pivot and pivot–target). Figure 2.3 illustrates this process with examples. In this case, two lexicons are passed to a mixing model that combines two different entries based on a specific method such as exact string matching.

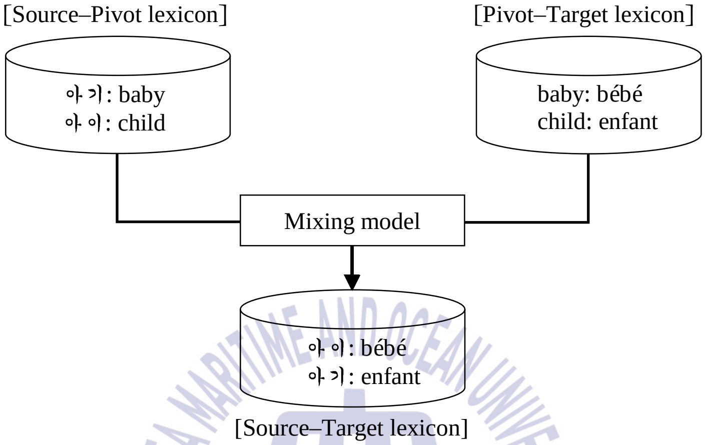


Figure 2.3: Example of combining two lexicons

However, there is a critical disadvantage, namely, a polysemy problem. To solve this problem, Tanaka and Umemura (1994) utilized the structures of dictionaries to measure the nearness of the senses of words. Bond *et al.* (2001) proposed using semantic classes to rank translation equivalents; in their method, word pairs with compatible semantic classes are preferred to those with dissimilar classes. Shirai and Yamamoto (2001) measured the degree of similarity between two words (i.e., in source and target languages) based on the number of pivot words shared by the words. Paik *et al*. (2001) used multiple pivot languages (i.e., English and Chinese) to improve the accuracy of bilingual lexicon extraction. The method proposed by Paik *et al*. is applicable to a specific language pair such as Korean–Japanese because Korean and Japanese share Chinese characters for most words. Schafer and Yarowsky (2002) presented a method to induce translation lexicons without parallel corpora or a direct bilingual seed dictionary by combining iteratively trained similarity measures such as string similarity, context similarity, date distributional similarity, and the similarity of word frequency and burstiness statistics. Goh *et al.* (2005) attempted to construct a bilingual lexicon between Japanese and Chinese by building a dictionary for Kanji words with simple conversion from Kanji to Hanzi. This method begins from the fact that most Japanese Kanji characters are similar to Chinese ideographs. Goh *et al*. assumed that, *since most of the kanji characters are originally from China, the usage should remain unchangeable in certain contexts*. They performed several experiments for nouns and verbal nouns and showed that the proposed method could improve performance.

2.2 Extraction of Bilingual Translations of Multi-Word Expressions

Many theoretical and practical studies on multi-word expressions (MWEs) have been undertaken (Nunberg *et al*., 1994; Manning & Schutze 1999; Sag *et al*., 2002). However, identifying and treating MWEs is difficult due to the lack of adequate linguistic resources such as parallel corpora in various languages. This problem has increasingly attracted the attention of the NLP community. Various NLP applications have been proposed that are based on bilingual MWE lexicons such as building ontologies (Venkatsubramanyan & Perez-Carballo, 2004), information retrieval (Doucet & Ahonen-Myka, 2004), text alignment (Venkatapathy & Joshi, 2006), and machine translation (Baldwin & Tanaka, 2004; Uchiyama *et al*., 2005).

An MWE has various definitions depending on the focus. MWEs can be defined as expressions that consist of two or more words that correspond to some conventional way of expressing an idea (Manning & Schutze, 1999), as the co-occurrence of sequences of words that tend to co-occur more frequently than chance and are either decomposable into multiple simple words or idiosyncratic (Baldwin *et al*., 2003), and as groups of two or more words or terms in a language lexicon that generally convey a single meaning (Monti *et al*., 2011). The latter definition conveys the basic role of MWEs. In human language, MWEs appear very frequently, either verbally or literally. They can be noun phrases such as *my best friend*, *a beautiful red dress*, and *the dog on the sofa;* collocations such as *alcoholic drink* and *nuclear family*; poly-words such as *by the way*, *of course,* and *in a flash;* idioms such as *take action* and *pulling my leg*; and phrasal verbs such as *give up* and *break up*. The wide range of possible usages accounts for the various definitions of MWEs (Rayson *et al*., 2009). MWE identification and alignment methods based on identified MWE candidates are discussed in the following sections.

2.2.1 MWE identification

Many methods of identifying various types of MWEs in different domains have been proposed. Some have focused on collocational behavior of MWEs (Church & Hanks, 1990). Pecina (2008) evaluated 55 different association measures such as PMI and mixed them to determine their influences on each other. He showed that mixing different types of association measures is more effective than using one standard measure. Other studies based on association measures have been conducted (Chang *et al*., 2002; Villavicencio *et al*., 2007; Bouma, 2010) to determine the measure that shows the highest efficiency for identifying several types of MWEs in several languages. However, Piao *et al*. (2003) reported that approximately 68% of MWEs occur only once or twice in their corpora; thus, statistical approaches may return less than satisfactory results when infrequent MWEs are considered.

As well as identifying the usage of linguistic properties of MWEs as an important issue, Piao *et al*. (2005) also contended that considering linguistic information and word statistics together is better than considering them independently. The research performed by Ramisch *et al*. (2008) supports this idea. Ramisch *et al*. showed that statistical measures on their own are generally sufficient to identify MWEs. However, for different languages and MWE types, such measures would have limited success in capturing specific linguistic features such as compositionality. Moreover, the study reported that some measures such as PMI usually show good performance; however, they may return different results for different types of MWEs. In addition, the study reported that adding type-specific linguistic information such as part-of-speech (POS) sequence patterns can significantly improve performance over that achievable by considering statistic measures alone.

Several studies concentrated on syntactic or semantic properties of MWEs. Wermter and Hahn (2004) explored the (non-)modifiability of preposition–noun–verb combinations in German, and Fazly and Stevenson (2006) and Bannard (2007) quantified the syntactic fixedness of English verb-noun phrases. Recently, Green *et al*. (2011) used a parsing module, specifically, tree substitution grammars, to identify French MWEs of arbitrary lengths. Using syntactic or semantic properties could achieve higher accuracy or better coverage of MWEs. However, such linguistic information is highly domain-, language-, or even type-specific; therefore, significant effort would be required to adapt such information to different types of MWEs.

Several studies have addressed specific linguistic features of MWEs (i.e., compositionality and non-compositionality). Identification of non-compositional (or idiomatic) MWEs is very important for any computational system (Sag *et al*., 2002). Recently, many researchers have considered this feature of MWEs (Lin, 1999; Baldwin *et al*., 2003; Moirón & Tiedemann, 2006). Katz and Giesbrecht (2006) performed latent semantic analysis (LSA) to distinguish whether the meanings of expressions were literal (compositional) or non-literal (non-compositional; idiomatic). They estimated that a vector similarity score between an MWE as a whole and its constituents can represent a degree of compositionality. For example, the similarity score obtained by LSA between the MWE *hit the road* and the single word *leave* is much higher than scores between the MWE and its constituents. However, all methods have advantages and disadvantages. The method proposed by Katz and Giesbrecht (2006) relies on either sufficient non-compositional usage of idiomatic MWEs in the corpus or a bilingual dictionary containing such MWEs for evaluation; however, such information is generally not available. In addition, the performance of this method will be reduced when the (non-)literal meaning is overwhelmingly frequent. Manning and Schütze (1999, Chapter 5) argued that *a mere co-occurrence measure does not well distinguish compositional meaning from non-compositional expressions*; therefore, to achieve better identification of such idiomatic expressions, external linguistic resources should be considered.

Taken together, many rule-based or hybrid identification studies using syntactic or semantic properties of MWEs have shown better results than those obtained using only word statistics. However, a particular approach does not always guarantee a successful result. Adapting existing language-specific resources to other languages, domains, or even different types of MWEs requires considerable time and effort. This adaptation becomes more difficult when dealing with resource-poor language pairs such as Korean–French and Korean–Spanish. Linguistic resources such as parallel corpora for such pairs are very rare. Therefore, this thesis focuses on using either the collocational behavior of MWEs or simple linguistic information such as POS sequence patterns corresponding to noun phrases to identify nominal MWEs. In addition, deeper linguistic processing such as syntactic parsing and linguistic resources such as bilingual dictionaries are ignored. Furthermore, to avoid time-consuming tasks such as building a bilingual dictionary containing idiomatic MWEs for evaluation, non-compositional (or idiomatic) MWEs are not considered.

The general method of identifying MWEs (Seo *et al*., 2014) and other approaches (Daille *et al*., 1994; Piao & McEnery, 2001; Kunchukuttan, 2007) are illustrated in Figure 2.4.

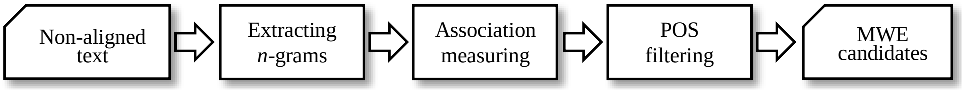


Figure 2.4: General flow of MWE identification

In this method, all possible n-grams ($2 \leq n \leq 3$) from each monolingual corpus are extracted first. Then, reasonable collocations by an association measure such as PMI are extracted. Finally, potential MWE candidates are extracted according to their specific POS sequence patterns. This identification method requires only morphological analyzers and noun phrase patterns for each language, which are readily available for general languages. This identification method is used in the PCAM and is described in the next section. The CTA will be described in Section 6.

2.2.2 MWE alignment

The alignment of MWEs in bilingual parallel corpora is important in NLP domains (Piao & McEnery, 2001). Several approaches based on association measures (Smadja *et al.*, 1996), n-grams, approximate string matching, finite state automata (McEnery *et al.*, 1997), and bilingual parsing matching (Wu, 1997) have been proposed. Piao and McEnery (2001) used n-grams, linguistic POS patterns, and collocation measures together to align nominal MWEs in an English-Chinese parallel corpus. They assumed that nominal MWEs in the source language generally tend to be translated to nominal MWEs in the target language; therefore, their occurrences in the parallel corpus are correlated. Based on this assumption, they align nominal MWEs using their collocational behavior. This approach aligns nominal English/Chinese MWEs with high precision but relatively low recall.

Recently, Seo *et al.* (2014) aligned nominal Korean/French MWEs in terms of collocational behavior of words. They proposed the pivot context-based approach for multi-words (PCAM), which uses a pivot language such as English to bridge the source and target languages. The most important reason why the PCAM considers a pivot language is that linguistic resources such as bilingual (parallel/comparable) corpora or evaluation dictionaries for resource-poor language pairs are very rare and may not be publicly available. Most approaches for extracting various types of bilingual MWEs from parallel corpora (Daille *et al.*, 1994; Wu & Xia, 1994; McKeown *et al.*, 1996) and comparable corpora (Lu & Tsou, 2009) have been based on resource-rich pairs such as English–French or English–Chinese because most of these language pairs (English–*) are readily available online.

The PCAM considers the collocational behaviors of words to select the greatest number of co-related words in source and target languages. As Nazar (2008) stated, co-related words in both languages can be translations of each other, i.e., *if a pair of words co-occurs more often than expected by chance in the aligned pair of sentences of both languages, then it is to expect that they are translations of each other*. To identify potential MWE candidates, the PCAM uses the general method described in Figure 2.4. It assumes that extracted MWE candidates are sufficient to be treated as actual MWEs. In addition, MWE candidates in a pivot language are nonessential because a single word in a pivot language is sufficient to bridge both the source and target words.

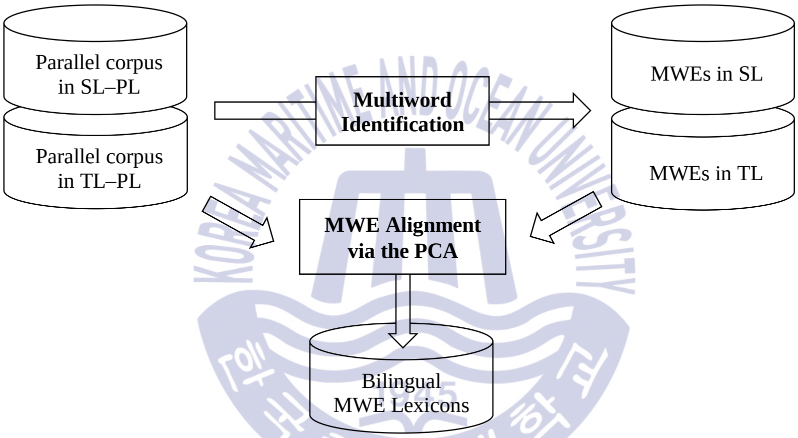


Figure 2.5: General flow of PCAM

After all MWE candidates have been extracted, the actual alignment task, which is derived from the PCA (Seo *et al*., 2013a) is performed with these candidates. The general flow of the PCAM is illustrated in Figure 2.5. The PCAM alignment task will be described in more detail in Chapter 3. Note that the PCA is used for a single word. Nonetheless, it performs well with MWEs because all extracted MWE candidates are transformed into single tokens. These MWE tokens are then added to each input corpus. The PCAM assumes that these modified single tokens, which are added to each sentence in which they occur, can act as single words.

A major drawback of the PCAM is that it can have limited success when common context words in the pivot language are insufficient, which can occur if the domains of two parallel corpora differ or if words do not occur sufficiently frequently to build context vectors in their corpora. Several types of errors stem from the deficiency of common context words. First, translation equivalents are similar with a correct translation but incorrect. For example, when *point de vue* (point of view) is given as a source word, the incorrect translation equivalent 세계관 (*segyegwan*, world view; *vue métaphysique, vision du monde*[3]) is extracted. Second, constituents of translation equivalents as a part are extracted as the top *x* translation equivalents. For example, when 언어학과 (*eoneohakgwa*, department of linguistics) is given as a source word, the translation equivalent *département* (or *linguistique*) is extracted rather than the MWE *département de linguistique* as a whole. This phenomenon can be observed when the context of the MWE (e.g., *département de linguistique*) has insufficient common contexts, but that of its constituents (i.e., *département* or *linguistique*) is much richer. Most such MWEs are infrequent and originally not enough more than their constituents in a corpus whether the multi-word is a high-frequent word.

In Section 6, this thesis proposes the CTA, which focuses on the latter error types mentioned above (i.e., extracting constituents as translation equivalents). Basically, the CTA uses the PCAM; however, it reinforces performance by measuring similarity scores for each target constituent.

2.3 Self-Organizing Maps

A SOM (Kohonen, 1982, 1995) can be a general unsupervised or a competitive learning network. SOMs are used to visualize large amounts of input data in lower-dimensional space. They can be used in pattern recognition (Li *et al*., 2006; Ghorpade *et al*., 2010), signal processing (Wakuya *et al*., 2004), 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 the training of an unsupervised network is entirely data-driven and no target results for the input vectors are provided, a SOM can be used to cluster input vectors and identify features inherent to the problem. It can represent high-dimensional data as low-dimensional map units or neurons, usually as a two-dimensional lattice. A SOM attempts to maintain the topological properties of the input data; therefore, semantically and geometrically similar input vectors are mapped to neighboring units.

[3] French translations of the Korean word are presented in italics.

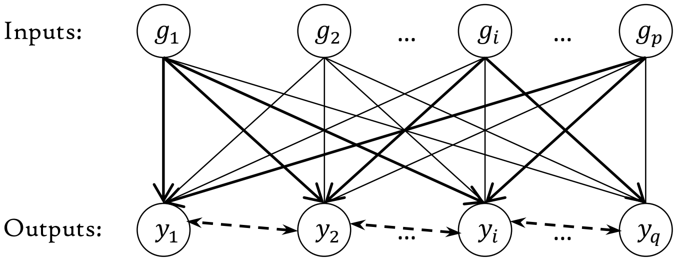


Figure 2.6: Inputs and outputs of SOM model

Figure 2.6 illustrates the overall architecture of a SOM. In this case, the SOM has p training vectors and q units (number of categories). Note that each output unit has its own weight vector of length p to be compared with input vectors. The overall SOM algorithm for training a two-dimensional map can be described as follows.

i. **Initialization**: Set initial weight vectors $w(0)$ with small random values [0, 1]. Let iteration t be 1 and learning rate $\eta(t)$ be a small positive value ($0 < \eta(t) \leq \eta(t-1) \leq 1$).

ii. **Sampling**: Select a sample training input g_i from the input space, where i is the index of the input data.

iii. **Competition**: Find the winning neuron c_g using the minimum Euclidean distance as the identification criterion. The Euclidean distance d between vectors is typically measured using Equation 2.2, where $w_{c,i}$ is a synaptic weight between input g_i and neuron c, and r is the number of neurons in the SOM.

$$d = \|g - w_c\| = \sqrt{\sum_{i=1}^{p} \left(g_i - w_{c,i}\right)^2}, \quad c = 1, \ \ldots, \ r \tag{2.2}$$

The neuron whose weight vector with the minimum value at iteration t, i.e., $c_{\boldsymbol{g}}(t) = \text{argmin}_c \{\|g - w_c(t)\|\}$, wins.

iv. **Updating**: Update weights to all nodes within a topological distance given by $d(t)$ with the update rule given below:

$$w_c(t+1) = w_c(t) + \ \eta(t) \ h_c(t) \left[g_i(t) - w_c(t)\right] \tag{2.3}$$

where $h_c(t)$ denotes the Gaussian neighborhood function kernel around the winner-takes-all neuron c at iteration t. Neighbor weights are updated with values less than that of the winning neuron to preserve the topological characteristics of the map.

v. **Continuation**: Increment t and return to step ii until the weight vector (feature map) stops changing.

When the SOM converges, the weight vector presents crucial statistical characteristics of the input space. The SOM algorithm used in the SOM-based approach will be described in Chapter 5.

2.4 Evaluation Measures

Here, several metrics for evaluating the proposed approaches are described. The metrics are used to determine whether the extracted translation equivalents are correct. The metrics used in this thesis are accuracy (Chatterjee *et al*., 2010), precision (Koehn & Knight, 2002; Chatterjee *et al*., 2010), recall (Haghighi *et al*., 2008), mean reciprocal rank (MRR) (Voorhees, 1999; Koehn & Knight, 2002; Chatterjee *et al*., 2010), and rated recall (Seo *et al*., 2013b). Accuracy and MRR are used when there is one correct answer, while the others are closely related to multiple translations in evaluation dictionaries.

In general, accuracy is defined as the rate of correct translations from the given translation equivalents. However, in this thesis, the top *x* accuracy, that is, the rate of correct source words, is used. This rate is counted when at least one acceptable translation within the top x ranks is discovered. Top *x* accuracy, $\mathrm{Accuracy}_x$, is expressed as follows:

$$\mathrm{Accuracy}_x = \frac{1}{N}\sum_{i=1}^{N} \max_{1\le j\le x} z_{ij}, \qquad \text{where } z_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise} \end{cases}, \tag{2.4}$$

where N denotes the number of evaluated source words s, A_i denotes a set of translations for source word s_i in the evaluation dictionary, t_{ij} denotes the j-th translation equivalent for s_i, and z_{ij} denotes the evaluated translation equivalent t_{ij} (1 or 0).

The MRR is derived from question answering (Voorhees, 1999) and the average of the reciprocal ranks of correct translation equivalents, and it takes the best correct translation equivalent if there are multiple correct equivalents. An MMR value of 0 is used if there is no correct translation equivalent, and greater weights are given to higher ranks than to lower ranks. The top *x* MMR, MRR_x , is expressed as follows:

$$\mathrm{MRR}_x = \frac{1}{N}\sum_{i=1}^{N} \max_{1\le j\le x} z'_{ij}\,, \qquad \text{where } z'_{ij} = \begin{cases} \frac{1}{i} & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise} \end{cases} . \tag{2.5}$$

Precision (also known as positive predictive value), which is the rate of correct translation equivalents within the top *x* ranks, is widely used in information retrieval. In contrast to accuracy, precision evaluates multiple equivalents. In other words, it allows multiple counts. The top *x* precision, $\mathrm{Precision}_x$, is expressed as follows:

$$\mathrm{Precision}_x = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{x}\sum_{j=1}^{x} z_{ij}\,, \qquad \text{where } z_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise} \end{cases} . \tag{2.6}$$

Recall (also known as sensitivity) is also widely used in information retrieval and is defined as the rate of retrieved translations in the evaluation dictionary. Recall at the top *x*, Recall_x, is expressed as follows:

$$\mathrm{Recall}_x = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|A_i|}\sum_{j=1}^{x} z_{ij}\,, \qquad \text{where } z_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise} \end{cases} . \tag{2.7}$$

While recall projects the rate of retrieved translations in the evaluation dictionary, rated recall (RR), which was proposed by Seo *et al*. (2013b), focuses on how many retrieved translation equivalents occur in the corpus. In other words, high-frequency translation equivalents are considered more important than rare equivalents. RR adds the frequency rate of a translation equivalent in the corpus rather than adding 1 if a translation equivalent is correct. In this case, the frequency rate is based on each set of translations A_i for source word s_i. Note that some examples are presented later. RR_x is expressed as follows:

$$\mathrm{RR}_x = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{x} z_{ij}\, r\big(t_{ij}\big), \qquad \text{where } z_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise} \end{cases}. \tag{2.8}$$

Here, $r(t_{ij})$ denotes the frequency rate of t_{ij} in the evaluation dictionary. Tables 2.1, 2.2, and 2.3 show examples of how rated recall works. Here, examples of correct Spanish translations for the Korean word 뜻 (*ddeut*, sense) are presented with some errors. Table 2.1 presents Spanish translations A_i in the evaluation dictionary and their frequencies. As can be seen, the RR contains the meaning in terms of importance (or a weight). Thus, computing averages is again unnecessary in this case. Table 2.2 shows the top *x* translation equivalents for the Korean word 뜻 with correctness. Fortunately, all translations in the evaluation dictionary for 뜻 are retrieved by the system (i.e., Ranks 1, 2, 6, 7, and 11). In this case, recall for 뜻 will be 1 on the top 11 based on the formula (4). Each recall (including this result) is presented in Table 2.3. When only the top 1 is considered, rated recall (0.44) and recall (0.2) express its importance differently. The RR score shows that the translation equivalent retrieved from its corpus is more important than any other equivalents. In other words, the most important translation is retrieved rather than the five top translations.

Table 2.1: Examples of Spanish translations of 뜻 with frequency rate $r(t)$

Translation	**Gloss**	**Frequency**	$\boldsymbol{r(t)}$
intención	intention, intent	3,595	0.44
voluntad	will	2,888	0.36
propósito	purpose, intention	902	0.11
mente	mind	374	0.05
significado	meaning, significance	354	0.04
Total	5	8,113	1.00

Table 2.2: Examples of automatically retrieved translation equivalents of 뜻

Rank	**T. equivalent**	**Gloss**	$\boldsymbol{r(t)}$	**Correct**
1	*intención*	intention, intent	0.44	**True**
2	*propósito*	purpose, intention	0.11	**True**
3	*preparación*	preparation	0.52	False
		…		
5	*consideración*	consideration	0.47	False
6	*voluntad*	will, intention	0.36	**True**
7	*mente*	mind	0.05	**True**
		…		
11	*significado*	meaning, significance	0.04	**True**

Table 2.3: Comparison of rated recall and recall

Rank	**Rated Recall**	**Recall**
1	0.44	1/5 = 0.20
2	0.44 + 0.11 = 0.55	2/5 = 0.40
6	0.55 + 0.36 = 0.91	3/5 = 0.60
7	0.91 + 0.05 = 0.96	4/5 = 0.80
11	0.96 + 0.04 = 1.00	5/5 = 1.00

Chapter 3

Pivot Context-Based Approach

This chapter discusses the use of the pivot context-based approach (PCA) to construct bilingual lexicons efficiently when only resource-poor language pairs are considered. The PCA was developed from the CA (Section 2.1.1). The CA constructs context vectors to present the characteristics of context words by considering contextually relevant words in a small window. However, the CA needs comparable corpora to build context vectors as well as a seed dictionary to translate source words into target words. Unfortunately, there is no publicly available dictionary for some resource-poor language pairs such as Korean–French and Korean–Spanish. In contrast, the PCA uses two parallel corpora with English as a pivot language. Although it relies on fragmentarily on the CA, the PCA does not use an external linguistic resource such as a seed dictionary. Nonetheless, the PCA has shown good performance.

3.1 Concept of Pivot Context-Based Approach

All approaches using a pivot language (Section 2.1.3) combine two existing bilingual lexicons to construct a single lexicon. The performances of these approaches are affected significantly by two distant lexicons (i.e., source–pivot and target–pivot languages). However, the PCA does not use two existing bilingual lexicons. As mentioned previously, it is based on the CA (Section 2.1.1). The PCA does not combine existing bilingual lexicons. The approach does not require direct parallel corpora or comparable corpora; instead, it uses parallel corpora that share one pivot language. Therefore, translation from one to another is unnecessary. Therefore, the PCA can build bilingual lexicons without any external linguistic resources. English words are sufficient to connect both source and target words. This approach is not a language-specific method; therefore, any resource-poor language pair can be considered.

The overall structure of the approach is illustrated in Figure 3.1. Only two parallel corpora are required as input. The sequence of the algorithm can be presented in three steps as follows.

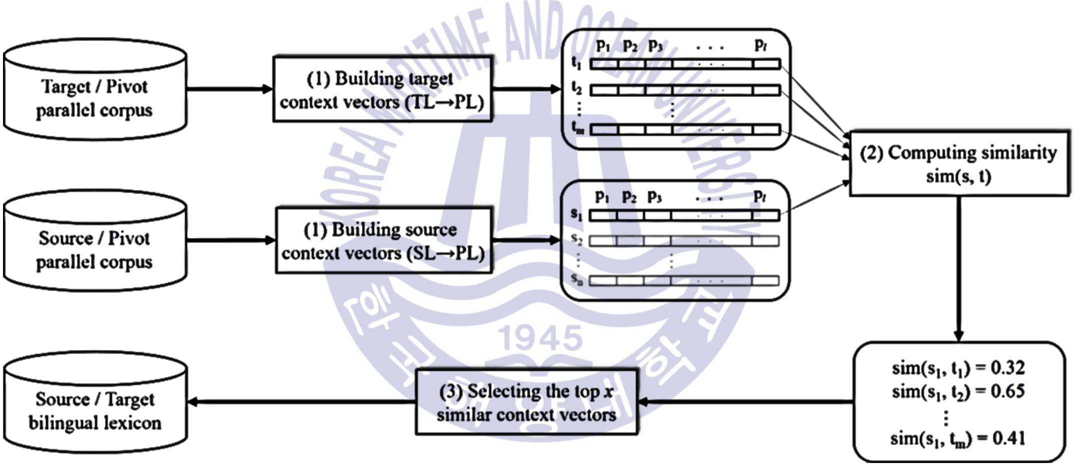
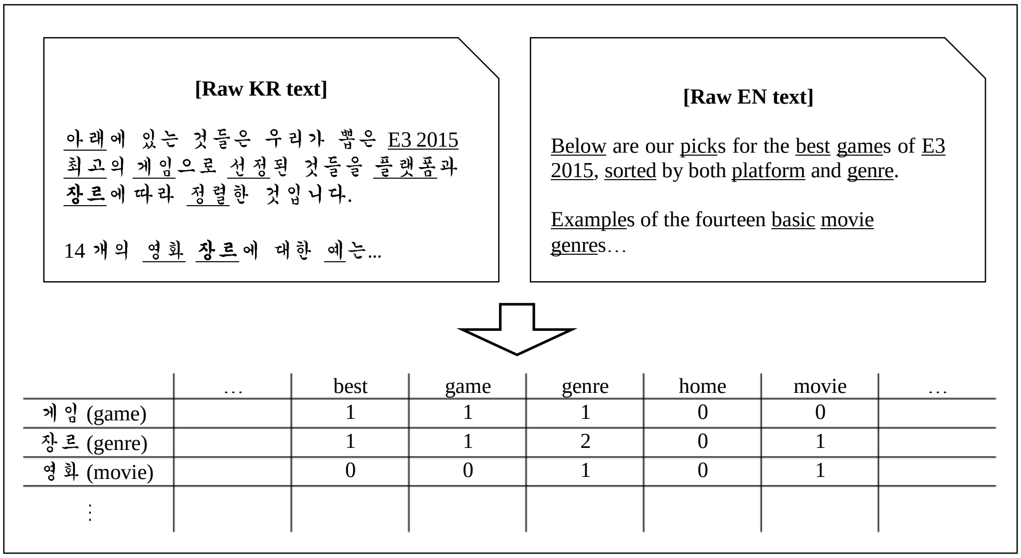

Target / Pivot parallel corpus
(1) Building target context vectors (TL→PL)
P1 P2 P3 . . . Pl
t1
t2
tm
(2) Computing similarity sim(s, t)
Source / Pivot parallel corpus
(1) Building source context vectors (SL→PL)
P1 P2 P3 . . . Pl
s1
s2
sn
sim(s1, t1) = 0.32
sim(s1, t2) = 0.65
sim(s1, tm) = 0.41
(3) Selecting the top x similar context vectors
Source / Target bilingual lexicon

Figure 3.1: Overall structure of the PCA

(1) **Building context vectors**: In this step, context vectors from two parallel corpora are collected separately. The overall flow of building a context vector can be summarized as follows: POS tagging → co-occurrence counting without stop-words → measuring association scores → building vectors based on the scores. The examples presented in Figure 3.2 represent co-occurrences of several words from KR–EN parallel corpora. In these examples, each corpus contains two sentences and each sentence has underlined content words (i.e., nouns, verbs, adverbs, and adjectives). In this work, only content words are considered when words are represented in vector spaces. To identify content words from the corpus, the raw text should be annotated with POS tags. Then, all stop-



[Raw KR text]

아래에 있는 것들은 우리가 뽑은 E3 2015 최고의 게임으로 선정된 것들을 플랫폼과 **장르**에 따라 정렬한 것입니다.

14 개의 영화 **장르**에 대한 예는...

[Raw EN text]

Below are our picks for the best games of E3 2015, sorted by both platform and genre.

Examples of the fourteen basic movie genres…

	…	best	game	genre	home	movie	…
게임 (game)		1	1	1	0	0	
장르 (genre)		1	1	2	0	1	
영화 (movie)		0	0	1	0	1	
⋮							

Figure 3.2: Examples of counting co-occurrences

words such as determiners, punctuation, and cardinal numbers are eliminated, and the word co-occurrences are counted. When these steps are completed, the window size that defines the range of word contexts should be determined. The PCA uses parallel corpora that contain parallel sentences; therefore, the context size is set as one sentence. Consequently, the co-occurrence is the number of parallel sentences containing both the source and target words.

As shown in Figure 3.2, the source word 장르 (*jangreu*[4], genre, bolded) has several co-occurred words in the target languages. The association measure[5] can be computed based on these numbers and additional information such as the number of observed sentences, which is the number of sentences containing source or target words. Of course, various association measures such as PMI, LL, and CHI can be considered here. After all association scores have been calculated, a context vector is constructed with the scores. To calculate the word associations among words in different languages, co-occurrence frequencies should be counted in each parallel sentence.

(2) **Computing similarity scores**: After the context vectors have been built, similarity scores between one source word and all target words are computed. Using the previous

[4] A gloss is shown in italics in parenthesis.
[5] The specific methods to compute association measures are not described in this thesis.

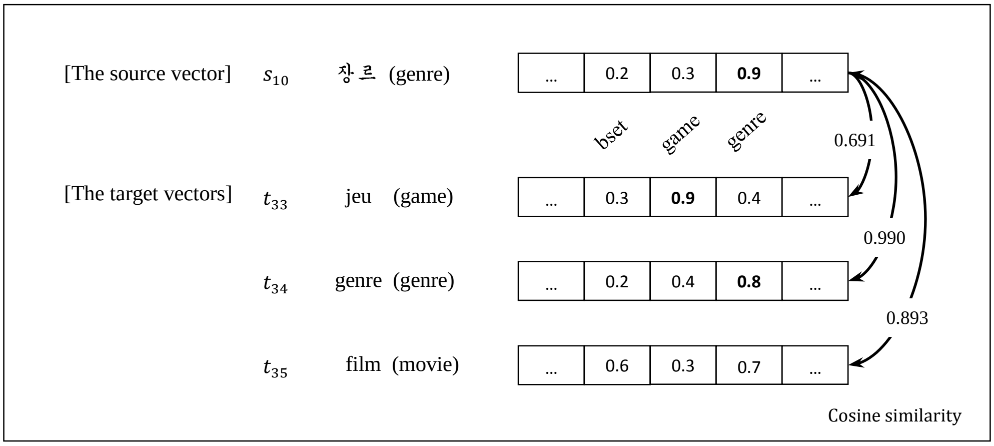


Figure 3.3: Examples of similarity score calculations

example, the source vector for 장르 and all target words should be considered to extract correct translations. More detailed examples of similarity score calculations are provided in Figure 3.3.

As shown in Figure 3.3, similarity scores between source word s_{10} and target words $s_{33} \dots s_{35}$ are computed independently. In this case, the similarity scores can be calculated because the target vectors are also represented in the pivot language. Most measures representing the degree of similarity or difference between two vectors (e.g., cosine similarity or Jaccard coefficient) can be considered (cosine similarity is used in Figure 3.3).

(3) **Selecting similar context vectors**: After all similarity scores for the source word have been calculated, the top *x* candidates are selected and added to the bilingual lexicon. In the experiments described in this thesis, *x* was empirically determined to be 20.

As shown, the overall sequence of the approach appears to be simple. Since parallel corpora share a common pivot language, a seed dictionary is not required. Unfortunately, the approach ignores polysemy problems. Therefore, many heuristic techniques or previous studies to solve the polysemy problem introduced in Section 3.1 can be considered or adapted here.

3.2 Experiments

3.2.1 Resources

Many linguistic resources were used to experimentally evaluate the proposed approach. Three parallel corpora (i.e., Korean–English, French–English, and Spanish–English) were used. The KMU parallel corpus[6] (Seo *et al.*, 2006) for the Korean–English pair, which consists of several bilingual news articles and is aligned at a sentence level, was also used.

The Europarl parallel corpora[7] (Koehn, 2005) for French–English and Spanish–English pairs extracted from the proceedings of the European Parliament were also used. In this work, sub-corpora sampled randomly from the Europarl parallel corpora that contained approximately the same numbers of sentences as the KMU parallel corpus were used to maintain balance with the corpora. The parallel corpora statistics are listed in Table 3.1.

Table 3.1: Parallel corpora statistics

	Korean–English		**French–English**		**Spanish–English**	
Sentences	433,151		500,000		500,000	
Words	8,283,222	13,381,739	13,292,137	12,750,062	13,196,180	12,713,067
Types	1,110,499	374,175	185,815	144,457	210,485	145,531
Avg. words*	19.1	30.9	26.6	25.5	26.4	25.4

* Avg. words is the average number of words per sentence.

As can be seen in Table 3.1, the distributions of the word types and average numbers of words per sentence for the Korean–English pair differ from those for the other language pairs. This phenomenon occurs due to a difference between domains (i.e., news articles and European parliament proceedings). The average number of words per sentence for Korean is lower due to a characteristic of the Korean language. On average, a Korean word usually contains one or more morphemes (2.3 morphemes per word in the experiment). The number of morphemes depends on the domain or corpus). For example, when the Korean POS tags such as NNG: general noun, XSV: verb-derivational suffix, EC: conjunctive ending, VX: auxiliary verb, ETM: adnominal ending, NNB: bound noun, VC: predicative case particle, and EF: final ending, are given, Korean

[6] https://sites.google.com/site/nlpatkmu/Resources/Corpora

[7] http://www.statmt.org/europarl/

tri-grams 부담해야 한다는 겁니다 (*budamhaeya handaneun geopnida*; have to pay) can be tagged as 부담/NNG + 하/XSV + 어야/EC 하/VX + ㄴ다는/ETM 것/NNB + 이/VC + ㅂ니다/EF. In this example, three Korean words are separated into eight morphemes after POS tags are annotated morphologically. Therefore, a Korean word should be separated into several morphemes because it contains one or more meanings.

In addition, two sets of evaluation dictionaries (Korean–French and Korean–Spanish) were compiled semi-automatically for this evaluation. That is, only the grammatical correctness of the collected entries was determined manually. The primary consideration was whether each sense of collected entries was correct and that all translations of source synonyms were included. Each dictionary was unidirectional, and all translations were obtained from the Web dictionary[8]. The Web dictionary contains 100 high-frequency words and 100 low-frequency words. The words were sampled from parallel corpora randomly based on their frequencies. The statistics of their translations in the evaluation dictionaries are presented in Table 3.2. The numbers can be considered as the degrees of ambiguity as well as the numbers of polysemous words.

Table 3.2: Evaluation dictionaries statistics for PCA (per one source word in evaluation dictionaries)

	Korean–French		**Korean–Spanish**	
	KR→FR	**FR→KR**	**KR→ES**	**ES→KR**
HIGH	5.79	10.42	7.36	10.31
LOW	2.26	6.32	3.12	5.49

Before using the parallel corpora, pre-processing was performed. As mentioned previously, a morpheme token is the base unit for Korean, and lemmatized word tokens are the base units for the other languages (i.e., English, French, and Spanish). For these languages, lemmas of word tokens were collected to reduce the sizes of the context vectors. Note that deeper pre-processing such as syntactic or semantic parsing was not necessary; however, morphological analysis with POS tagging was performed. The following tools were used to prepare the input materials automatically. U-tagger[9] (Shin & Ock, 2012) was used to tokenize Korean sentences and induce POS tags of morpheme tokens. For the other languages, TreeTagger[10] (Schmid, 1994) was used

[8] http://dic.naver.com/
[9] http://nlplab.ulsan.ac.kr
[10] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

to lemmatize the word tokens and induce their POS tags. All word/morpheme tokens were annotated and transformed into lowercase letters. The statistics for the preprocessed texts are listed in Table 3.3.

Table 3.3: Preprocessed texts statistics

	Korean–English		French–English		Spanish–English	
Tokens	19,054,681	15,171,888	15,357,708	14,083,616	14,293,198	14,073,076
Types	115,628	218,113	70,749	81,607	104,605	81,782
Avg. tokens	44.0	35.0	30.7	28.2	28.6	28.2

The statistics presented in Table 3.3 differ significantly from earlier statistics in that the distributions of the Korean–English corpus are roughly equal to those of the others because the Korean words were separated into morpheme tokens. Thus, the number (resp. the average number) of tokens for Korean is increased by more than two times from 8,283,222 (avg. 19.1 per sentence) to 19,054,681 (avg. 44.0 per sentence). The numbers of word tokens for the other languages are also greater after pre-processing. Note that repeatedly used suffixes (Eomi in Korean) are separated by words; therefore, the number of word/morpheme types is decreased significantly. The tokenizing/lemmatizing/POS-tagging tasks, eliminated all of the words, except for stop-words and content words (i.e., nouns, verbs, adverbs, and adjectives).

3.2.2. Results

In this section, the experimental results for two language pairs (i.e., Korean–French and Korean–Spanish) are presented. In addition, the PCA is evaluated from several perspectives. Settings for PCA evaluation are as follows.

i. Association measure
ii. High- and low-frequency words
iii. Different language pairs

Figure 3.4 shows four types of accuracy measurements based on different association measures such as CHI, LL, log-odds (LO), and PMI for Korean–Spanish translations when only the top translation equivalent is considered. The association measure was used to build the context vectors (step (1), Section 3.1). For appropriate comparison, the other conditions were fixed. As can be seen, using context vectors based on CHI scores yielded the highest accuracy, 48%. Based on this experiment, the CHI test measures word associations. Of course, this result is not absolute, and selecting an appropriate association measurement depends on various factors such as the languages, domains, and documents.

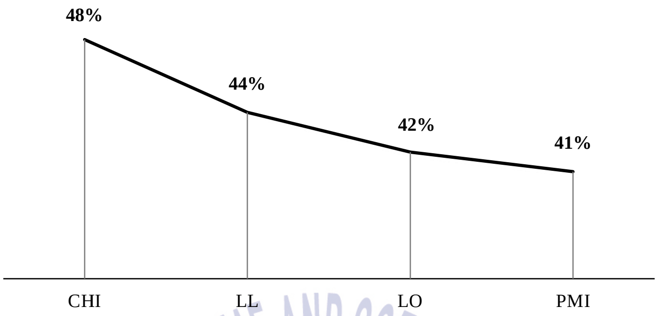
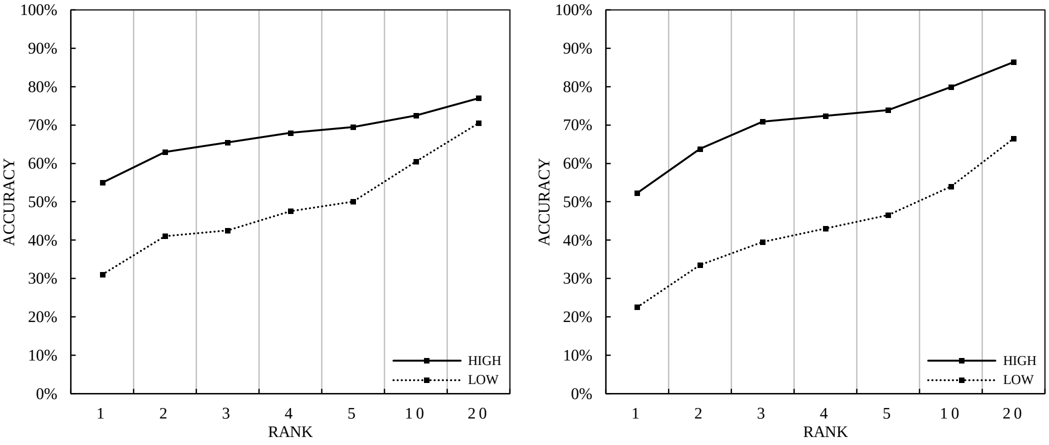


Figure 3.4: Comparison of accuracy measurements by different association measurements

The proposed approach was evaluated by using the CHI test for different language pairs and different word distributions. Figures 3.5 and 3.6 present four accuracies for different language settings. In this work, as mentioned in Section 2.4, accuracy is defined as the percentage of the number of source words that have at least one correct translation within a specified rank. The highest accuracy (88.1%) was obtained for Spanish → Korean translations. The accuracies for high-frequency words at the top 20 in Figures 3.5 (b) and 3.6 (b) show higher performance than the opposite cases. Note that the fact that more translations are included in the * → Korean evaluation dictionaries (Table 3.2) than opposite cases can affect these results. This difference indicates that Korean translation equivalents are more likely to be recognized as correct translations than are Korean → * translations.

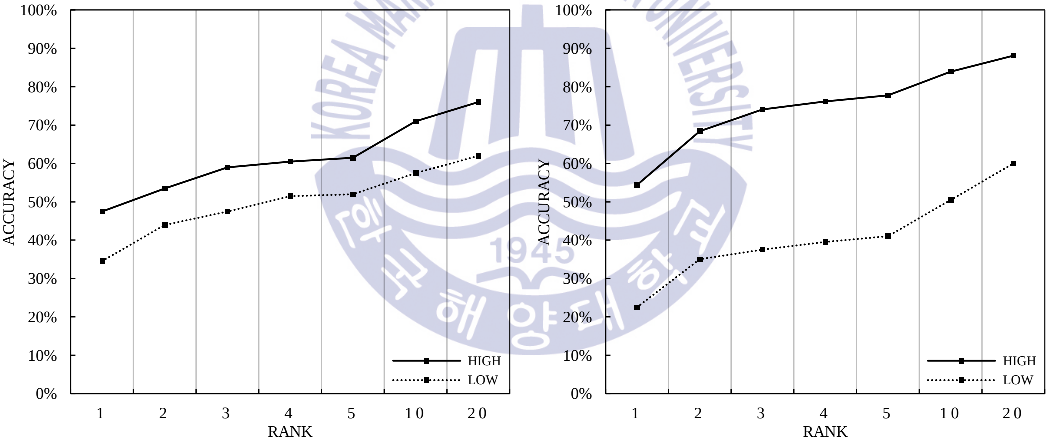
Another characteristic is that, for * → Korean translations (Figures 3.5 (b) and 3.6 (b)), the gaps between high-frequency words and low-frequency words are greater than those for other language settings (Figures 3.5 (a) and 3.6 (a)). In addition, the overall accuracies for low-frequency words within the top 20 are somewhat lower than they are for Korean → * translations. Because low-frequency words have relatively meager contexts, the availability of many more translations in the * → Korean evaluation dictionaries could not affect performance[11].



(a) Korean → French

(b) French → Korean

Figure 3.5: Accuracies for Korean–French translations



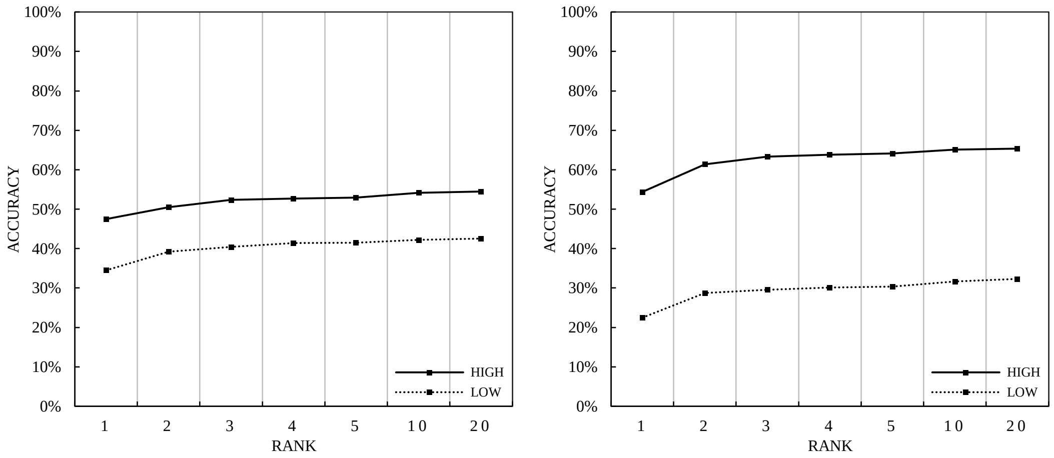
(a) Korean → Spanish

(b) Spanish → Korean

Figure 3.6: Accuracies for Korean–Spanish translations

Figures 3.7 and 3.8 present the MRR results of the proposed approach. As mentioned previously, the MRR takes the best rank when there are multiple correct translations. Although some translation equivalents at the highest ranks among multiple correct equivalents are taken, four MRR results exhibit unimpressive curves. All graphs gently rise over the entire region. In

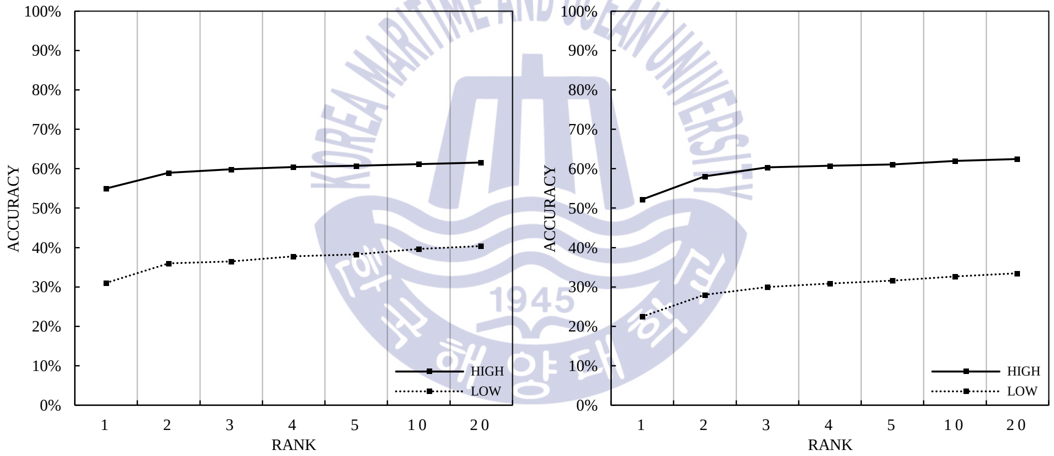
[11] As mentioned in Section 3.2.1, the number of Korean morphemes (opposite to French or Spanish words) is larger than others. That is, more candidates lead to confusion.



(a) Korean → French

(b) French → Korean

Figure 3.7: MRRs for Korean–French translations



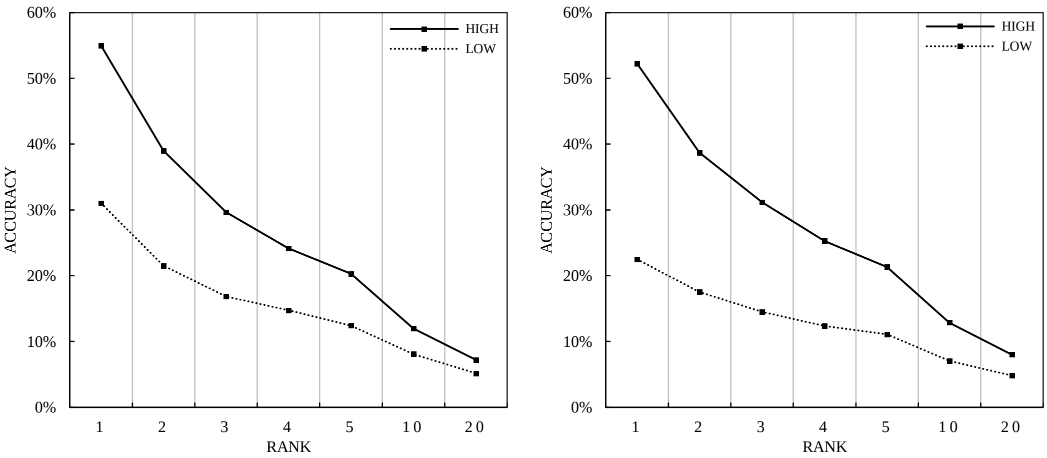
(a) Korean → Spanish

(b) Spanish → Korean

Figure 3.8: MRRs for Korean–Spanish translations

particular, the curves for the top 2 and 3 for all language settings rise very gradually. This trend indicates that most of the correct translations along with the most frequent translations in each corpus are found within the top 2 or 3. In other words, translation equivalents over the top 3 or maybe 5 are very rare. Thus, the proposed approach is promising.

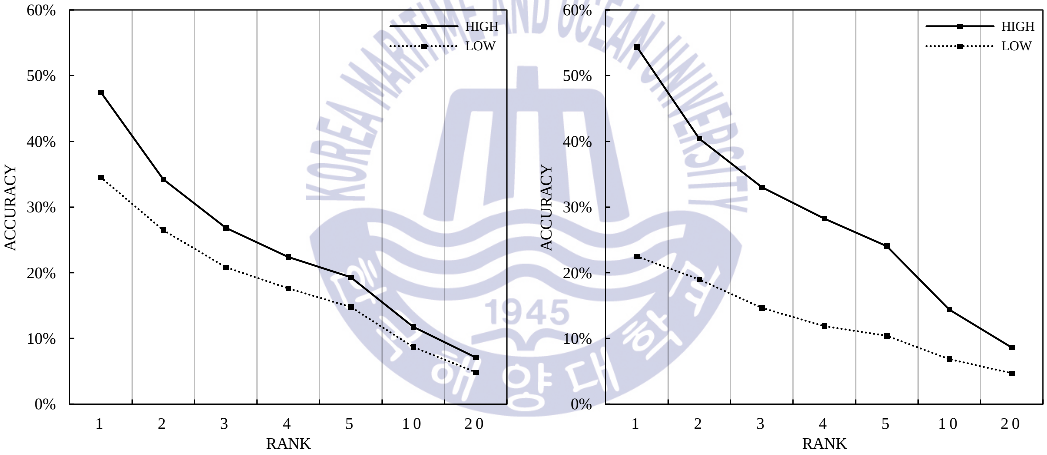
Different characteristics are observed when precision is considered. Since precision is closely related to the number of translations in evaluation dictionaries, the performance of multiple correct translations can be used as a precision measure (Figures 3.9 and 3.10). Except for Figure



(a) Korean → French

(b) French → Korean

Figure 3.9: Precisions for Korean–French translations



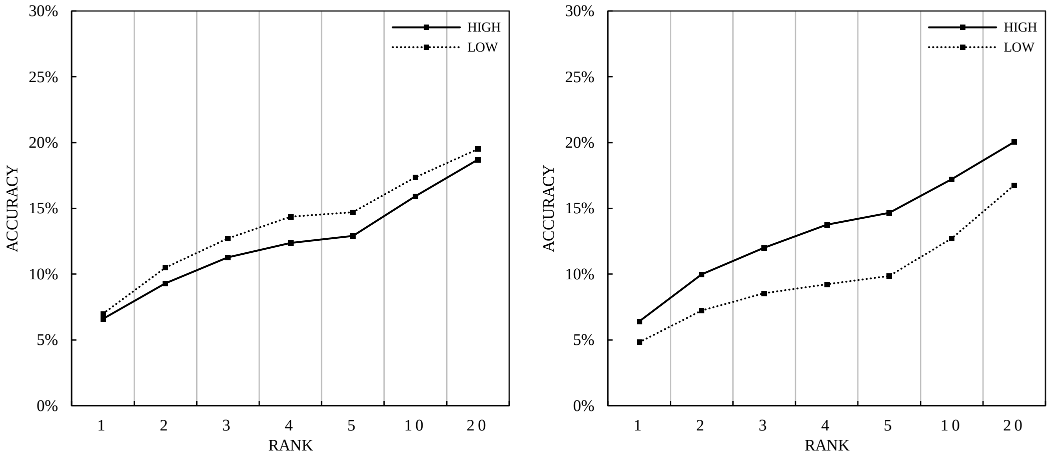
(a) Korean → Spanish

(b) Spanish → Korean

Figure 3.10: Precisions for Korean–Spanish translations

3.9 (a), high-frequency words at the top 1 show accuracies greater than 50%. This characteristic indicates that translation equivalents for over 50% of high-frequency source words can be found by the proposed approach. In contrast, when more translation candidates are considered (i.e., not the top 1 but the top 20), performance decreases. This phenomenon is caused by the rich translations in the evaluation dictionaries (Table 3.2). The proposed approach does not consider homonym (i.e., homograph) problems. Another proof is that the recalls of low-frequency words for * → Korean translations are generally lower than those for Korean → * translations. The results are directly proportional to the distributions in Table 3.2 (i.e., the average number of

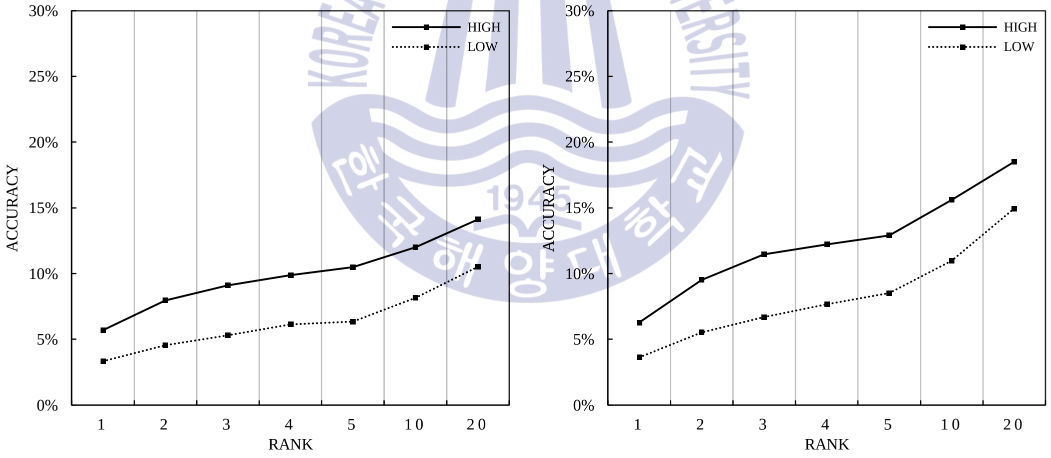
Korean translations is higher than those of the other translations). Therefore, rich translations in evaluation dictionaries were contrary to what was expected.



(a) Korean → French

(b) French → Korean

Figure 3.11: Recalls for Korean–French translations



(a) Korean → Spanish

(b) Spanish → Korean

Figure 3.12: Recalls for Korean–Spanish translations

Nonetheless, multiple translations in evaluation dictionaries are not useless or meaningless. Figures 3.11 and 3.12 present recalls for different language settings. Recall in this work means the percentage of the translations in an evaluation dictionary that are recalled by the system along with the translation equivalents. As can be seen in Figures 3.11 and 3.12, when more candidates are considered, recall is higher. Of course, it is obvious that steeply rising curves cannot be confirmed.

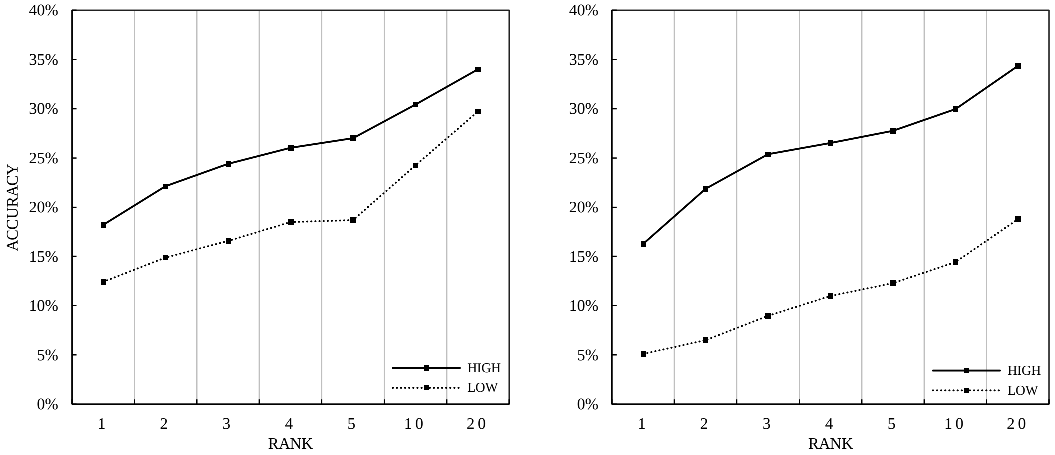
To address this issue, RRs are provided. RRs adapt distributions of recalled translations in their corpora. In other words, an RR reveals the degree to which translation equivalents appear in a corpus. The evaluation dictionaries contain translations that do not appear in the parallel corpora. From this perspective, having proper translations as entries in evaluation dictionaries is meaningless because the translations are not in the parallel corpora. RR reveals this phenomenon and assigns higher weights to more frequent translations. For example, the French word *législation* has several translations (Table 3.4) in the French → Korean evaluation dictionary. Each translation has its own frequency, and its percentage based on its frequency in the corpus is annotated. Thus, each translation can have an independent percentage depending on the source words. RR can reveal how frequent words are recalled/retrieved by adding such percentages rather than adding the integer 1[12].

Table 3.4: Korean translations for French word *législation* in evaluation dictionary

Korean Translations	Frequency	Rated Recall
법 (*beop*, law, rule)	6784	0.847
입법 (*ipbeop*, legislation)	818	0.102
법제 (*beopje*, legislation, the legislative system)	230	0.029
법학 (*beophak*, law, jurisprudence)	120	0.015
입법권 (*ipbeopgwon*, legislative power)	61	0.008
법제론 (*beopjeron*, theory of legislation)	0	0.000
Total	8013	1.000

Regarding the performance at the top 1, all graphs in Figures 3.13 and 3.14 show much better results than those for recall in Figures 3.11 and 3.12. This difference indicates that recalled translations are more often high-frequency words than other translations in evaluation dictionaries. Therefore, regarding the percentages of words occurring in corpora, the proposed approach can

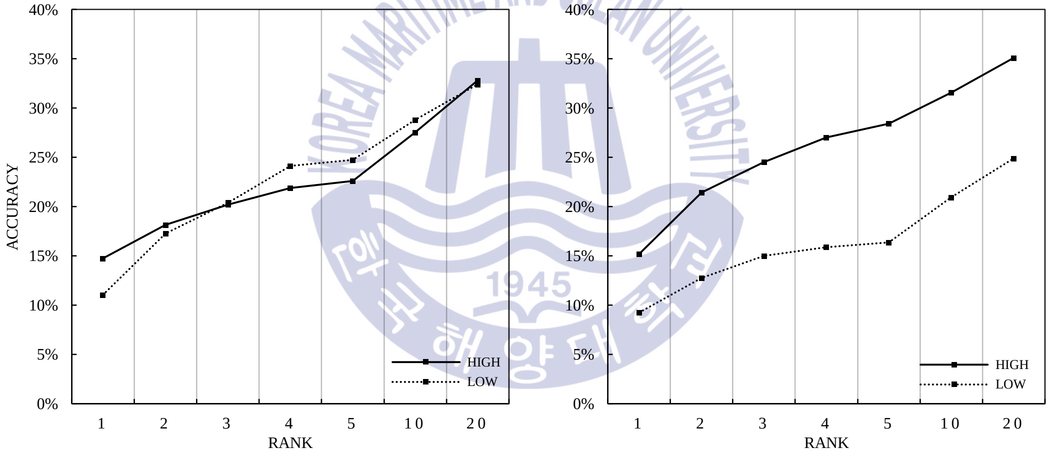
[12] In general, a recall measurement sums integer 1 when a retrieved word is relevant.



(a) Korean → French

(b) French → Korean

Figure 3.13: Rated recalls for Korean–French translations



(a) Korean → Spanish

(b) Spanish → Korean

Figure 3.14: Rated recalls for Korean–Spanish translations

also recall many meaningful translation equivalents.

In summary, the PCA shows the highest accuracy (88%) for Spanish–Korean translations. Through precision scores, translations that occur only several times in evaluation dictionaries can disrupt rather than improve overall performance. Moreover, the proposed approach can also retrieve most high-frequency words in the corpora through MRRs and RRs.

Based on these experimental results, four types of errors are described and examples of errors are provided. The error types can be summarized as follows.

i. Homonyms (i.e., homographs)

ii. Synonyms and weak gold standard

iii. Transliterated words

iv. Word segmentations and compound nouns

First, many homonyms have more than two senses in different concepts. A Korean word is usually derived from hanja-eo (i.e., a compound of a Chinese character word) which consists of its own meaning. In fact, the use of hanja-eo out of Korean in real life is specified to be 66.3% (Heo, 2010). Derivations from hanja-eo result in many ambiguity problems. For example, the Korean word 과장 (*gwajang*, section chief, exaggeration, overstatement) has two different meanings, 課長 (section chief) in Chinese and 誇張 (exaggeration, overstatement) in Chinese. Thus, some translation equivalents with poor contexts could not be retrieved even though other translation equivalents with different contexts are marked as correct. Unless rich word context is considered, the proposed approach did not extract all translation equivalents correctly. However, context is closely related to the domains of the corpora. If the domains of two parallel corpora are the same, the senses of each word are in common use, and this type of problem can be handled.

Second, using weak evaluation dictionaries as gold standards can result in false positive results. For example, the Korean word 당파[13] (*dangpa*, faction) (Table 3.5) has the meaning of the word *faction*. However, this translation is evaluated as incorrect because the evaluation dictionaries only include the synonymous entry 파 (*pa*, group, party, sect, faction). For the same reason, 병력 (*byeongryeok*, troop) is treated as an incorrect answer. The weakness of gold standards results from one of the base limitations of the proposed approach; in bilingual lexicon extraction, there is no publicly acceptable gold standard. Therefore, as mentioned in Section 3.2.1, in this work, practical evaluation dictionaries were constructed semi-automatically (Section 3.2.1) and were therefore incomplete. If synonyms for all words occurring in the corpora are considered as much as possible, then this type of problem can be handled.

Third, there are transliterated words. For example, the Korean word 그룹 (*geurup*, group) (Table 3.5) is retrieved from actual text; however, it does not exist in the gold standard (i.e., the French → Korean evaluation dictionary). This type of error occurs infrequently; therefore, eliminating it or compensating for it either automatically or manually is difficult.

[13] Literally, 당파 (黨派 in Chinese) consists of two characters; 당 (黨, political party) and 파 (派, group).

Finally, there is a word segmentation problem. As can be seen in Table 3.5, both the Korean words (i.e., 집단활동 (*gipdanhwaldong*, group activity) and 자선단체 (*jaseondanche*, charity organization)) should be separated into several words (i.e., 집단활동 → 집단 (*gipdan*, group, mass, organization) + 활동 (*hwaldong*, activity), 자선단체 → 자선 (*jaseon*, charity, philanthropy, benevolence) + 단체 (*danche*, group, organization, party)), because both words are compound nouns. This type of problem can be handled by performing word segmentation in a different way or by exercising human judgement. However, manual methods (i.e., human judgement) are time-consuming; therefore, other methods (e.g., multi-word expression identification or evaluation via a different method) should be considered.

Table 3.5: Examples of translation equivalents for the French word *groupe*

Korean (Romanization)	Gloss	Correct	Type
단체 (*danche*)	group, organization, party	True	right answer
활동 (*hwaldong*)	activity	False	true negative
당파 (*dangpa*)	faction, party	False	synonym
병력 (*byeongryeok*)	troop	False	synonym
그룹 (*geurup*)	group	False	transliterated
집단활동 (*gipdanhwaldong*)	group activity	False	compound word
자선단체 (*jaseondanche*)	charity organization	False	compound word
Korean translations for *groupe* in evaluation dictionary	무리(*muri*, group, crowd), 떼 (*tte*, flock, herd), 집단 (*gipdan*, group, mass, organization), 단체 (*danche*, group, organization, party), 집합 (*giphap*, gathering, meeting, set), 파 (*pa*, group, party, sect, faction), 계열 (*gyeyeol*, affiliation, faction)		

3.3 Summary

This chapter presents the PCA. The PCA builds comparable context vectors from two parallel corpora sharing an intermediary language. The proposed approach constructs two types of context vectors: one from a source–pivot parallel corpus and one from a target–pivot parallel corpus. Then, an association measure such as CHI can be considered to present the vector entries. The experimental results indicate that the CHI method performed better than LL, LO, and PMI. This is a type of language-/domain-specific matters; therefore, it depends on the environment or problem considered. Following this, comparable context vectors are compared using a vector distance measure such as cosine similarity or the Jaccard coefficient. Based on their similarity scores, the most similar *x* translations were selected to be included in a bilingual lexicon.

The most prominent advantage of this approach is that it does not use linguistic resources such as a seed dictionary because it uses parallel corpora sharing an intermediary language. Of course, bilingual lexicon extraction using parallel corpora generally does not require seed dictionaries to translate from one corpus to another. However, such parallel corpora are usually not available to the public. Furthermore, in resource-poor language pairs such as Korean–French and Korean–Spanish the situation is even more serious. However, since two parallel corpora sharing one pivot language such as English are considered, the proposed approach can extract bi-/multi-lingual corpora easily. This idea is very useful even though resource-poor language pairs are considered. Therefore, the proposed method is very attractive when public bilingual corpora between two languages are unavailable but public parallel corpora, e.g., with English as one language, are available.

However, the proposed approach also has a few disadvantages. The first challenge involves homonyms. The proposed approach is based on context vectors; therefore, the impact of related contexts is very important. Consequently, homonyms can result in context vectors that have several types of meanings or contexts. If all the contexts were strong or the domains of the two parallel corpora were the same, the situation would not be problematic. However, if the domains were to differ, the translation equivalents would not be retrieved. Unfortunately, the parallel corpora used in the experiments had different domains (i.e., news articles for Korean–English parallel corpora and European parliament proceedings for French–/Spanish–English parallel corpora). Therefore, this problem was evident in the experimental results. Second, neither rich gold standards to cover most synonyms nor specific evaluation measures to consider false positive translation equivalents exist. To consider these issues, all evaluation dictionaries must be constructed manually by experts, or external linguistic resources such as well-made thesauri are required. For these reasons, evaluation dictionaries should be extended automatically in the future. Third, there are several transliterated words in the corpora. This type of error occurs infrequently; therefore, dealing with it would be difficult. In addition, many compound nouns were not handled by the word segmentation task. Since the number of compound nouns has been increasing, adding all compound nouns to evaluation dictionaries for every language pair would be extremely difficult. Therefore, other evaluation metrics or segmentation skills to handle compound nouns and MWE identification should be considered. Some disadvantages mentioned here are addressed in the following chapters.

Chapter 4

Extended Pivot Context-Based Approach

This chapter presents the extended pivot context-based approach (EPCA). The EPCA combines the PCA (Chapter 3) with the EA (Section 2.1.2) to improve performance. The EA extracts bilingual words from comparable corpora, and its goal is to reduce dependence on initial seed dictionaries. However, the PCA uses parallel corpora to extract such lexicons. Nevertheless, the main idea of the EA is to reinforce context vectors. The experimental results demonstrate that the proposed approach can extract the most proper translation equivalents and increase the ranks of such equivalents.

4.1 Concept of Extended Pivot Context-Based Approach

The EPCA was derived from the PCA to improve performance. As mentioned in the previous chapter, the PCA has some weaknesses (i.e., lack of contexts). To overcome this problem, the proposed approach described in this chapter considers similar context vectors. The basic idea originates from the EA (Section 2.1.2). The core idea of the EA is the assumption that synonyms share the same contexts (Déjean & Gaussier, 2002). In fact, the EA reduces the load of seed dictionaries. However, the PCA qualitatively differs from the EA. The former uses parallel corpora, while the latter uses comparable corpora. Therefore, seed dictionaries are unnecessary when the PCA is used. However, the core idea of the EA can be used to augment the density of context vectors.

The overall flow of the EPCA can be presented in three steps (Figure 4.1).

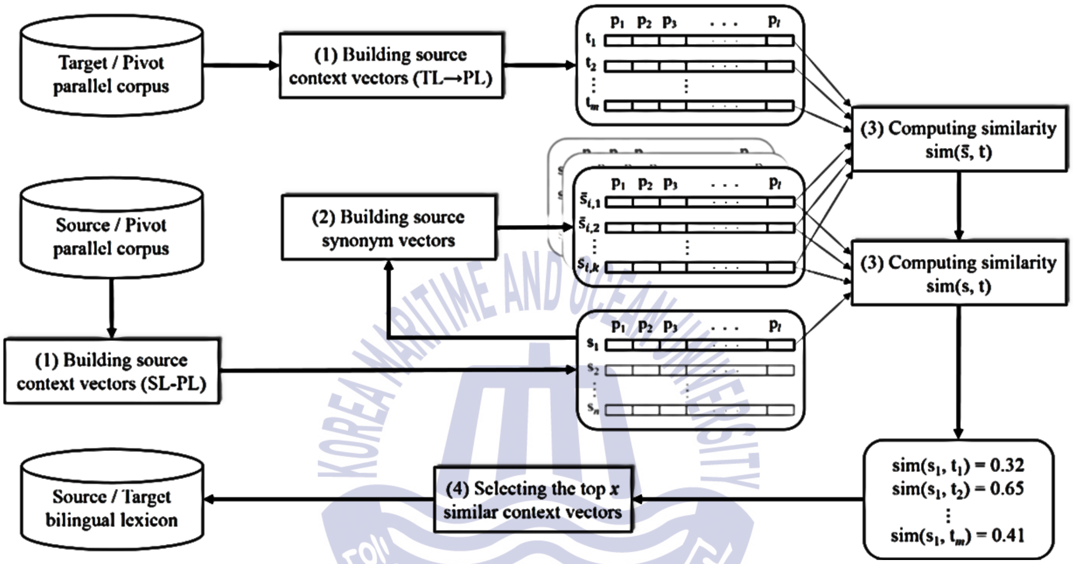

Target / Pivot parallel corpus
(1) Building source context vectors (TL→PL)
P₁ P₂ P₃ . . . Pₗ
t₁
t₂
tₘ
(3) Computing similarity sim(s̄, t)
Source / Pivot parallel corpus
(2) Building source synonym vectors
s̄ᵢ,₁
s̄ᵢ,₂
s̄ᵢ,ₖ
(3) Computing similarity sim(s, t)
(1) Building source context vectors (SL-PL)
s₁
s₂
sₙ
Source / Target bilingual lexicon
(4) Selecting the top x similar context vectors
sim(s₁, t₁) = 0.32
sim(s₁, t₂) = 0.65
sim(s₁, tₘ) = 0.41

Figure 4.1: Overall structure of EPCA

(1) **Building context vectors**: Two types of context vectors are constructed separately from two parallel corpora (i.e., $\vec{s}$ from source–pivot corpus and $\vec{t}$ from target–pivot corpus). This task is exactly the same as step (1) in Section 3.1. All entries of the context vectors are presented in the pivot language and weighted via word association scores between source/target words and pivot words. The CHI method is used to calculate associations among words in different languages.

(2) **Building the nearest source context vectors**: For the source vector $\overrightarrow{s_i}$, its k nearest context vectors $\{\overrightarrow{s_{i,1}}, \overrightarrow{s_{i,2}}, \ldots, \overrightarrow{s_{i,k}}\}$ are collected in this step. For example, let the Korean word 아기 (*agi*, baby) be s_i; then its nearest words could be 젊음 (*jeolmeum*, youth), 젊은이 (*jeolmeuni*, young), and 아이 (*ai*, child). The k nearest vectors are several vectors presenting such nearest words in vector spaces. As mentioned in step (1), all source vectors are weighted with association scores between the source words and pivot words.

By computing similarity scores among source vectors $\vec{s}$ via a vector distance measure such as cosine similarity $\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$, the k nearest context vectors $\overrightarrow{\bar{s}_i}$ that satisfy the threshold condition are identified. Figure 4.2 shows these nearest context vectors.

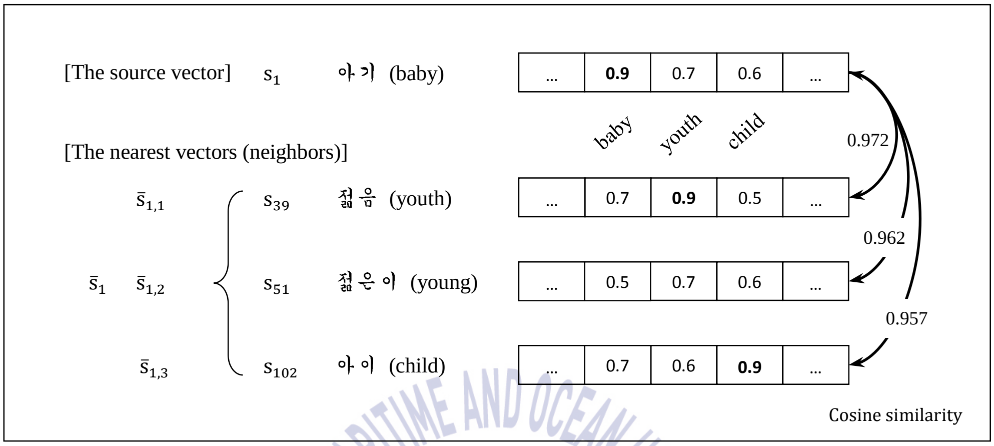


Figure 4.2: Examples of nearest context vectors

In Figure 4.2, several Korean words (i.e., those from the source language) are given and their relations are indexed. The given context vectors show their association scores for three pivot words (i.e., *baby*, *youth*, and *child*). As shown, these scores have the highest values when the source word and component of the vector are closely related. The cosine similarity scores between the source word and its neighbors are also shown in Figure 4.2. These similarity scores demonstrate how closely they are related. The k nearest words for each source word are determined based on these scores.

Basically, this thesis assumes that the collected k nearest words are semantically related and can augment the similarity score between $\overrightarrow{\bar{s}_i}$ and target context vectors $\vec{t}$.

(3) **Computing similarity scores**: After the k nearest context vectors $\overrightarrow{\bar{s}_i}$ for the source word s_i have been collected, the similarity score $\mathrm{sim}(\overrightarrow{s_i}, \overrightarrow{t_j})$ should be calculated. Here, two similarities, $\mathrm{sim}(\overrightarrow{s_i}, \overrightarrow{\bar{s}_{i,k}})$ and $\mathrm{sim}(\overrightarrow{\bar{s}_{i,k}}, \overrightarrow{t_j})$, are considered, where k is the number of nearest context words for s_i. The final score among the similarity scores can be calculated by Equation 4.1:

$$\mathrm{sim}(\overrightarrow{s_i}, \overrightarrow{t_j}) = \sum \mathrm{sim}(\overrightarrow{s_i}, \overrightarrow{\bar{s}_{i,k}}) \times \mathrm{sim}(\overrightarrow{\bar{s}_{i,k}}, \overrightarrow{t_j}). \tag{4.1}$$

Using Equation 4.1, all similarity scores between a source word s_i and all target equivalents t can be computed. As stated previously, the top k nearest context vectors $\overrightarrow{\bar{\mathrm{s}}_{\iota,k}}$ can reinforce the similarity score $\mathrm{sim}\left(\vec{\mathrm{s}_\iota}, \vec{\mathrm{t}_J}\right)$. Here, the similarity scores were calculated by using cosine similarity; however, other measures could also be considered. Note that the similarity score $\mathrm{sim}\left(\vec{\mathrm{s}_\iota}, \overrightarrow{\bar{\mathrm{s}}_{\iota,k}}\right)$ was already calculated, as shown in Figure 4.2. Therefore, the scores $\mathrm{sim}\left(\overrightarrow{\bar{\mathrm{s}}_{\iota,k}}, \vec{\mathrm{t}_J}\right)$ should be computed first in this step.

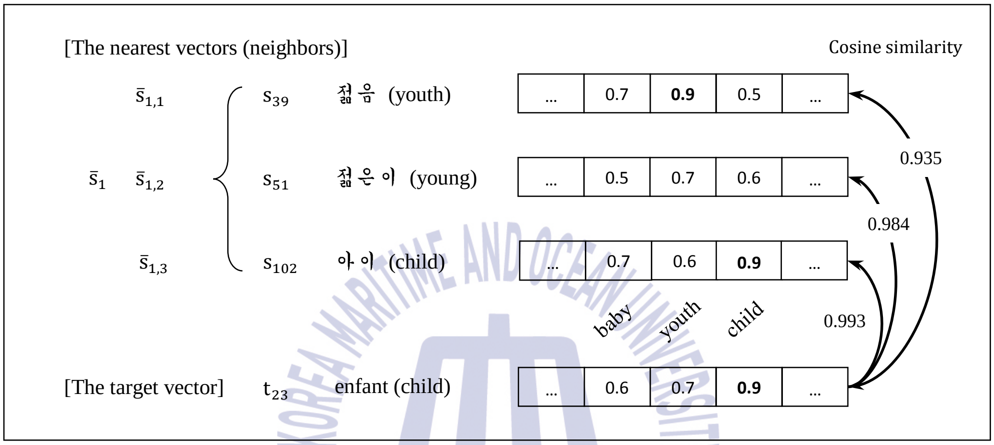


Figure 4.3: Examples of relationship between nearest vectors and the target vector

Figure 4.3 shows the relationships between the k nearest vectors and a target word. Note that all target words should be considered; however, in this example, only one target word is considered. Based on both Figures 4.2 and 4.3, calculating the similarity scores can be represented as shown in Figure 4.4. Here, all similarity scores described, including $\mathrm{sim}(\mathrm{s}_1, \mathrm{t}_{23})$, can be represented as follows.

- $\mathbf{sim}(\mathbf{s_1}, \mathbf{t_{21}}) = \mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,1}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,1}, \mathrm{t}_{21}) + \mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,2}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,2}, \mathrm{t}_{21}) +$ $\mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,3}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,3}, \mathrm{t}_{21}) = 0.972 \times 0.1 + 0.962 \times 0.05 + 0.993 \times 0.1 = \mathbf{0.24}$
- $\mathbf{sim}(\mathbf{s_1}, \mathbf{t_{22}}) = \mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,1}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,1}, \mathrm{t}_{22}) + \mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,2}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,2}, \mathrm{t}_{22}) +$ $\mathrm{sim}(\bar{\mathrm{s}}_1, \mathrm{s}_{1,3}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,3}, \mathrm{t}_{22}) = 0.972 \times 0.1 + 0.962 \times 0.1 + 0.957 \times 0.2 = \mathbf{0.38}$
- $\mathbf{sim}(\mathbf{s_1}, \mathbf{t_{23}}) = \mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,1}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,1}, \mathrm{t}_{23}) + \mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,2}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,2}, \mathrm{t}_{23}) +$ $\mathrm{sim}(\mathrm{s}_1, \bar{\mathrm{s}}_{1,3}) \times \mathrm{sim}(\bar{\mathrm{s}}_{1,3}, \mathrm{t}_{23}) = 0.972 \times 0.935 + 0.962 \times 0.984 + 0.957 \times 0.993 = \mathbf{2.80}$

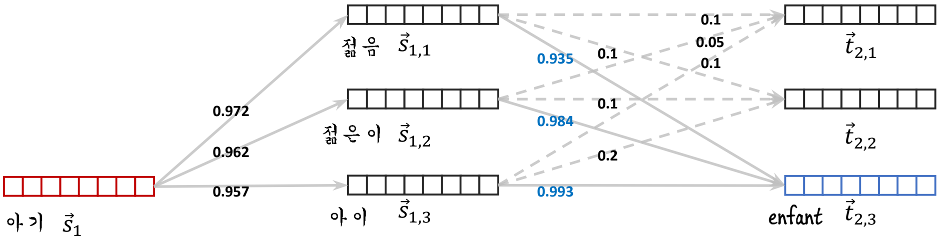


Figure 4.4: Examples of similarity score calculation in EPCA

(4) **Selecting similar context vectors**: After all similarity scores for a single source word and all target words have been calculated, the top *x* translation equivalents with the highest scores are selected and added to the bilingual lexicon as bilingual pairs.

4.2 Experiments

4.2.1 Resources

The same corpora described in Section 3.2.1 were used to evaluate the EPCA. Here, the two sets of parallel corpora (i.e., for the Korean–English pair), the KMU parallel corpus (Seo *et al.*, 2006) consisting of news articles and that for French–/Spanish–English pairs, and the Europarl parallel corpora (Koehn, 2005) consisting of European parliament proceedings are used. All content words (nouns, verbs, adjectives, and adverbs) were POS-tagged, and all stop-words (EN, KR, and FR) were removed (Section 3.2.1).

Because the evaluation dictionaries mentioned in the previous chapter have only 100 entries, the test sets could be considered slightly insufficient. Therefore, the test sets were renewed using other corpora. To cover domain-specific terms, comparable corpora in the new domains[14] were included. The corpora shown in Table 4.1 originated from international news domains dealing with the same events. Articles published over a two-year period (2011.10.28–2013.11.4) were collected. The Korean and French corpora each contain almost 400k sentences, and the Spanish corpus has approximately 270k sentences. The average numbers of words per sentence are relatively larger than those of the parallel corpora discussed in Section 3.2.1.

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

Table 4.1: Comparable corpora statistics

	Korean	**French**	**Spanish**
Sentences	418,474	426,341	268,384
Types	214,484	153,083	112,534
Avg. words/sentence	35.65	32.12	31.89

Consequently, 200 entries were randomly sampled from these corpora based on their frequencies in the corpora. All source entries were required to occur in the parallel corpora so that they could be retrieved by the proposed approach. The collected test sets are compared with those of the parallel corpora in Table 4.2. The average numbers of translations mostly increased, particularly for low-frequency words. The next section discusses the overall experimental results by comparing them with those of the PCA.

Table 4.2: Evaluation dictionaries statistics for EPCA

	Korean–French		**Korean–Spanish**	
	KR→FR	**FR→KR**	**KR→ES**	**ES→KR**
Avg. # of translations (100 entries from parallel corpora)				
HIGH	5.79	10.42	7.36	10.31
LOW	2.26	6.32	3.12	5.49
Avg. # of translations (200 entries from comparable corpora)				
HIGH	8.42	10.79	10.35	12.03
LOW	6.90	7.15	5.43	6.72

4.2.2 Results

This section compares the previously reported results (i.e., for the PCA) to those of the proposed approach in order to demonstrate the performance of the latter. The following figures present several accuracies for various language settings. Translation equivalents within the top 20 ranks are considered in order to observe the overall characteristics.

As shown in Figures 4.5–4.8, there appears to be no significant difference in terms of performance. The performance of the proposed approach decreases slightly in the lower ranks.

However, there is little improvement in the highest rank. The proposed approach augment similarity scores between source word and target equivalents by its nearest words. In addition, it shows the worst performance over the top 10 and better performance within the top 2. In other words, the EPCA performs better than the PCA mostly at the top 1 or 2. Initially, it is difficult to observe the advantages of the EPCA. Thus, the following figures describe the MRR results.

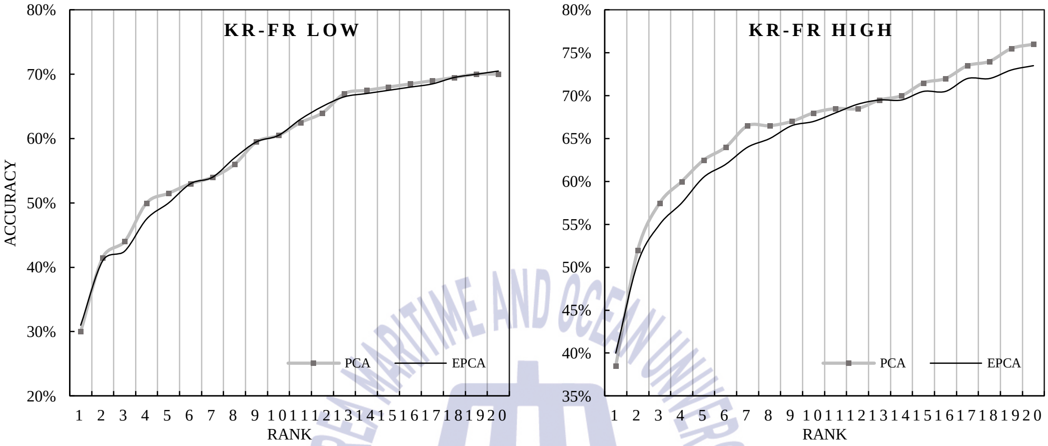


Figure 4.5: Accuracies for Korean → French translations

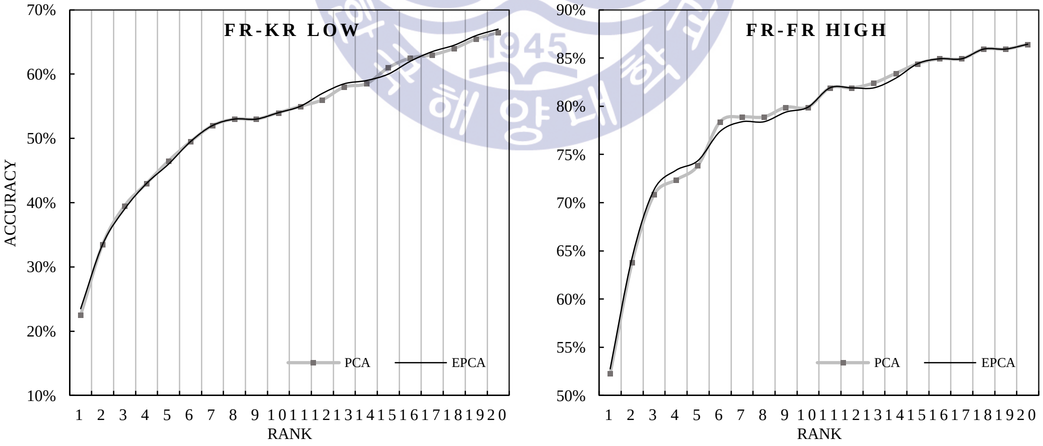


Figure 4.6: Accuracies for French → Korean translations

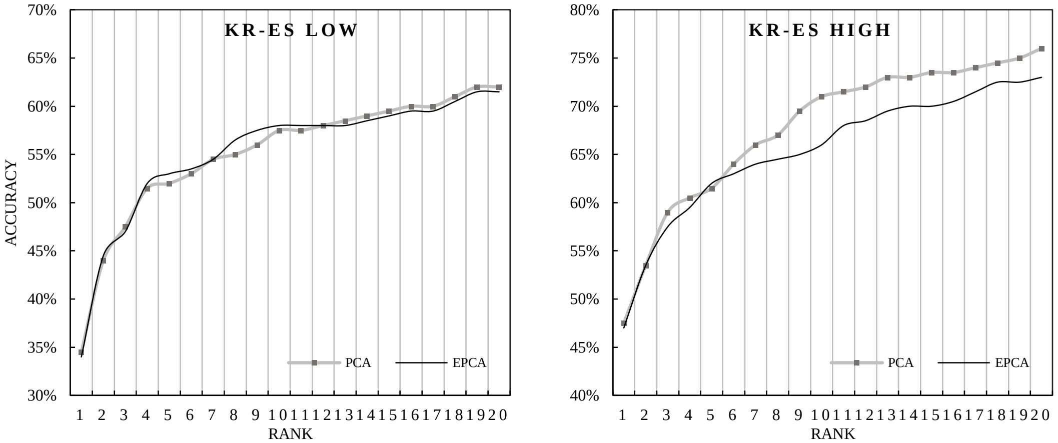


Figure 4.7: Accuracies for Korean → Spanish translations

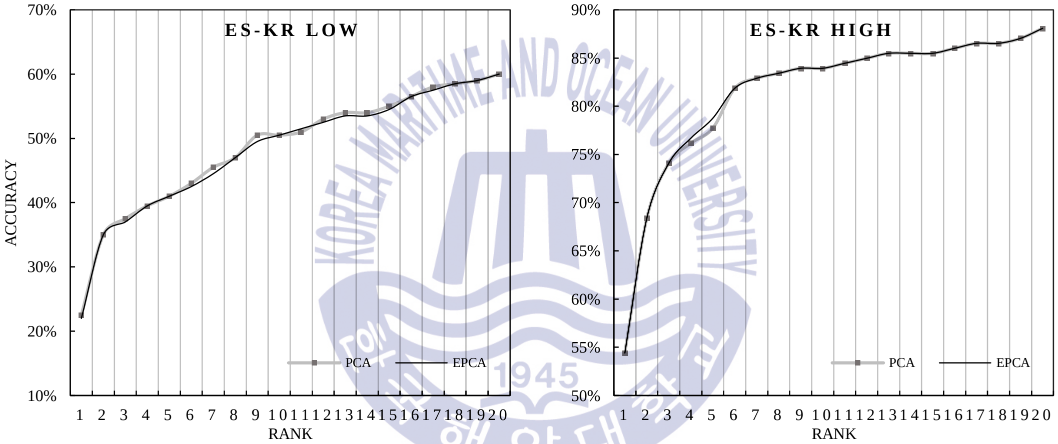


Figure 4.8: Accuracies for Spanish → Korean translations

Figures 4.9 and 4.10 show the MRR results for Korean–French translations, and Figures 4.11 and 4.12 show the MRR results for Korean–Spanish translations. As can be seen in Figures 4.9 and 4.10, the MRR results for Korean–French translations of the EPCA show better performance than the PCA. In fact, the numbers of high rankings increase, while the numbers of correct translation equivalents in low ranks decrease (Figures 4.5–4.8). These characteristics indicate that the proposed approach reinforces the similarity scores of several low rankings.

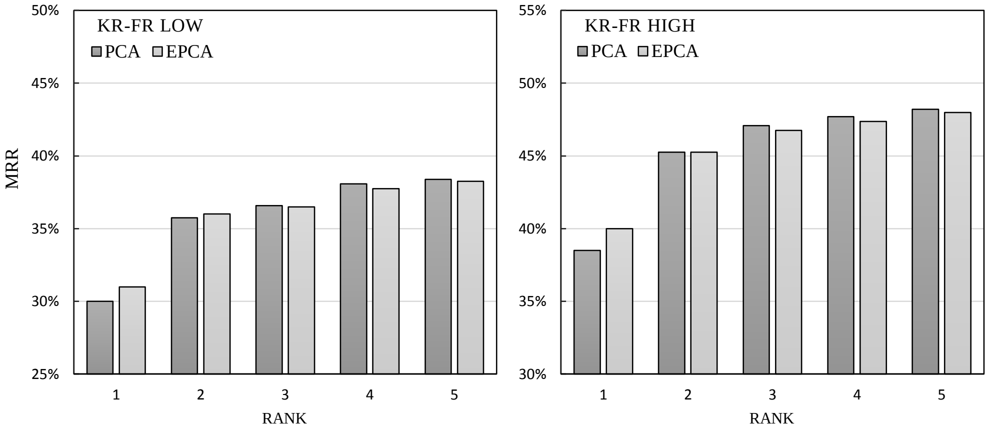


Figure 4.9: MRRs for Korean → French translations

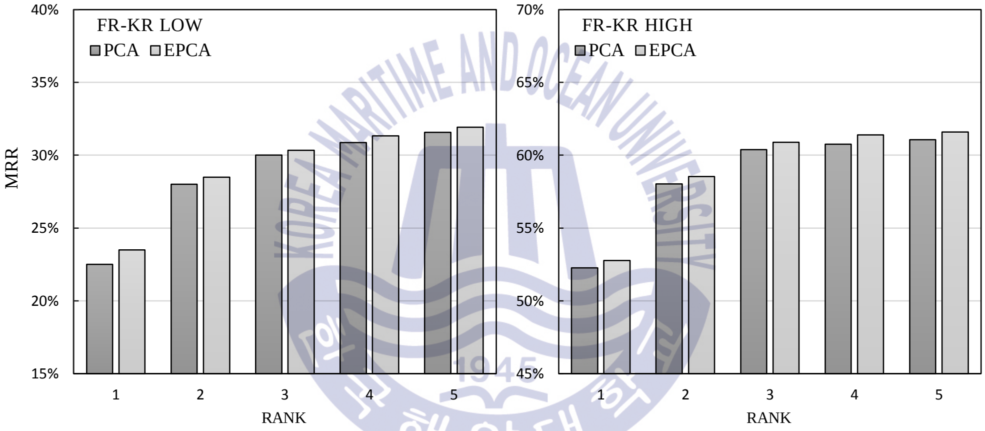


Figure 4.10: MRRs French → Korean translations

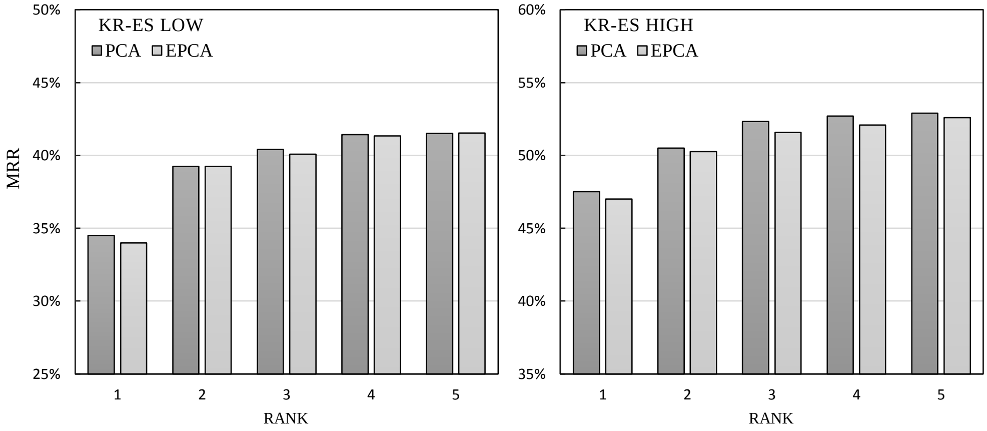


Figure 4.11: MRRs for Korean → Spanish translations

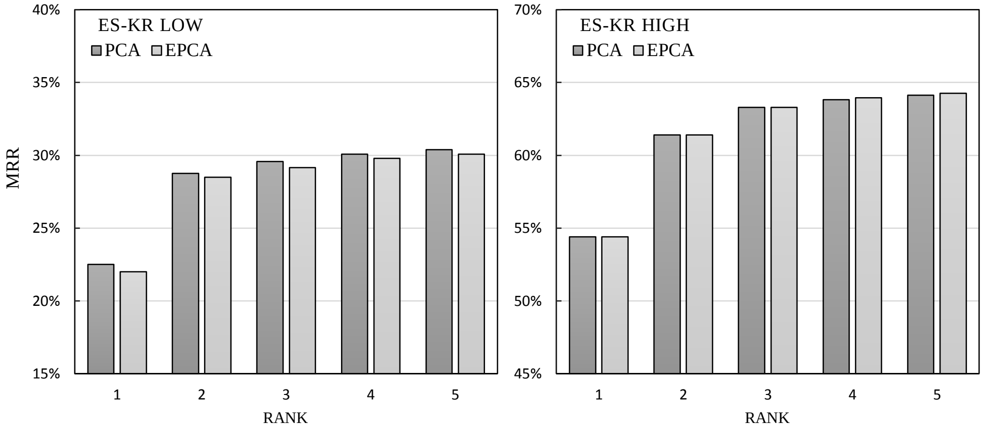


Figure 4.12: MRRs for Spanish → Korean translations

On the other hand, the performance for Korean–Spanish translations is mostly equal or less than that achieved by the PCA (not among the top 5 of the LOW for Korean → Spanish and top 4 and 5 of the HIGH for Spanish → Korean). The proposed approach appears to perform poorly for Korean–Spanish translations. However, in terms of RR, the EPCA is meaningful. With the exception of the translation equivalents at the top 1 of the LOW (Figures 4.13 and 4.14), the EPCA generally shows better performance. Thus, the proposed approach mainly demonstrates lower accuracies for Korean–Spanish translations. In fact, it yields important (or more frequent) translation equivalents with higher similarity scores. However, considered collectively, the performance of the EPCA is lower than expected. Nevertheless, the proposed approach can perform meaningful jobs, particularly augmenting similarity scores supposed to be much lower.

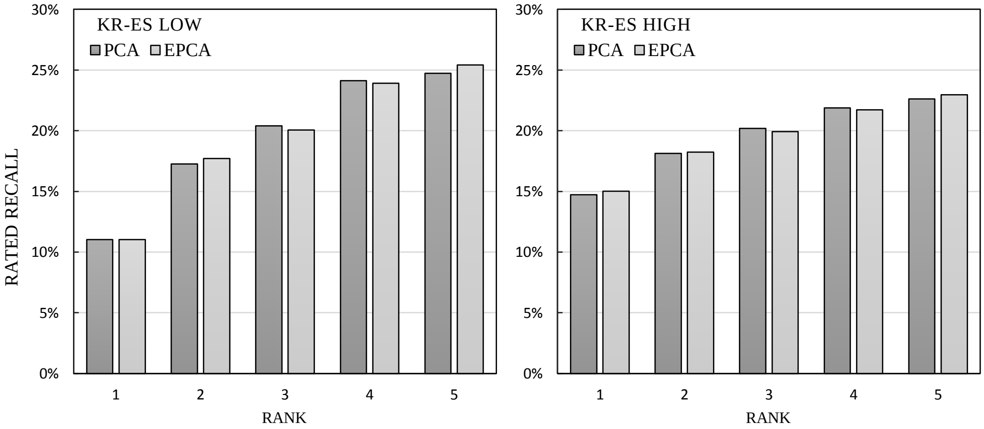


Figure 4.13: Rated recalls from Korean → Spanish translations

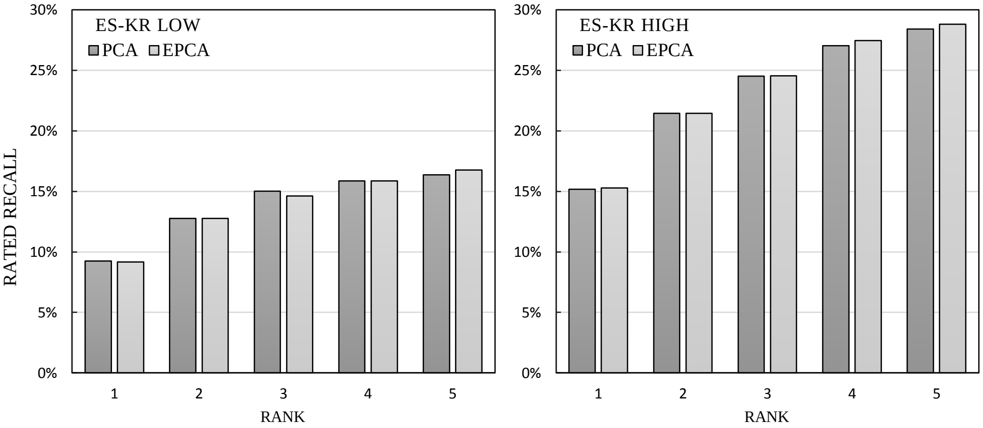


Figure 4.14: Rated recalls from Spanish → Korean translations

Thus, translation equivalents that would occur in low ranks can be retrieved at higher ranks. To improve performance, the qualities of the contexts should be augmented. In addition, the domains of the parallel corpora should be identified.

4.3 Summary

In this chapter, the EPCA was proposed to improve the performance of the PCA. The proposed approach collects *k* nearest context words of source words and adds their similarities and all translation equivalents. While the EA (Section 2.1.2) requires both comparable corpora and a seed dictionary, the EPCA requires only parallel corpora as linguistic resources. Thus, the overall structure of the proposed approach is much simpler than that of the EA. However, the performance is poorer than expected. For Korean–French translations, the accuracy is high. However, for Korean–Spanish translations, the accuracy is somewhat low. Nevertheless, the performance in terms of how important translation equivalents are retrieved is very meaningful. Obviously, meticulous error analysis is required (particularly for Korean–Spanish translations). Furthermore, different domains should be unified unless synonyms from corpora are extended.

The next chapter describes the SOM-based approach, which avoids the strong dependence upon word context. This approach uses SOMs (Section 2.3) and trains feature maps that represent the properties of words to extract bilingual pairs.

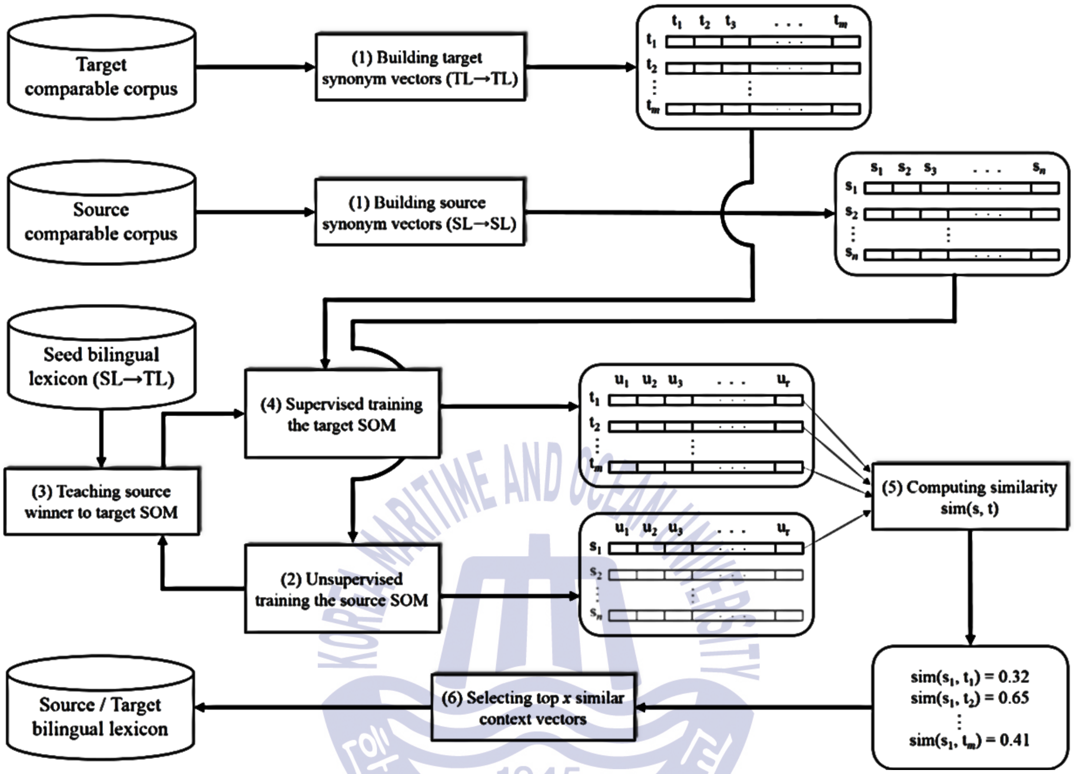
Chapter 5

SOM-Based Approach

This chapter presents an approach that extracts multilingual lexicons using a SOM, which is an artificial neural network algorithm. This SOM-based approach (SA) is very similar to the CA (Section 2.1.1) in terms of its comparison of two types of vectors and use of the same types of linguistic resources (i.e., comparable corpora and seed dictionaries). To estimate the SA, various experiments using Korean–French/–Spanish translations were performed, and the proposed approach demonstrated very good performance.

5.1 Concept of SOM-Based Approach

To apply the SOM algorithm to the main problem (i.e., finding translations in different languages), this thesis assumes that similar words have a common winner (i.e., neuron or unit) and that these words are mapped nearby when semantically or geometrically similar. Based on this property, the SA constructs two types of SOMs (i.e., source and target SOMs) and ensures that two words in different languages have one common winner through the SOMs. Each map is trained in a unique manner; however, they are not necessarily independent. The two different SOMs are trained interactively. Figure 5.1 illustrates the overall flow of the SA.



Target comparable corpus
(1) Building target synonym vectors (TL→TL)
t1 t2 t3 . . . tm
Source comparable corpus
(1) Building source synonym vectors (SL→SL)
s1 s2 s3 . . . sn
Seed bilingual lexicon (SL→TL)
(4) Supervised training the target SOM
u1 u2 u3 . . . ur
(3) Teaching source winner to target SOM
(5) Computing similarity sim(s, t)
(2) Unsupervised training the source SOM
Source / Target bilingual lexicon
(6) Selecting top x similar context vectors
sim(s1, t1) = 0.32
sim(s1, t2) = 0.65
sim(s1, tm) = 0.41

Figure 5.1: Overall structure of SA

The overall structure is organized in six steps.

(1) **Building synonym vectors**: Two types of synonym vectors should be built from monolingual corpora. In this case, synonym vectors are not the same as the nearest context vectors discussed in the previous chapter. Essentially, constructing synonym vectors begins by finding similar context vectors. Figure 5.2 illustrates a difference between the (nearest) context vectors and the synonym vectors from the previous example.

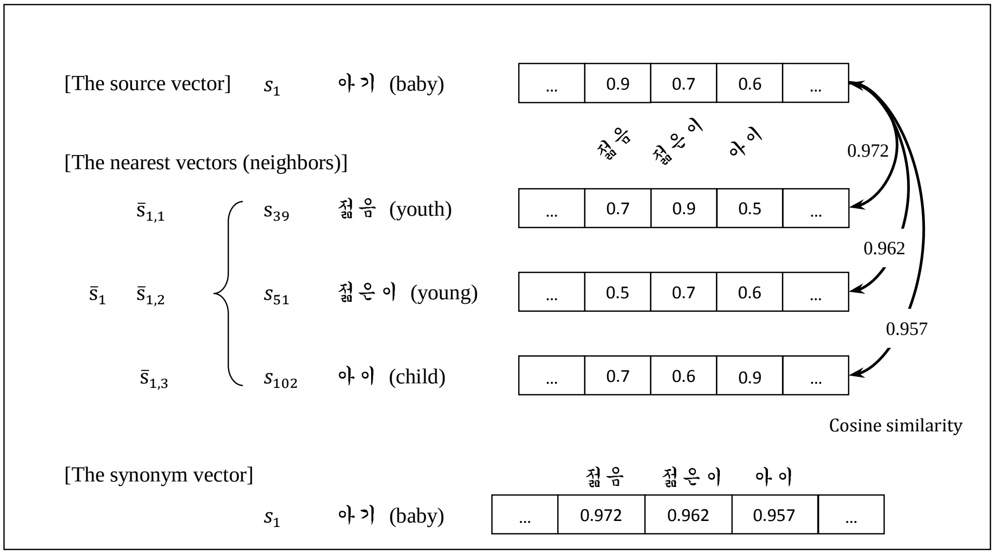


Figure 5.2: Synonym vector examples

As shown, a synonym vector consists of similarity scores among source/target context vectors rather than association scores between source/target words and pivot words. To build synonym vectors, monolingual context vectors (i.e., in a source or target language, a pivot language is not considered) should be constructed first. All entries of the context vectors are weighted by association scores such as CHI values. The method of building context vectors differs from that discussed in Section 3.1. Primarily, the context vectors are constructed from comparable corpora rather than parallel corpora. Thus, the context window size should be adjusted. Based on the empirical results, the context window size was determined to be 5 in this work. After co-occurrences have been counted and all association scores among words have been determined, context vectors are built on the basis of these scores. This step represents the most significant difference between the construction of context vectors from comparable corpora and from parallel corpora (resp. described here and in the PCA). Then, the similarity scores among the context vectors can be calculated by cosine similarity to build synonym vectors. Finally, synonym vectors are composed with these similarity scores. At this point, a specific threshold should be considered to eliminate irrelevant words. This method reduces the dimensions of the vector.

Note that synonym vectors in both the source and target languages are not comparable (i.e., each vector entries indicates different senses). In other words, a seed dictionary should be available when the vectors are compared. These synonym vectors are not comparable; however, two different SOMs (i.e., source and target SOMs) can generate different forms of vectors (i.e., a SOM vector) based on the synonym vectors. To achieve this objective, these synonym vectors should be represented semantically. Consequently, building synonym vectors is very important in the SA.

(2) **Unsupervised training – source SOM**: After synonym vectors are built, all source synonym vectors $\vec{s}$ are taken as inputs for the source SOM. Figure 5.3 depicts the input and output of the SOM (which is also partly included in Figure 5.1).

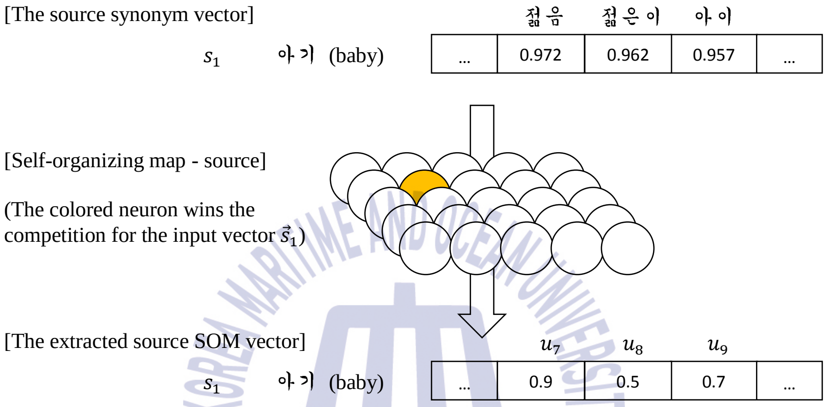

[The source synonym vector]
젊음
젊은이
아이
s_1
아기 (baby)
...
0.972
0.962
0.957
...
[Self-organizing map - source]
(The colored neuron wins the competition for the input vector $\vec{s}_1$)
[The extracted source SOM vector]
u_7
u_8
u_9
s_1
아기 (baby)
...
0.9
0.5
0.7
...

Figure 5.3: Example of building SOM vector

As can be seen, the SOM is trained on the basis of the input synonym vector for s_1. The SOM updates weights by choosing a winner and its neighbors based on their Euclidean distances. The neuron with the minimum score wins the competition for the input. Following this, the weights of the winner and its neighbors are updated by using Equation 2.3. The updating process, in which weights are updated immediately when a winner is chosen, is called an online mode. In addition, these selection and updating processes are repeated until a specific iteration converges. In the experiments described in this work, the source SOM training used unsupervised learning because it updated itself without indicating a winner. After the source SOM has been trained, the SOM vectors can be constructed based on the source SOM. As can be seen in Figure 5.3, entries in the SOM vectors originate from the dot product of two vectors (i.e., an input vector and each weight vector corresponding to every neuron). Note that the meaning of each entry is not estimated. In addition, the source SOM vectors should be constructed based on a well-trained source SOM.

(3) **Teaching source winners to target SOM**: After the source SOM has converged at a specific iteration, well-trained weight vectors in the source language are preserved to train the target SOM with specific winners. The specific winners correspond to source words included in the seed dictionary. If a source word (i.e., an input sample of the source SOM) is included in the seed dictionary, its winning neuron is preserved for training its translation in the dictionary in the next step.

(5) **Supervised training – target SOM**: In this step, the target SOM is trained in a supervised manner. If a target word is contained in the seed dictionary as a translation of a source word, the target input is learned based on the winner of the parallel source word in a supervised fashion. This process is illustrated in Figure 5.4.

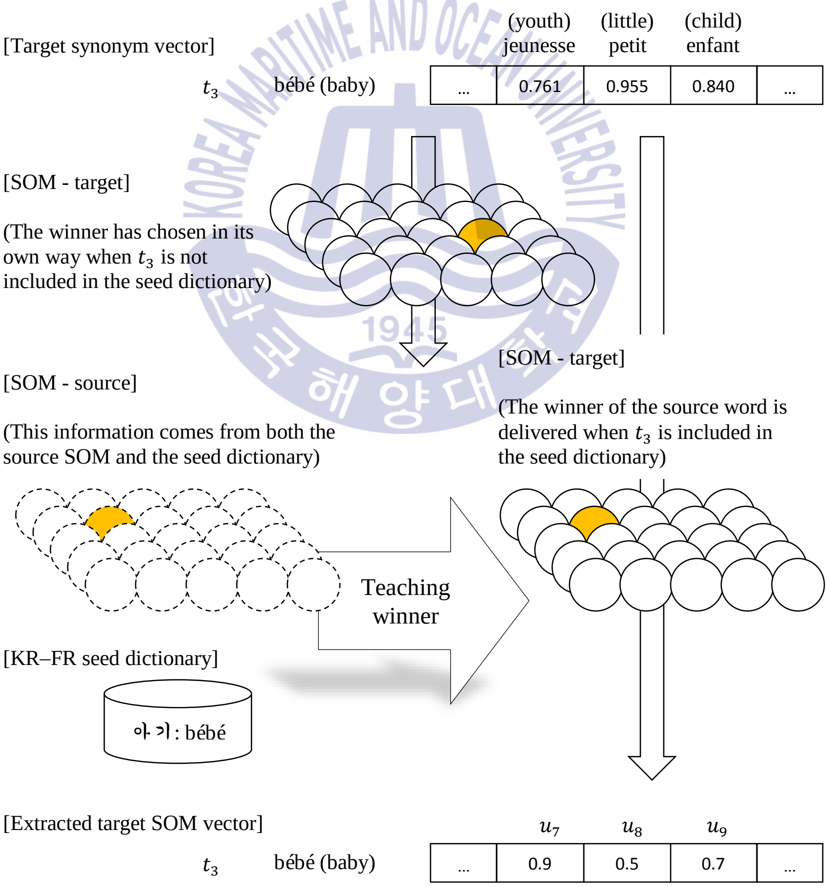

[Target synonym vector]
(youth) jeunesse
(little) petit
(child) enfant
t_3
bébé (baby)
…
0.761
0.955
0.840
…
[SOM - target]
(The winner has chosen in its own way when t_3 is not included in the seed dictionary)
[SOM - source]
(This information comes from both the source SOM and the seed dictionary)
[SOM - target]
(The winner of the source word is delivered when t_3 is included in the seed dictionary)
Teaching winner
[KR–FR seed dictionary]
아기 : bébé
[Extracted target SOM vector]
u_7
u_8
u_9
t_3
bébé (baby)
…
0.9
0.5
0.7
…

Figure 5.4: Example of determining winner from SOM vector

As shown in Figure 5.4, the target SOM can be affected by the source SOM. If target words are in the seed dictionary, the target SOM is not updated in a natural way. Accordingly, target topologies of the map increasingly resemble those of the source SOM. This approach assumes that two words in different languages have the same winner when they are translations. Furthermore, this feature is the key characteristic of the SA.

(6) **Computing similarity scores**: After SOM vectors have been built, similarity scores between one source SOM vector and all target SOM vectors are computed. This step is exactly the same as step (2) in Section 3.1.

(7) **Selecting similar context vectors**: After all similarity scores for the source word have been calculated, the top *x* candidates are selected and added to the bilingual lexicon. This step is exactly the same as step (3) in Section 3.1.

As shown, generating two different words to face one winner is the most important issue in this approach. If both SOMs are well trained, they and their semantic neighbors can be located in the same position of the SOM. Several experimental settings and results using the proposed approach are presented in the next section.

5.2 Experiments

5.2.1 Resources

In this chapter, the proposed approach is evaluated using the same language pairs as in the previous experiments (i.e., Korean–French and Korean–Spanish). For comparison, the CA discussed in Section 2.1.1 is implemented as the baseline.

Two types of linguistic resources were used to analyze the proposed approach. First, three comparable corpora (Kwon *et al.*, 2014) (i.e., Korean, French, and Spanish) were used. Each corpus contained 800k sentences from the Web. The Korean corpus consisted of news articles combined with other news corpus (Seo *et al.*, 2006), and the others consisted of either news articles or European parliament proceedings (Koehn, 2005). The statistics of the comparable corpora are described in Table 5.1. The table presents the statistics of news articles both before and after being combined with comparable external corpora. As can be seen, newly gathered news articles contain greater numbers of average words per sentence.

Table 5.1: Combined comparable corpora statistics

	Korean	French	Spanish
Before combined (only news articles)			
Sentences for news	418,474	426,341	268,384
Word types	214,484	153,083	112,534
Avg. words per sentence	35.7	32.1	31.9
After combined (news articles with non-category news & Europarl corpora)			
Sentences	800,000	800,000	800,000
Word types	281,026	179,389	184,963
Avg. words per sentence	16.2	15.9	16.1

As mentioned previously, nouns were the focus of the investigations. Table 5.2 presents the total frequencies of words in each corpus and the distributions of nouns. As can be seen in Table 5.2, the rates of nouns for each corpus can be predicted. A striking point is that Korean nouns comprise approximately one-third of the total words. This rate is a well-marked difference between the corpora.

Table 5.2: Statistics of nouns and their frequencies

	Korean	French	Spanish
Word types	281,026	179,389	184,963
Total word frequency	33,067,681	28,793,031	22,750,343
Noun types	192,268	46,643	58,324
Noun frequency	10,268,456	5,795,622	4,743,043

This thesis defines sections of word frequencies to determine the effect of seed dictionary size. This is presented in Table 5.3. For Korean, 11,910 of 192,268 nouns are contained in 95% of the total words in the corpus. That is, 180,358 nouns (i.e., approximately 94% of nouns) are contained in only 5% of the corpus. Only 6% of nouns (17.4% for French and 12.8% for Spanish) are high-frequency words; the rest are extremely rare (i.e., low-frequency words) in the corpus. There are two issues. The first involves theme unity, which can be lessened by including many different news article subjects even though the documents come from common keywords. Many articles deal with various themes; therefore, a great variety of rare nouns can be collected, for example, hapax-legomenons, neologisms, compound nouns, and transliterated words. The second issue relates to errors in a corpus, for instance, segmentation errors and POS-tagging errors such as named entities annotated as general nouns. These issues occur more frequently in the Korean comparable corpus than in the others.

Table 5.3: Number of seed words in each interval (except 200 high-frequency test words)

Frequency intervals	**85%**	**90%**	**95%**
Korean	2,536	4,794	11,910
Korean–French	296	787	2,399
Korean–Spanish	563	1,388	4,387
French	1,833	3,404	8,105
French–Korean	388	835	2,138
Spanish	1,805	3,240	7,458
Spanish–Korean	345	736	1,813

The other striking point is that very few entries are actually extracted in seed dictionaries because the true translations do not appear in their corpora. Thus, 11,910 to 2,399 entries (i.e., source words) for Korean–French translations (4,387 for Korean–Spanish translations) were extracted. Based on these numbers of nouns, both seed dictionaries and evaluation dictionaries were built. Each evaluation dictionary contained 200 source words. The statistics of the evaluation dictionaries are presented in Table 5.4.

Table 5.4: Evaluation dictionaries statistics for SA

Language pairs	**Korean-French**		**Korean–Spanish**	
Source language	**KR**	**FR**	**KR**	**ES**
# of source words	200	200	200	200
# of translations	447	209	456	509
# of translation types	420	189	369	421

As can be seen in Table 5.4, there are several duplicate translations in the evaluation dictionaries, indicating that no heuristic process to make each source word have a unique sense, winner, or translation was used in advance. This thesis assumes that one translation can be a part of two different source words even if the words are similar. Finally, the accuracies of the proposed approach were methodically evaluated on the basis of these resources.

5.2.2 Results

Unfortunately, there are no publicly accepted gold standards or experimental guides. In fact, the best performance depends on the experimental settings, including the languages, document domains, seed dictionaries, and so on. Above all, the input samples and the relationship between seed words and evaluation words are the most important factors that determine the quality of the results. Input samples should semantically related, i.e., synonyms can be extracted on the basis of the synonym vectors (Section 5.1). Alternatively, the relationship between seed words (i.e., training data) and evaluation words (i.e., test data) is also very important. They should be close in the vector space to retrieve correct translation equivalents. However, the similarity scores between these words were not considered in this study; only their frequencies in the corpora were considered.

Since all parameters such as the learning rate, Gaussian functions, epochs, and SOM size cannot be tuned simultaneously, the following three experimental settings were selected.

i. Size of SOMs
ii. Epoch
iii. Comparisons with the base approach (i.e., the CA)

In this study, the most efficient learning rates were briefly investigated. Figure 5.2 presents the accuracies according to learning rates in the top 20 for French → Korean translations where the size of the SOM is 300 and the epoch is 2000. Based on this result, the learning rate is fixed at 0.1 in the following discussion.

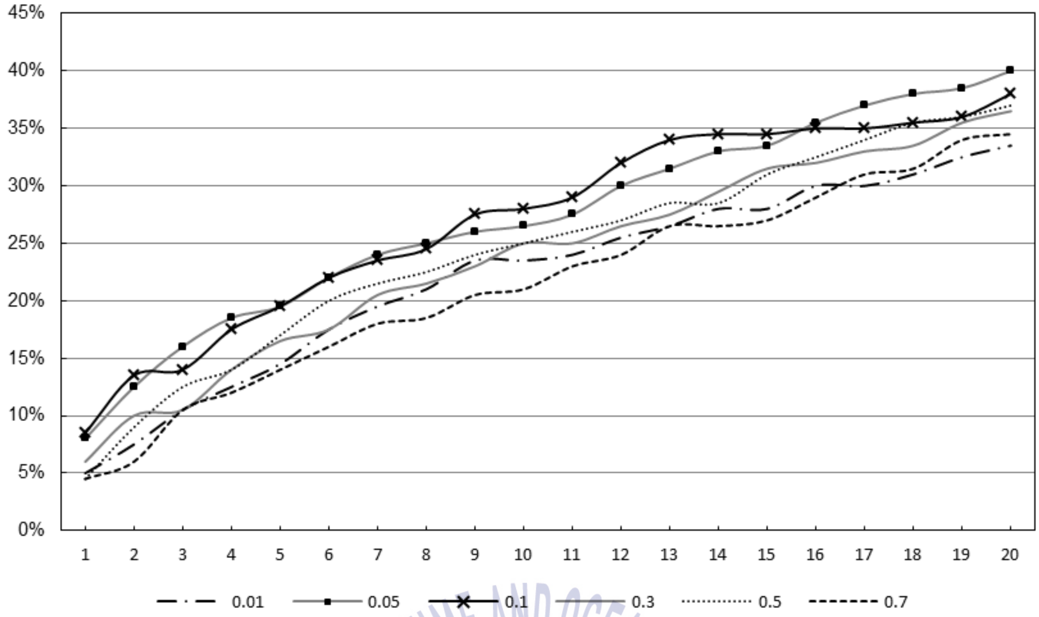


Figure 5.5: Accuracies according to learning rates (x-axis: rank, y-axis: accuracy)

To demonstrate the differences in performance according to the SOM size, the proposed approach was evaluated for 200 high-frequency words, where the size of the seed words was 90% (787 source words; Table 5.3) for Korean–French translations and the epoch was fixed at 2000. In addition, the learning rate was fixed at 0.1, and the Gaussian size was fixed at 25 (5×5).

As can be seen in Figure 5.3, the biggest SOM does not always yield the best performance. On average, sizes of 600 and 800 show reasonable performances within the top 10. Over the top 10, sizes of 700 and 800 exhibit adequate performances on average. Based on these results, it is difficult to see direct or inverse proportions in these experiments. In addition, it is difficult to observe a direct relationship between the SOM size and the Gaussian function. In order to check this feature, various experiments using different Gaussian functions should be performed. Unfortunately, the objective of the work described in this thesis was not fine-tuning to achieve the best performance. In this sense, the main conclusion is that too enough size of SOMs is rather wasteful. Furthermore, the Gaussian function should be modified to handle larger SOMs. The sizes of the SOM and Gaussian function have some specific relationships that have not been determined yet (these relationships will be considered in future work). However, it is known that this relationship depends on data in some manner.

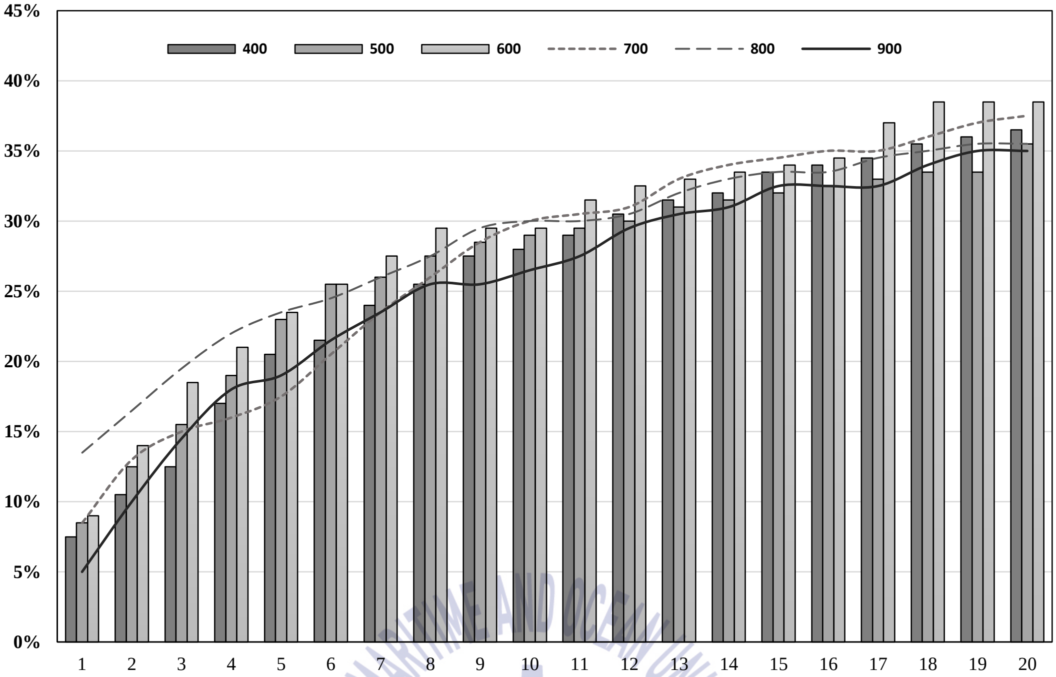


Figure 5.6: Accuracies according to SOM size (x-axis: rank, y-axis: accuracy)

In this work, an epoch is defined as the moment in time at which every source word (resp. target words) participates in the source (resp. target) SOM training. If sufficient epochs are given, every single weight must be converged. This section focuses on how the epoch affects the performance.

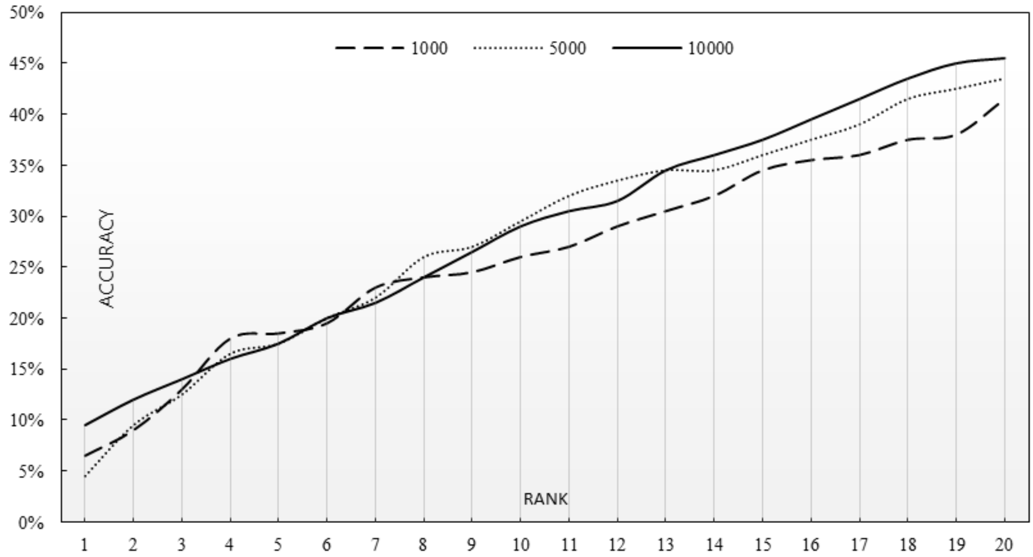


Figure 5.7: Accuracies according to epochs (x-axis: rank, y-axis: accuracy)

Figure 5.4 shows accuracies according to different epochs. This experiment was conducted for Spanish → Korean translations (SOM size: 300, learning rate: 0.1). As can be seen in the figure, higher performance on average is obtained when more epochs are considered. Of course, this performance depends on the data. Therefore, adaptive tuning tasks should be considered to achieve optimum performance.

Figures 5.5–5.16 describe the accuracies for two sets of language pairs, Korean–French and Korean–Spanish, where different numbers of nouns are considered. Each percentage is the percentage of word frequency in a corpus. For example, 90% indicates that 90% of all words in a certain corpus (duplication is allowed) are considered, and all nouns of the 90% are considered to be seed words. Therefore, seed words for the same percentages actually present different forms according to the given corpus.

As can be seen in Figures 5.5–5.16, the SA outperforms the CA when the same linguistic resources are considered. Based on these results, it is considered that the proposed approach is valid for resource-poor language pairs. In addition, the sizes of SOMs are slightly smaller in most cases than the numbers of considered seed words, with the exception of the French–Korean translations (835 seed nouns in 90% of the total words). This case does not represent the optimum performance because fine-tuning of various parameters has not been considered. The best tuning depends on many factors such as the sizes of the Gaussian filter and SOM. Thus, the tuning settings can be improved. Previously reported results (i.e., for the SOM sizes) indicate that somewhat duplicated input samples for each winner result in better performance than unique winners that correspond to each sample. Basically, the Gaussian function controls many neighbors around centroids for each iteration, and similar seed words should adjust their weights reciprocally. Thus, the area under the Gaussian function should be sufficiently wide for this interactive process. However, in these experimental conditions (i.e., 5 × 5 for various SOM sizes), somewhat small SOM sizes would yield better performance. Of course, the performance depends on the data.

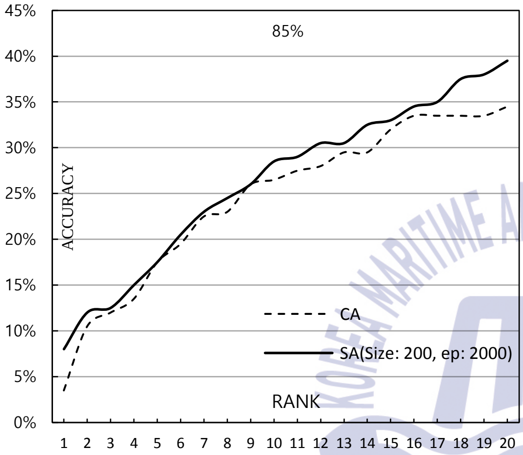


Figure 5.8: ACC for KR → FR (296 nouns)

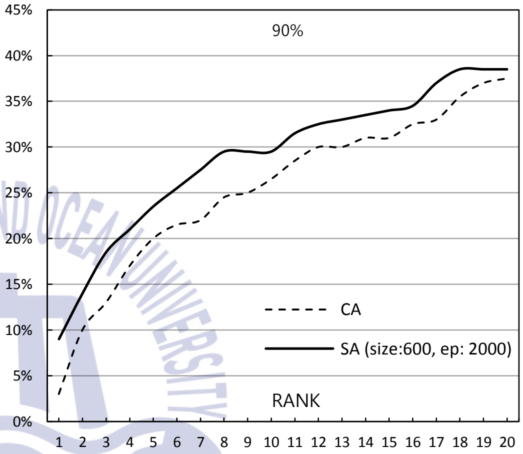


Figure 5.9: ACC for KR → FR (787 nouns)

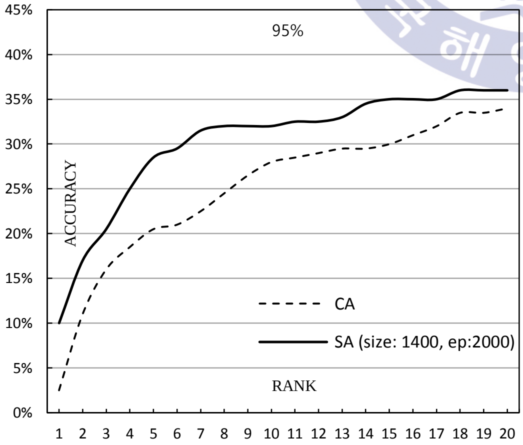


Figure 5.10: ACC for KR → FR (2399 nouns)

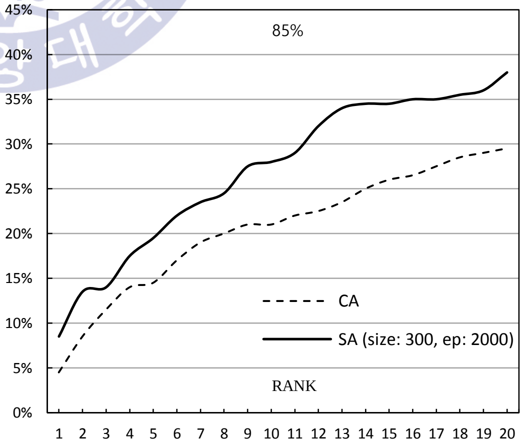


Figure 5.11: ACC for FR → KR (388 nouns)

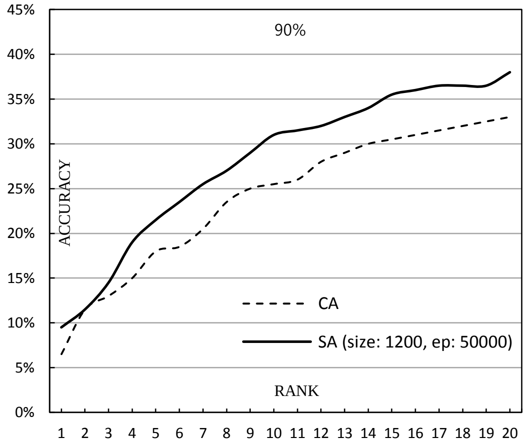


Figure 5.12: ACC for FR → KR (835 nouns)

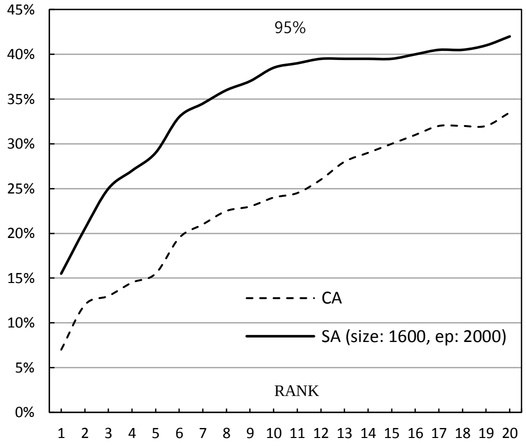


Figure 5.13: ACC for FR → KR (2138 nouns)

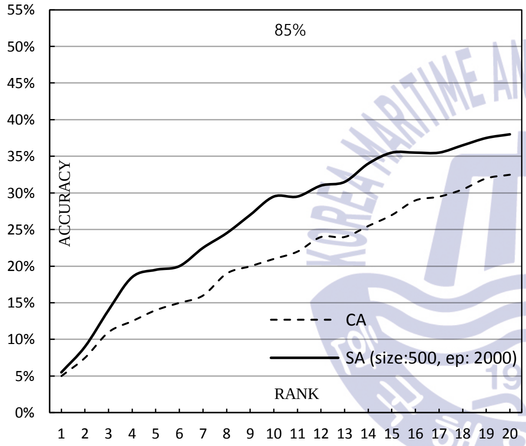


Figure 5.14: ACC for KR → ES (563 nouns)

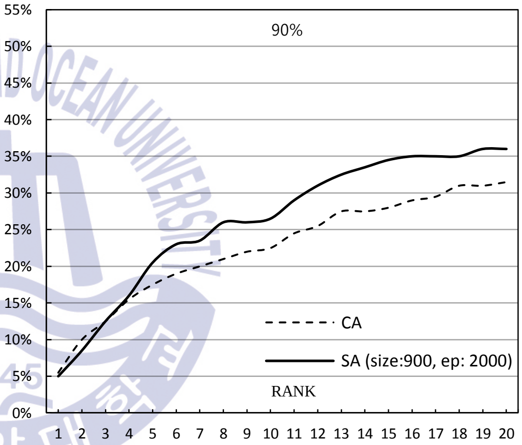


Figure 5.15: ACC for KR → ES (1388 nouns)

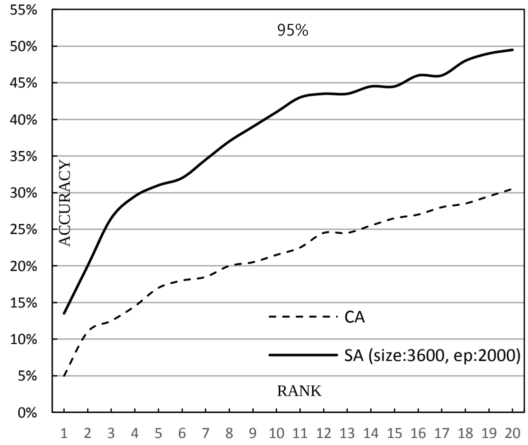


Figure 5.16: ACC for KR → ES (4387 nouns)

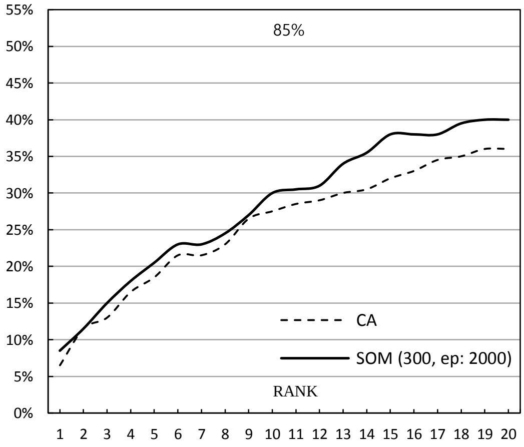


Figure 5.17: ACC for ES → KR (345 nouns)

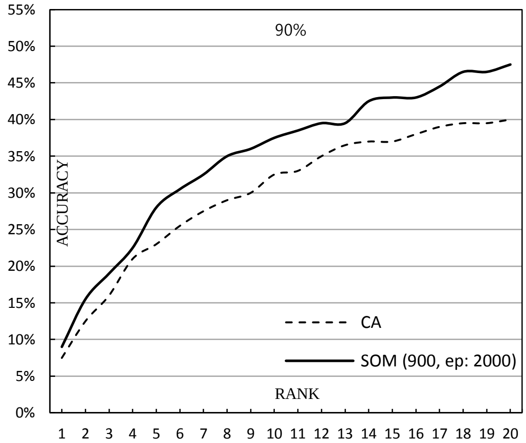


Figure 5.18: ACC for ES → KR (736 nouns)

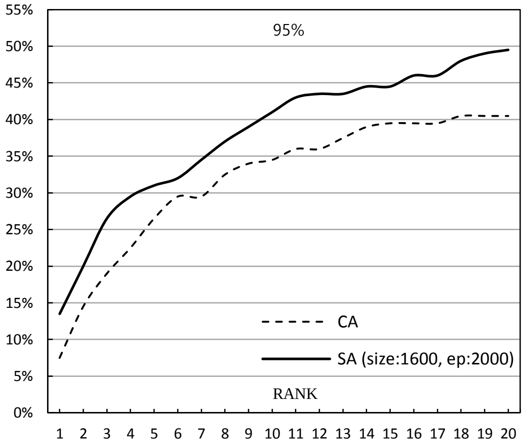


Figure 5.19: ACC for ES → KR (1813 nouns)

5.3 Summary

This chapter has presented a method of extracting multilingual lexicons from comparable corpora using a machine learning technique (i.e., the SOM) in unsupervised and semi-supervised manners. The key idea is to set two types of SOM vectors for comparison.

First, the proposed approach builds two types of synonym vectors from each monolingual corpus. These synonym vectors come from context vectors weighted by a word association measure such as the CHI value. The *k* nearest context vectors with the highest similarity scores are considered synonyms. Thus, these synonym vectors are weighted by their similarity scores. In this investigation, it was expected that semantically similar words could be collected via these synonym vectors. After the synonym vectors have been built in the proposed approach, the source SOM is trained using the source synonym vectors in an unsupervised fashion. That is, the winner (i.e., winning neurons, nodes, or units) with the minimum Euclidean distance score in each phase is selected in a natural manner. After a single winner is selected, the weights for the winner and its neighbors are updated in an online mode. After the SOM reaches convergence at a specific iteration, the weight vectors are preserved to train the target SOM with specific winners. If there is no corresponding translation for a source word in the seed dictionary, the target words are essentially trained in an unsupervised fashion. However, target words with corresponding source entries in the seed dictionary are trained in a supervised fashion. All winners whose words are included in the seed dictionary are updated in an online mode. Thus, the target SOM can be treated as a supervised model. In this sense, this method assumes that two words in different languages are translations and that each has the same winner. After the SOM vectors have been built, similarity scores between one source word and all target words are computed by cosine similarity. Finally, after all similarity scores for the source word have been calculated, the top *x* candidates are selected and added to the bilingual lexicon.

The most prominent advantage of the proposed approach is that it outperforms the CA under the same experimental settings, specifically, the same seed dictionaries and corpora. Our experimental results show that the proposed approach can extract multilingual lexicons for resource-poor language pairs. However, there is some room for improvement with various parameter factors such as the size of the SOM, learning rate, Gaussian function, and epochs.

Tuning such parameters to obtain optimal performance is planned for future work. Furthermore, in our simplified experiments, only nouns were used; thus, other parts of speech could also be considered in future. Finally, thorough error analysis is also required.

Chapter 6

Constituent-Based Approach

This chapter addresses a method that automatically extracts bilingual MWEs in resource-poor language pairs such as Korean–French/–Spanish. The PCAM is used as the baseline, and the performance of the PCAM is reinforced. The PCAM has difficulty when the MWE contexts are insufficient. To mitigate its shortcomings, a method to compute constrained similarity scores between source words and translation equivalents and between source words and constituents of the translation equivalents is presented. Based on this idea, the reinforced approach (the constituent-based context approach) significantly outperforms the baseline in terms of accuracy. This chapter also evaluates the proposed approach through several types of tests.

6.1 Concept of Constituent-Based Approach

As mentioned in Section 2.2, the earlier approach (i.e., PCAM) results in several types of errors. The CTA described in this chapter focuses on one of the error types. That is, the CTA solves the problem in which one of the constituents from a translation equivalent is extracted as the top *x* translation equivalent. For the previous example, the translation equivalent *département* (or *linguistique*) is extracted as a single result when *département de linguistique* should be extracted for the source word 언어학과 (*eoneohakgwa*, department of linguistics). This type of error can be caused when MWEs have poor contexts in common but contain constituents with much richer independent contexts. Most MWEs of this type are infrequent and originally not significantly more than their constituents in a corpus whenever the multi-word is a high-frequency word.

The contribution of the CTA is that it considers the relationships between source MWEs and translation equivalents and between the MWEs and constituents of the translation equivalents. This primarily occurs for low-frequency words because their context vectors are not sufficient to yield high similarity scores. Note that this thesis defines a source MWE as a landmark case; therefore, the constituents of source MWEs are ignored. That is, only target constituents are considered. The overall structure of the CTA is illustrated in Figure 6.1.

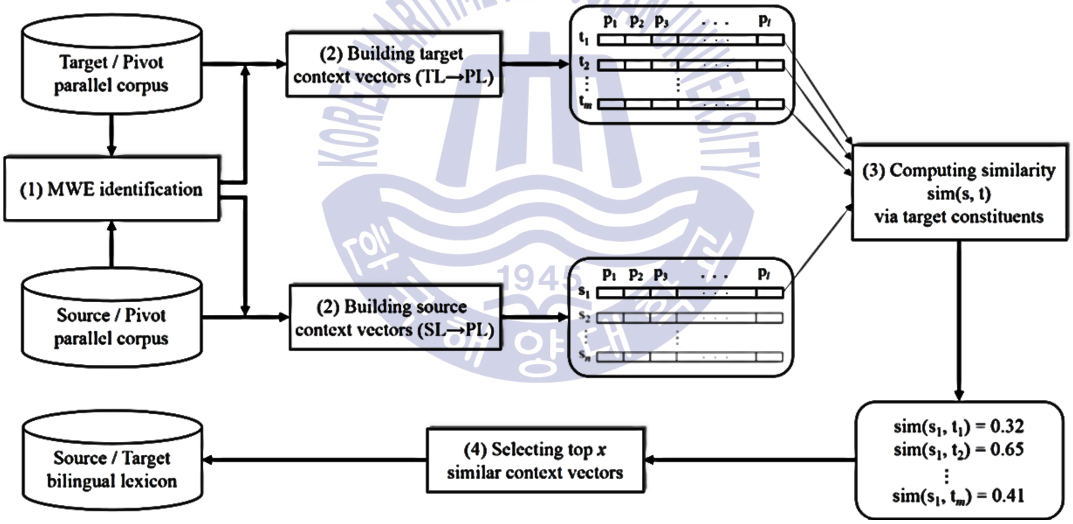

Target / Pivot parallel corpus
(1) MWE identification
Source / Pivot parallel corpus
(2) Building target context vectors (TL→PL)
(2) Building source context vectors (SL→PL)
P1 P2 P3 . . . Pl
t1
t2
tm
s1
s2
sn
(3) Computing similarity sim(s, t) via target constituents
sim(s1, t1) = 0.32
sim(s1, t2) = 0.65
sim(s1, tm) = 0.41
(4) Selecting top x similar context vectors
Source / Target bilingual lexicon

Figure 6.1: Overall structure of CTA

As can be seen in Figure 6.1, the structure of the proposed approach is very similar to that of the PCA.

(1) **MWE identification**: First, MWE candidates should be extracted via the identification method described in Section 2.2.1 (see Figure 2.4 for more detail). Then, all possible n-grams ($2 \leq n \leq 3$) can be independently extracted from each of the monolingual corpora (i.e., the source and target languages). Next, reasonable collocations are extracted from

the n-grams by an association measure (PMI was empirically determined in this work). Some collocations with scores lower than a specific threshold are eliminated. After that, several POS sequence patterns are provided to remove irrelevant MWE candidates. This identification method requires morphological analyzers and noun phrase patterns for each language. Removing irrelevant MWE candidates is relatively easy because this information is readily available for general languages. This thesis assumes that extracted MWE candidates are accepted as actual MWEs.

(2) **Building context vectors**: After the MWE candidates have been extracted, context vectors from two parallel corpora are constructed separately. This process is the same as that described in step (1) in Section 3.1. Note that added MWE candidates are also involved in building context vectors before extraction. The MWE candidates are first converted to single tokens by concatenating them with a specific symbol such as "_" Such converted MWEs are treated the same as other single words in this work. As mentioned in Section 2.2.2, MWEs in the pivot languages are unnecessary. Single pivot words are sufficient to connect the source and target languages.

(3) **Computing similarity scores**: After the context vectors have been built, similarity scores between one source word and all target words are computed. Note that this step differs from step (2) in Section 3.1, while the PCAM uses the same step. The biggest difference between the PCAM and the CTA is whether or not all constituents of translation equivalents are considered. The CTA considers all constituents when similarity scores are measured. This method is not measure-specific; therefore, any similarity measurement can be used. In this thesis, only cosine similarity is considered. The modified measurement is described below.

$$\cos\theta = \mathrm{sim}(\mathrm{s},\mathrm{t}) = \alpha\left(\frac{\vec{\mathrm{s}}\cdot\vec{\mathrm{t}}}{|\vec{\mathrm{s}}||\vec{\mathrm{t}}|}\right) + \beta\left(\frac{1}{|\mathrm{t}|}\sum_{o=1}^{|\mathrm{t}|}\frac{\vec{\mathrm{s}}\cdot\vec{t}_o}{|\vec{\mathrm{s}}||\vec{t}_o|}\right) \qquad (6.1)$$

As can be seen in Equation 6.1, this measure computes the similarity between two vectors (i.e., $\vec{\mathrm{s}}$ and $\vec{\mathrm{t}}$), where $|\mathrm{t}|$ denotes the number of translation equivalent constituents. For example, the similarity score sim (언어학과, *département de linguistique*) between the Korean word 언어학과 (*eoneohakgwa*, department of linguistic) and the French word *département de linguistique* can be scored as follows (note that two parameters are empirically determined as $\alpha = 0.6, \beta = 0.4$):

$$\mathrm{sim}\left(\text{언어학과}, département\ de\ linguistique\right) =$$

$$0.6 \times \left(\frac{\overrightarrow{\text{언어학과}} \cdot \overrightarrow{département\ de\ linguistique}}{\left|\overrightarrow{\text{언어학과}}\right| \left|\overrightarrow{département\ de\ linguistique}\right|} \right) +$$

$$0.4 \times \frac{1}{2} \left(\frac{\overrightarrow{\text{언어학과}} \cdot \overrightarrow{département}}{\left|\overrightarrow{\text{언어학과}}\right| \left|\overrightarrow{département}\right|} + \frac{\overrightarrow{\text{언어학과}} \cdot \overrightarrow{linguistique}}{\left|\overrightarrow{\text{언어학과}}\right| \left|\overrightarrow{linguistique}\right|} \right).$$

As can be seen, only content words (i.e., nouns, verbs, adjectives, or adverbs) are included in the translation equivalents. This measurement, which considers constituents to augment the MWE scores, is the key feature of the CTA.

(4) **Selecting similar context vectors**: After all similarity scores for each source word have been computed, the top *x* candidates are selected and added to the bilingual lexicon. This step is exactly the same as step (3) in Section 3.1.

6.2 Experiments

6.2.1 Resources

To implement the CTA, several linguistic resources (i.e., stemmers, lemmatizers, POS taggers, and parallel corpora for the source–pivot and pivot–target language pairs) were required. For the Korean–English pair, the KMU parallel corpus (Seo *et al*., 2006) was used. For the French–English and Spanish–English pairs, Europarl parallel corpora (Koehn, 2005) were used. The parallel corpora used in these experiments were the same as those used for the PCA (see Table 3.1 in Section 3.2.1 for more details).

As can be seen from Table 3.1, the distributions of word types and of the average number of words per sentence for the Korean–English pair differ from those of the other language pairs, as a result of the difference between their domains (i.e., news articles and European parliament proceedings). The average number of words per sentence for Korean is less than it is for any of the other languages due to a particular characteristic of Korean. On average, a Korean word usually contains one or more morphemes (2.3 morphemes per word in this experiment), depending on the domain or corpus.

Before using the corpora, the same pre-processing tasks as those conducted for the PCA were performed. POS tagging for Korean morphemes and lemmatizing for English, French, and Spanish were performed using the following tools[15]: the U-tagger was used to tokenize sentences and induce POS tags of morpheme tokens in Korean, and the TreeTagger was used to lemmatize word tokens and induce their POS tags in other languages. All word/morpheme tokens were annotated and then transformed to lowercase letters. The statistics for the pre-processed texts are listed in Table 3.2.

After the texts had been pre-processed, the MWE candidates were extracted. Note that only the MWE candidates for Korean, French, and Spanish were collected. For this task, all stop-words, numeric strings, or punctuation marks were excluded. Then, word/morpheme *n*-grams ($1 \leq n \leq 3$) that occurred three or more times in each monolingual corpus (i.e., in Korean, French, and Spanish) were extracted by applying light POS filters and computing the association scores between them. This identification method is described in step (1) in Section 6.1. As mentioned previously, only noun phrases are considered as MWEs to simplify large-scale experiments. The POS filters used to extract noun phrases are listed in Table 6.1.

Table 6.1: Noun phrase patterns for three languages

Korean	French and Spanish
N-N / N-N-N	N-N / N-N-N
V-E-N	J-N / N-J
J-E-N	J-N-J / N-N-J / N-J-J
N-G-N	N-P-N

The noun phrase patterns used in this work for French originated from the approach proposed by Bouamor *et al*. (2012), and those for Spanish/Korean were based on the French list. The French list was adapted to that of Korean in order to extract as many similar POS sequences as possible by considering Korean characteristics. As shown in Table 6.1, the POS filter for Korean contains five patterns, while that for French/Spanish contains eight patterns where N is a noun, G is a

[15] Both tools are described in Section 3.3.

genitive case marker, V is a verb, J is an adjective, E is an adnominal ending, and P is a preposition. Most French/Spanish patterns consist of a noun and an adjective. To maintain balance with the French filter, the Korean filter was designed to include as many similar POS sequences as possible. The Korean bigrams V-E, J-E, and N-G usually function as noun modifiers such as unconjugated adjectives. Thus, this thesis assumes that these Korean POS sequences can act as POS filters that extract Korean MWE candidates similar to extracted French/Spanish MWE candidates. Finally, single content word/morpheme tokens (i.e., nouns, verbs, adjectives, or adverbs) including extracted MWE candidates (i.e., POS sequences listed in Table 6.1) remained as the input text. The input text statistics are listed in Table 6.2.

Table 6.2: Input text statistics

	Korean–English		**French–English**		**Spanish–English**	
Single-words	43,550	41,626	22,364	18,299	28,722	18,126
Multi-words	3,640	-	1,606	-	1,345	-

For evaluation, dictionaries that consisted of source MWEs and target translations, which were manually constructed from the Web[16] dictionary, were necessary. Four evaluation dictionaries, specifically, Korean → French, French → Korean, Korean → Spanish, and Spanish → Korean, were constructed. The form *A* → *B* indicates that *A* is a source word and *B* is its translation(s). The case of "one source MWE: one target translation or more" was considered as the evaluation set, whereas the target translation could be a single word or an MWE. Compiling the evaluation dictionaries involved the following steps.

i. All noun words from the source monolingual corpora (resp. target monolingual corpora) were extracted.

ii. Extracted nouns were queried to the Web dictionary, and the results were collected.

iii. Pre-processing was performed with some heuristics to fit the collected results to the experimental data.

The query results had the form "one French/Spanish single-/multi-word or more: one Korean single-/multi-word or more." In addition, the results presented all entries containing the queried

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

words and consisted of noun compounds, idioms, adages, and so on. Thus, the results should be focused to extract correct pairs in an appropriate manner. After all pairs had been extracted, pre-processing (i.e., tokenizing/lemmatizing and POS tagging) was performed to extract the same POS sequences as those of the MWEs of the input texts. Finally, some of the morphologically constructed source MWEs collected from the Web dictionary were selected for evaluation. In this work, source MWEs from the evaluation dictionaries that occurred at least once in the source corpora were selected. In this case, one of the translations was required to also occur in the target corpora. However, it was not necessary for all of the translations to occur in the target corpora. The number of source MWEs used for evaluation and the average numbers of their translations are listed in Table 6.3.

Table 6.3: Source MWEs in evaluation dictionaries statistics

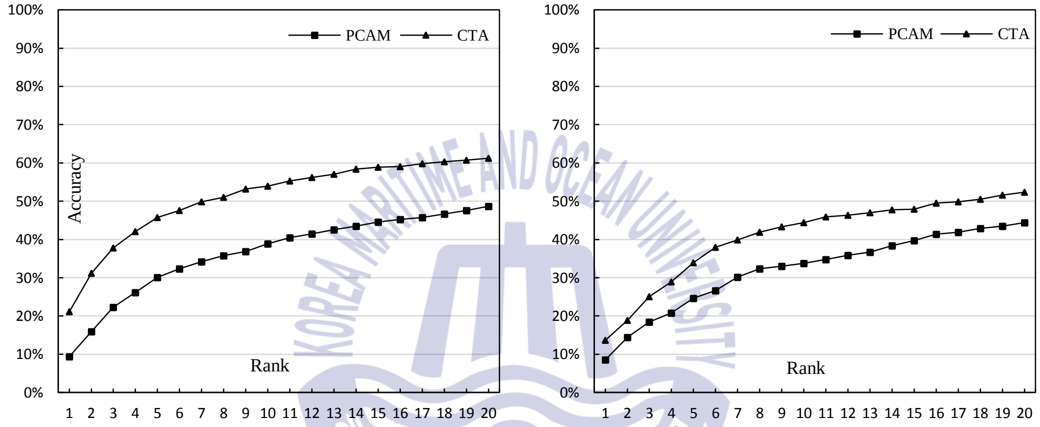
	Korean–French		**Korean–Spanish**	
Collected	15,287	28,961	8,489	15,540
Selected	754	630	426	529
Avg. translations	1.6	1.2	1.4	1.2

6.2.2 Results

In this section, the results of experiments conducted on the parallel corpora for MWE extraction are presented. The experiments were performed with the source MWEs in the evaluation dictionaries described in Table 6.3. Note that the MWEs and their translations were neither domain-specific nor over-fitted (i.e., they are considered general terms) because the source MWEs originated from Web dictionaries. Therefore, the MWEs could occur frequently or infrequently in their corpora; however, each MWE and at least one of its translations were required to occur at least once.

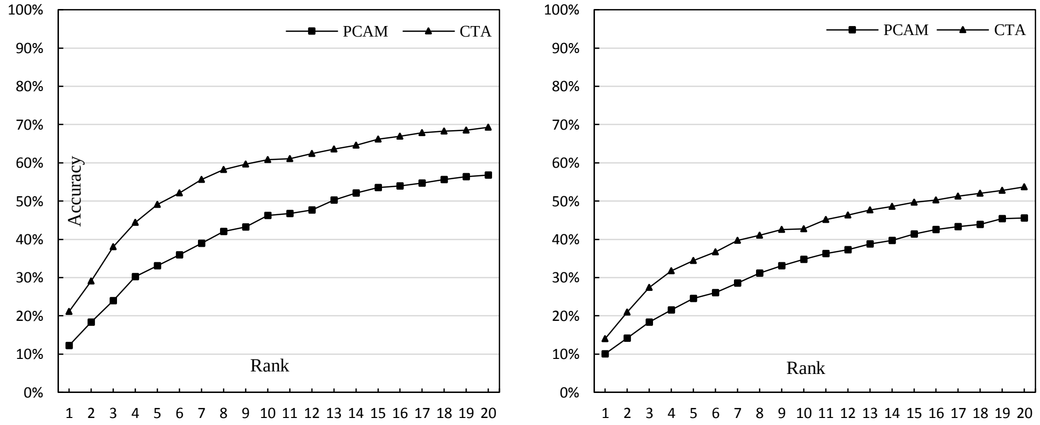
To simplify the comparison, the PCAM (Seo *et al*., 2014) is referred to as the baseline in the remainder of this section. The PCAM measures a general cosine similarity score between two context vectors (i.e., similarity scores for constituents are ignored). Alternatively, the CTA considers the relationship between one source word and the constituents of the translation equivalent. Figure 6.2 (resp. Figure 6.3) shows the accuracy from the top 1 to 20 for a Korean–French pair (resp. Korean–Spanish pair), that is, the percentage of source words that had at least one exact translation in the top *x* candidate translations.

As can be seen in Figures 6.2 and 6.3, the CTA significantly outperforms the PCAM for the Korean–French pair. With regard to Korean → French translations, the best accuracy, 61.3%, (455 out of 754 Korean source MWEs) was obtained in the top 20 by the CTA, while 48.7% (367 out of 754 Korean source MWEs) was obtained by the PCAM. For French → Korean translations, the best accuracy, 52.4%, (330 out of 630 French source MWEs) was obtained in the top 20 by the CTA, while 44.4% (280 out of 630 French source MWEs) was obtained by the PCAM. These results are very meaningful because they clearly demonstrate that considering constituents improve the PCAM performance.



(a) Korean → French (b) French → Korean

Figure 6.2: Accuracy results for Korean–French parallel corpora



(a) Korean → Spanish (b) Spanish → Korean

Figure 6.3: Accuracy results for Korean–Spanish parallel corpora

The experimental results for the Korean–Spanish pair also support this claim. The results for the Korean–Spanish pair (Figure 6.3) show better performance than those for the Korean–French pair. The best accuracy, 69.3%, (295 out of 426 Korean source MWEs) was obtained on the top 20 by the CTA, while 56.8% (242 out of 426 Korean source MWEs) was obtained by the PCAM. For Spanish → Korean translations, the best accuracy, 53.7%, (284 out of 529 Spanish source MWEs) was obtained on the top 20 by the CTA, while 45.6% (241 out of 529 Spanish source MWEs) was obtained by the PCAM. These results indicate that the CTA generally shows the best performance. Note that the evaluated words were not high-frequency words or were general terms not fitted to specific domains.

Table 6.4 shows the overall error statistics from the evaluated methods for two parallel corpora. The statistics are observed where accuracies on top 20 are considered. On average, the CTA reduced errors by 10.4%, indicating that considering constituents can improve the MWE alignment performance for resource-poor language pairs, even if the approach generates other types of errors.

Table 6.4: Error statistics for evaluated methods

	Korean–French		Korean–Spanish	
# of source MWEs	754	630	426	529
# of source MWEs with no translation				
from the PCAM	387 (51.3%)	354 (56.2%)	184 (43.2%)	288 (54.4%)
from the CTA	292 (38.7%)	300 (47.6%)	131 (30.8%)	245 (46.3%)

Even though the CTA outperformed the PCAM in the experiments, the performance of proposed approach still requires improvement. In particular, generating Korean translations, specifically, French/Spanish → Korean translations, is comparatively more difficult (47.6% error rate for French → Korean, 46.3% for Spanish → Korean translations; Table 6.4) than generating Korean → French translations (38.7% error rate) and Korean → Spanish translations (30.8% error rate).

There are several reasons for this problem. First, Korean translations in the evaluation dictionaries are insufficient compared to French/Spanish translations. In the dictionaries, each French (resp. Spanish) source MWE has on average 1.17 (resp. 1.18) Korean translations, while each Korean source MWE has on average 1.59 French translations (resp. 1.36 Spanish translations). These numbers are only slightly different; nevertheless, French and Spanish source MWEs have less opportunity to be identified as correct. Second, the number of types of Korean MWE candidates (i.e., 3,640; Table 6.2) is relatively higher than those of the other languages: 1,606 for French and 1,345 for Spanish. Therefore, in French and Spanish, there are more source MWEs that can be aligned with target candidate translations than vice versa. Of course, if robust contexts for MWE candidates are supported, having various types of candidates is not particularly significant. To investigate how robust the contexts of MWE candidates are in the corpora, the error frequencies for MWE candidates were analyzed. Here, frequency is regarded as the number of sentences containing a specific MWE. The erroneous MWE candidate statistics are listed in Tables 6.5 and 6.6. Note that the frequency $\boldsymbol{f}$ in Tables 6.5 and 6.6 indicates the number of sentences containing a specific MWE.

For the PCAM, as shown in Tables 6.5 and 6.6, over 94% of erroneous source MWEs for Korean–French pairs (resp. over 98% for Korean–Spanish pairs) occur at most 100 times in their

Table 6.5: MWE error statistics for Korean–French translations

	Configuration	$\boldsymbol{f \leq 10}$	$\boldsymbol{f \leq 100}$	$\boldsymbol{100 < f}$	**Max $\boldsymbol{f}$**	**Avg. $\boldsymbol{f}$**
Korean to French	**PCAM** (387 errors): frequency (%)					
	KR source MWEs	233 (60.2)	**366 (94.6)**	21 (5.4)	1067	33.1
	Top 1 FR equivalent	196 (50.7)	277 (71.6)	110 (28.3)	8195	346.5
	FR translations	107 (27.7)	223 (57.6)	164 (42.3)	24064	1019.4
	CTA (292 errors): frequency (%)					
	KR source MWEs	159 (54.5)	**271 (92.8)**	21 (7.2)	1067	39.6
	Top 1 FR equivalent	87 (29.8)	137 (46.9)	155 (53.1)	34317	1509.4
	FR translations	74 (25.3)	158 (54.1)	134 (45.9)	11268	827.1
French to Korean	**PCAM** (354 errors): frequency (%)					
	FR source MWEs	185 (52.3)	**333 (94.1)**	21 (5.9)	3587	45.5
	Top 1 KR equivalent	223 (63.0)	296 (83.6)	58 (16.4)	13549	207.7
	KR translations	160 (45.2)	276 (78.0)	78 (22.0)	14209	276.7
	CTA (300 errors): frequency (%)					
	FR source MWEs	162 (54.0)	**281 (93.7)**	19 (6.3)	3587	42.0
	Top1 KR equivalent	110 (36.7)	166 (55.3)	134 (44.7)	20519	1564.9
	KR translations	106 (35.3)	222 (74)	78 (26.0)	14209	315.5

Table 6.6: MWE error statistics for Korean–Spanish translations

	Configuration	$f \leq 10$	$f \leq 100$	$100 < f$	Max f	Avg. f
Korean to Spanish	**PCAM** (184 errors): frequency (%)					
	KR source MWEs	100 (54.3)	**182 (98.9)**	2 (1.1)	795	36.2
	Top 1 ES equivalent	95 (51.6)	149 (81.0)	35 (19.0)	27903	563.5
	ES translations	73 (39.7)	162 (88.0)	22 (12.0)	26828	428.0
	CTA (131 errors): frequency (%)					
	KR source MWEs	70 (53.4)	**120 (91.6)**	11 (8.4)	596	32.0
	Top 1 ES equivalent	35 (26.7)	62 (47.3)	69 (52.7)	37580	2949.9
	ES translations	51 (38.9)	104 (79.4)	27 (20.6)	26828	582.1
Spanish to Korean	**PCAM** (288 errors): frequency (%)					
	ES source MWEs	126 (43.8)	**283 (98.3)**	5 (1.7)	1188	44.8
	Top 1 KR equivalent	194 (67.4)	255 (88.5)	33 (11.5)	9958	302.8
	KR translations	110 (38.2)	264 (91.7)	24 (8.3)	10908	216.4
	PCA (245 errors): frequency (%)					
	ES source MWEs	111 (45.3)	**227 (92.7)**	18 (7.3)	880	35.4
	Top1 KR equivalent	95 (38.8)	152 (62.0)	93 (38.0)	39357	1050.6
	KR translations	88 (35.9)	187 (76.3)	58 (23.7)	4648	768.3

corpora. It is difficult to define what constitutes a small number; however, this thesis assumes that only 100 out of nearly 0.4 million Korean sentences (resp. 0.5 million for French and Spanish sentences) is a very low number. In addition, over 50% of erroneous source MWEs (except for those occurring in Spanish → Korean translations) occur only 10 times in their corpora, and this number is extremely small. More importantly, nearly 95% of erroneous source MWEs might have very narrow contexts. In contrast, approximately 5% of errors might have relatively sufficient contexts. This phenomenon is also evident in the CTA configuration.

For the CTA, similar distributions, as well as a clear difference from the former method (i.e., the PCAM), are evident. Almost 93% of erroneous source MWEs occur at most 100 times for Korean–French pairs (resp. almost 92% for Korean–Spanish pairs). These distributions are similar to those of the former method. On the other hand, although MWEs that occur at most 100 times are considered low-frequency MWEs in the rest of this section, various types of low-frequency MWEs from the CTA decrease no matter what the types of MWEs are. Moreover, the number of MWEs that occur more often than low-frequency MWEs increases.

For this frequency alone, it not possible to predict how many content words composing context vectors are around the MWE. In addition, whether decreasing low-frequency erroneous MWEs is worthwhile has not been determined. However, it is assumed that most errors result from lack of context. Here, lack of context is defined as a context that does not include sufficient common features of the source and target words to allow them to be aligned correctly.

If the relationship between a frequency and a context size is observed, their effect on errors can be estimated. Figure 6.4 presents the MWE frequencies (dotted line) and context sizes (solid line) for French → Korean translation. As can be seen, the context size is not always directly proportional to the MWE frequency. A very small number of MWEs have context sizes smaller than their frequencies. Nevertheless, in general, as MWE frequency increases, context size also increases and does so at a faster rate. In addition, when MWE frequency is low, the gap between the frequency and context size is small. This phenomenon is very natural because context originates from sentences whose lengths are fixed or limited. Of course, a high-frequency MWE has abundant opportunities to obtain essential contexts. However, it is not possible to confirm that a low-frequency MWE has no alternative but to have a poor context. If a low-frequency MWE has certain crucial or essential contexts and shares them with its translation, alignment of the MWE can be performed successfully.

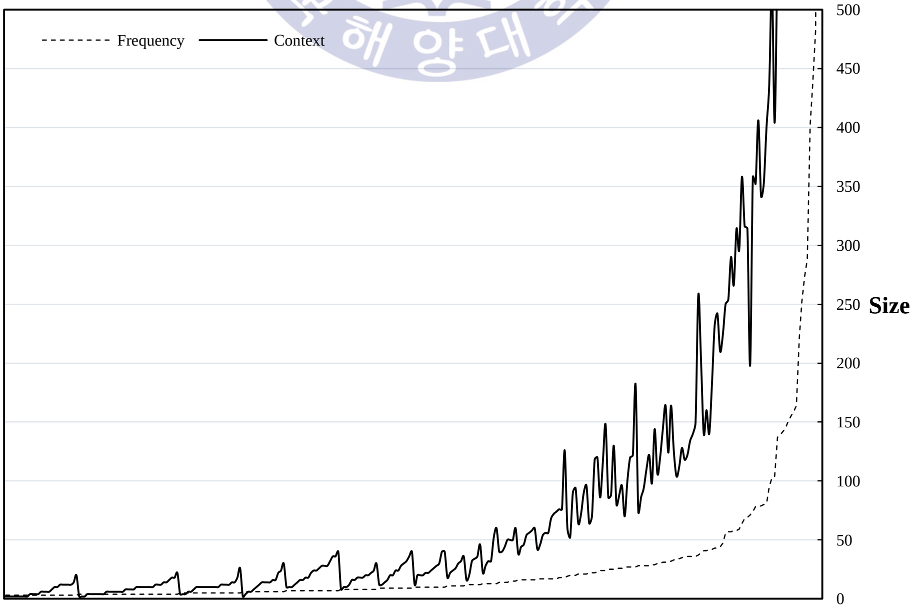


Figure 6.4: Comparison of frequency and context size for French → Korean translations

This investigation estimated the numbers of common contexts to reveal the relationship between common contexts and errors. Common context means that a common pivot (e.g., English) word exists in two monolingual corpora, for example, from the English corpus in the Korean–English parallel corpus and another English corpus in the French–English parallel corpus. It is assumed that, as the number of common contexts increases, MWE alignment errors become less frequent. The statistics for such common contexts are listed in Table 6.7. The table contains the numbers of common contexts between source MWEs and their target translations in the evaluation dictionary, denoted |source MWE ∩ translation| and (a), (b), or (c) according to the frequency, as well as between the translations and the top 1 equivalents, denoted as |Top 1 equivalent ∩ translation| and (d), (e), or (f) according to the frequency. Note that, if the number of common contexts between a source MWE and its translations is 0 such as in case (a), correct alignment can never be achieved using the context-based method. For example, 105 source MWEs share no common contexts with their translations when the PCAM for Korean → French translations is considered. With the PCAM, on average, 27.8% of errors for Korean–French pairs (resp. 21.5% for Korean–Spanish pairs) are included in this case. When the CTA is considered in case (a), the percentages (with regard to no common context) are reduced (27.1% to 22.6% for Korean → French translations, 28.5% to 19.0% for French → Korean translations, 22.8% to 16.8% for Korean → Spanish translations, and 20.1% to 19.6% for Spanish → Korean translations).

Table 6.7: Statistics of intersections between contexts in errors

	| src MWE ∩ translation |			**| Top 1 equivalent ∩ translation |**		
Freq. (case)	0 (a)	≤ 100 (b)	> 100 (c)	0 (d)	≤ 100 (e)	> 100 (f)
Korean to French						
PCAM (387)	105 (27.1%)	374 (96.6%)	13 (3.4%)	106 (27.4%)	339 (87.6%)	48 (12.4%)
CTA (292)	66 (22.6%)	278 (95.2%)	14 (4.8%)	49 (16.8%)	208 (71.2%)	84 (28.8%)
French to Korean						
PCAM (354)	101 (28.5%)	347 (98.0%)	7 (2.0%)	68 (19.2%)	343 (96.9%)	11 (3.1%)
CTA (300)	57 (19.0%)	288 (96.0%)	12 (4.0%)	69 (23.0%)	234 (78.0%)	66 (22.0%)
Korean to Spanish						
PCAM (184)	42 (22.8%)	180 (97.8%)	4 (2.2%)	57 (31.0%)	171 (92.9%)	13 (7.1%)
CTA (131)	22 (16.8%)	128 (97.7%)	3 (2.3%)	24 (18.3%)	105 (80.2%)	26 (19.8%)
Spanish to Korean						
PCAM (288)	58 (20.1%)	278 (96.5%)	10 (3.5%)	80 (27.8%)	277 (96.2%)	11 (3.8%)
CTA (245)	48 (19.6%)	237 (96.7%)	8 (3.3%)	59 (24.1%)	213 (86.9%)	32 (13.0%)

Thus, the CTA is considered an effective way to reduce such errors. It can also improve the alignment of low-frequency MWEs and consequently improve accuracy.

The top 1 equivalent listed in Table 6.7 is most likely considered a translation of a source MWE because it has the highest similarity score of the top *x* equivalents. It comes from system results; therefore, it can be considered an evaluation criterion to determine if the method improves performance. If the number of the cases of (d), which have no common context between a top 1 equivalent and a target translation in an evaluation dictionary, is reduced by the CTA compared to when the PCAM is used, then the CTA can help reduce errors. Except for French → Korean translations, the CTA generally reduces the number of errors in (d). Therefore, at minimum, source MWEs are more likely to be correctly aligned to translation equivalents by using the CTA. On the other hand, the cases of (c), in which there are more than 100 common contexts between source MWEs and target translations, or (f) for the top 1 equivalents and target translations, show that the CTA clearly extracts equivalents with rich contexts. Moreover, their parallel source MWEs also have rich contexts. Thus, the CTA volumizes the contexts of the MWE candidates, which improves the performance of the approach. However, there is also an adverse effect.

To analyze the overall impact of the CTA, all errors obtained by the two methods are addressed with reference to three cases: (I) the translation equivalent is a reference translation that is not included in an evaluation dictionary; (II) there is no correct translation; however, a translation equivalent and a translation in an evaluation dictionary originate from a same domain; and (III) the translation equivalent is one constituent of a correct MWE translation. These error types will be discussed and several examples will be provided. The statistics of these error types are presented in Table 6.8.

Table 6.8: Error-type statistics

Language Pair	Method	Type I	Type II	Type III
Korean → French	PCAM (387)	9 (2.3%)	80 (20.7%)	144 (37.2%)
	CTA (292)	13 (4.5%)	110 (37.7%)	187 (64.0%)
French → Korean	PCAM (354)	54 (15.3%)	156 (44.1%)	96 (27.1%)
	CTA (300)	35 (11.7%)	150 (50.0%)	69 (23.0%)
Korean → Spanish	PCAM (184)	11 (6.0%)	59 (32.1%)	64 (34.8%)
	CTA (131)	13 (9.9%)	58 (44.3%)	89 (67.9%)
Spanish → Korean	PCAM (288)	19 (6.6%)	89 (30.9%)	48 (16.7%)
	CTA (245)	37 (15.1%)	149 (60.8%)	47 (19.2%)

The results (except for French → Korean translations) corresponding to Type I errors obtained by the CTA show higher percentages, specifically, 2.3% to 4.5% for Korean → French translations, 6.0% to 9.9% for Korean → Spanish translations, and 6.6% to 15.1% for Spanish → Korean translations. Two examples are presented to illustrate these results. First, the Korean → French translation pair 비상 사태 (*bisang satae*, state of emergency) → *état d'urgence* already exists in the evaluation dictionary. The French word *état d'urgence* has related meanings for synonyms (or reference translations of 비상 사태) such as *situation d'urgence* (emergency situation), *situation critique* (plight), and *situation de danger* (dangerous situation). However, the acceptable translation equivalents *situation d'urgence*, *situation critique*, and *situation de danger* would be marked as incorrect because the evaluation dictionary contains neither synonyms of translations nor reference translations of source MWEs. If the evaluation dictionaries are extended either manually or automatically, the performance of the approach could be improved significantly. Alternatively, *col blanc* (white collar) has Korean translations for the literal meaning of 하얀색 깃 (*hayansaek git*) as well as for the idiomatic meaning of 사무 직원 (*samu jikwon*, clerical worker, office worker). As mentioned previously, such idiomatic expressions are ignored in this work; therefore, the latter example cannot be resolved with the CTA.

Type II errors indicate that extracted translation equivalents are incorrect; however, they achieve correct translation of a same topic. For example, when the Korean → Spanish evaluation dictionary includes the pair 민간 항공 (*mingan hanggong*, civil aviation) → *aviación civil*, Spanish translation equivalents such as avión (plane), *aeronave* (aircraft), and *línea internacional* (international line), are extracted as the top *x* equivalents. All of these equivalents, the target translation *aviación civil*, and the source MWE 민간 항공 are related to the same topic, “a flight.” In other words, these words share common context words in the pivot language, English such as *flight*, *airplane*, *international*, *domestic*, and so forth. However, the exact target translation *aviación civil* does not exist in the target monolingual corpus or has very poor context even though it exists. This shortcoming could be due to misalignment of parallel sentences or mismatching of domains between the source and target corpora. As mentioned in Section 6.2.1, each parallel corpus shares the same domain; however, the source/target monolingual (i.e., Korean/French or /Spanish, and vice versa) corpora do not. Fortunately, the CTA extracts words sharing common topics much more frequently than does the baseline. With regard to the number, the listed numbers obtained by the CTA decrease slightly (e.g., 156 to 150 and 59 to 58) or increase significantly (e.g., 80 to 110 and 89 to 149). However, the percentages indicate that the claim is true. Considering these results, it is evident that the CTA gathers increasingly more equivalents that share contexts that are as similar as possible.

Type III errors are also referenced in the second example in Section 6.1. This work attempts to improve this type of error. As mentioned previously, Type III indicates that an equivalent is not extracted as a whole; however, its constituent parts are extracted as the top *x* equivalents. Note that this phenomenon occurred the most frequently when the contexts of a translation equivalent as a whole were very poor and the contexts of the constituent parts were rich. Usually, low-frequency words exhibit this type of error. As can be seen from Type III in Table 6.5, the results (except for French → Korean translations) obtained by the CTA show higher percentages than do the results obtained from the baseline, specifically, 37.2% to 64.0% for Korean → French translations, 34.8% to 67.9% for Korean → Spanish translations, and 16.7% to 19.2% for Spanish → Korean translations. Taken as part of the error types, these results seem rather poor. However, taken as a whole, the error rates are decreased by the CTA.

6.3 Summary

This chapter has presented an efficient method for bilingual MWE alignment in resource-poor language pairs such as Korean–French/–Spanish. In general, bilingual corpora are essential to perform bilingual lexicon extraction; however, parallel corpora are unavailable for many domains. To address this issue, the PCAM, which uses two parallel corpora sharing one pivot language (e.g., English), was proposed by Seo *et al*. (2014). This idea is reasonable because corpora for certain language pairs such as English–* are available online. With two parallel corpora, the approach identifies MWE candidates in each monolingual corpus by a general collocation measure such as PMI and then aligns these candidates on the basis of the similarity scores of their context vectors. The PCAM performs well even without an external language resource such as a seed dictionary to translate from one language to another. However, alignment of bilingual MWEs in parallel corpora using the PCAM is difficult when the context words of either the source or target MWEs as a whole are insufficient. In this case, the context indicates that a single pivot word co-occurs with source or target MWEs in the aligned pair of sentences of both languages. This thesis assumes that single words in a pivot language are sufficient to act as bridges that connect the source and target languages. Therefore, extracting pivot MWEs is not required. The translation equivalent accuracy may decrease because of lack of context, and this shortcoming should be addressed.

In a situation like this, the CTA adequately addresses the poor-context problem. More specifically, it calculates vector similarity scores between source MWEs and complete translation equivalents and between source MWEs and the constituents of translation equivalents. The similarity scores are not considered independently; instead, they are summed into a single score that indicates which of translation equivalent as a whole. Again, with this approach, complete translation equivalent, rather than parts of equivalents, generally obtain higher similarity scores because this thesis assumes that, if a source word is a multi-word, its translation equivalent is also more likely to be a multi-word.

In the experiments, the CTA significantly outperforms the PCAM (baseline) in accuracy. For the CTA, the highest accuracy, 61.3%, was obtained using the top 20 for Korean → French translations and 52.4% for French → Korean translations. In addition, the CTA obtained the highest accuracy, 69.3%, using the top 20 for Korean → Spanish translations and 53.7% for Spanish → Korean translations. The proposed approach was evaluated with reference to three types of errors. Type I error occurs when an extracted equivalent is one of the reference translations excluded from the evaluation dictionary; thus, correctness could not be estimated. Type II error occurs when an extracted equivalent hints a same domain with a correct translation. Type III error occurs when only one constituent of a target translation is extracted rather than a complete translation. Note that all of the error types result from lack of contexts. Type I can be easily improved if evaluation dictionaries are extended. The other types of errors could be solved if the domains of two parallel corpora were the same or if the size of the corpora were increased. However, the experimental results indicate that the CTA performs very well for MWE alignment for resource-poor language pairs.

Of course, there are many opportunities for improvement. An important area for further research is fine-tuning of the parameters α and β (Section 6.2.2) to maximize accuracy. In addition, the evaluation dictionaries could be extended by extracting synonyms of target translations or by collecting reference translations of source MWEs.

Chapter 7

Conclusions and Future Work

This final chapter concludes the thesis by summarizing the approaches that were investigated. Moreover, future research related to each approach is suggested.

7.1 Conclusions

In this section, all of the approaches proposed in this thesis are summarized and compared. The main issue addressed in this thesis was the lack of availability of direct linguistic resources such as bilingual corpora or bilingual seed dictionary. For example, Korean, French, and Spanish are resource-rich languages. However, when paired, they are resource-poor. This thesis has proposed several approaches to improve upon earlier methods when applied to resource-poor language pairs under various conditions.

Table 7.1: Comparison of characteristics for all approaches

	Corpora[*]	Seed Dict	Pivot	Word Type	Vector Type	Key Features
Earlier approaches						
CA	CC	o	x	single	context	Baseline
EA	CC	o	x	single	context	CA + Nearest vectors
PA	x	o	o	single	x	Word-alignment models
PCAM	PC	x	o	multi	context	MWE identification + PCA
Proposed approaches						
PCA	PC	x	o	single	context	CA + Pivot language
EPCA	PC	x	o	single	context	PCA + Nearest vectors
SA	CC	o	x	single	synonym	CA + SOM algorithm
CTA	PC	x	o	multi	context	PCAM with constrained similarity measurement

[*] CC: comparable corpora, PC: parallel corpora

Table 7.1 presents the detailed characteristics of the proposed approaches. As can be seen, four different approaches have been proposed, and four earlier related approaches have been reviewed. The CA was identified as the baseline, and it directly influenced many of the approaches described here. The CA is highly dependent on the coverage of a seed dictionary. To address this limitation, many revised approaches have been proposed. The extended approach (EA) collects the *k* nearest words to augment context vectors and reduce the dependence on the seed dictionary. The proposed pivot-based approach (PA) collects bilingual lexicons when most language pairs are unavailable. This approach combines existing lexicons that share one pivot language (SL–PL and PL–TL). The PA uses some word-alignment models such as exact merging to generate a bilingual lexicon for SL–TL. However, it is not an effective solution because it starts with resource-poor languages. Furthermore, building or collecting such lexicons would be a huge burden.

The PCA, which extracts bilingual lexicons for resource-poor language pairs, has also been proposed. It gathers contextually relevant words from parallel corpora to compare two different types of context vectors (i.e., SL–PL and PL–TL). Such vectors are comparable because they are built from two parallel corpora sharing one common pivot language. In addition, external linguistic resources such as a seed dictionary are unnecessary. Based on the experimental results, the PCA performs well for resource-poor language pairs, particularly for Korean–French/–Spanish translations.

As can be seen in Table 7.1, the EPCA collects the *k* nearest context words for each source word to improve the PCA. All of the collected nearest vectors satisfying a specific threshold are added to the similarity computation task to enrich the contexts of words to be compared. The EPCA uses parallel corpora rather than comparable corpora; therefore, translating source context vectors into a target language is unnecessary. However, the EPCA did not perform as expected due to the lack of contexts for the experimental data sets, particularly for different domains of two parallel corpora. The overall accuracy for Korean–French was reasonable; however, the accuracy for Korean–Spanish was somewhat low. Nonetheless, in terms of obtaining translation equivalents to higher ranks, the performance of the proposed approach was meaningful.

The SOM-based approach (SA) extracts bilingual lexicons from comparable corpora using SOMs in an unsupervised or a semi-supervised fashion. The most prominent advantage of this approach is that it outperforms the CA under the same experimental conditions, that is, when the same seed dictionaries and corpora are used. The experimental results show that the proposed approach is sufficient for extracting multilingual lexicons from resource-poor language pairs. However, it should be noted that there is room for improvement in the parameter tuning.

The proposed constituent-based approach (CTA) handles MWEs for resource-poor language pairs. The CTA was compared to the PCAM, which was modified to consider MWEs. First, the PCAM identifies MWE candidates according to their PMI and then adds them into the input data to build context vectors as single units. Then, similar to the PCA, it computes similarity scores. It performs well even without an external linguistic resource such as a seed dictionary. However, PCAM has difficulties when the pivot context words of either the source or target MWEs as a whole are insufficient. The most important issue is that the translation equivalent accuracy is reduced due to the lack of contexts. The CTA can address this issue even though there are insufficient contexts. The experimental results show that the CTA significantly outperforms the PCAM. The best accuracy (69.3%) was obtained using the top 20 for Korean → Spanish translations.

7.2 Future Work

The PCA performs well for resource-poor language pairs; however, it has several limitations. First, homonyms (particularly homographs) such as *lead* or *tear* are troublesome. If all context types are strong and the domains of two parallel corpora are the same, this problem can be solved. Second, there are neither rich gold standards to cover most synonyms nor specific evaluation measures to consider false positive translation equivalents. To address this issue, manually created evaluation dictionaries or external linguistic resources such as a well-made thesaurus are required. However, this is a case of the tail wagging the dog. Thus, other alternatives are required. Third, there are several transliterated words in the corpora. This type of error is only a small minority; thus, attempting to overcome it either automatically or manually seems difficult. Fourth, many compound nouns are missed by the word segmentation task. To address these problems, other evaluation metrics, extra segmentations, or MWE identification should be considered in future work.

The most significant problem with the EPCA is that the collected nearest context vectors cannot augment the centroid vector. In other words, the *k* nearest context vectors should be used to obtain the similarity score between the centroid, that is, a source word, and a target equivalent. However, the set of the *k* nearest context vectors cannot reinforce the analysis to find appropriate translation equivalents due to the lack of context vectors. To use the proposed approach, a sufficient number of context words or synonyms should be prepared or supported, and, at least, their domains should be unified, unless many synonyms are extended in the corpora.

For the SA, diverse/incoherent parameters for acceptable performance could be problematic. In future work, the parameters, such SOM size, learning rate, Gaussian function, and epoch, should be optimized. In addition, different parts of speech such as verbs, adjectives, and adverbs should be considered, because only nouns were considered in this work. Moreover, additional experiments for MWEs could be valuable. Most importantly, more thorough error analysis should be conducted.

Most errors in the CTA result from the lack of contexts. Performance should be extended at least by extending the evaluation dictionaries. In addition, the domains of two parallel corpora should be the same, and their sizes should be increased. Furthermore, fine-tuning of Equation 6.1 should be performed for each variation.

Reference

(Baldwin *et al*., 2003) Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the Association for Computational Linguistic (ACL'03) Workshop on Multiword Expressions: Analysis, acquisition and treatment*, volume 18, pages 89-96.

(Baldwin & Tanaka, 2004) Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *Proceedings of the Association for Computational Linguistic (ACL'04) Workshop on Multiword Expressions: Integrating Processing*, pages 24-31.

(Bannard, 2007) Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1-8.

(Baobao *et al*., 2002) Chang Baobao, Pernilla Danielsson, and Wolfgang Teubert. 2002. Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing*, volume 18, pages 1-5.

(Bouamor *et al*., 2012) Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbeaum. 2012. Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective, In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)*, pages 95-108.

(Bond *et al*., 2001) Francis Bond, Ruhaida B. Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable japanese-malay dictionary. In *Proceedings of the 8th Conference on Machine Translation Summit (Mt Summit VIII)*, pages 53-58.

(Bouma, 2010) Gerlof Bouma. 2010. Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010 Conference*, short papers, pages 109-114.

(Bowker & Pearson, 2002) Lynne Bowker and Jennifer Pearson. 2002. Working with specialized language: A practical guide to using corpora. *Routledge*, London & New York.

(Brown *et al*., 1990) Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical approach to machine translation. *Computational Linguistics*, volume 16, issue 2, pages 79-85.

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

(Chen, 1993) Stanley Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics (ACL'93)*, pages 9-16.

(Chiao & Zweigenbaum, 2002) Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208-1212.

(Chu *et al*., 2014) Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge, In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING'14)*, pages 296-309.

(Church & Hanks, 1990) Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, volume 16, issue 1, pages 22-29.

(Dagan & Church, 1994) Ido Dagan and Keneth W. Church. 1994. Termight: Identifying and translation technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 34-40.

(Dagan *et al*., 1993) Ido Dagan, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1-8.

(Daille *et al*., 1994) Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics (COLING'94)*, volume 1, pages 515-521.

(Daille & Morin, 2005) Béatrice Daille and Emmanuel Morin. 2005. French-English terminology extraction from comparable corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'05)*, pages 707-718.

(Déjean & Gaussier, 2002) Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l’extraction de lexiques bilingues à partir de corpus comparables. In W. Teubert and R. Krishnamurthy (Ed.), 2007. *Corpus Linguistics: Critical Concepts in Linguistics*, Routledge, England (in French).

(Déjean *et al*., 2002) Hervé Déjean, Éric Gaussier, and Fatia 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)*, volume 1, pages 1-7.

(Doucet & Ahonen-Myka, 2004) Antoine Doucet and Helena Ahonen-Myka. 2004. Non-contiguous word sequences for information retrieval. In *Proceedings of the Association for Computational Linguistic (ACL’04) Workshop on Multiword Expressions: Integrating Processing*, pages 88-95.

(Dunning, 1993) Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, volume 19, issue 1, pages 61-74.

(Fano, 1961) Robert M. Fano. 1961. Transmission of information: A statistical theory of communications. *MIT Press*, Cambridge, MA, USA.

(Fazly and Stevenson, 2006) Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL’06)*, pages 337-344.

(Fung & Church, 1994) Pascale Fung and Kenneth W. Church. 1994. Kvec: A new approach for aligning parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING’94)*, pages 1096-1102.

(Fung, 1998) Pascale Fung. 1998. A Statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA’98)*, pages 1-17.

(Gale & Church, 1991) William A. Gale and Kenneth W. Church. 1991. Identifying word correspondences in parallel text. In *Proceedings of the 4th Darpa Workshop on Speech and Natural Language (HLT’91)*, pages 152-157.

(Ghorpade *et al*., 2010) Santaji Ghorpade, Jayshree Ghorpade, and Shamla Mantri. 2010. Pattern Recognition Using Neural Networks. International Journal of *Computer Science & Information*

Technology (IJCSIT'10), volume 2, issue 6, pages 92-98.

(Goh *et al*., 2005) Chooi-ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2005. Building a Japanese-Chinese dictionary using kanji/hanzi conversion. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'05),* Pages 670-681.

(Green *et al*., 2011) Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 725-735.

(Grefenstette, 1994a) Gregory Grefenstette. 1994a. Corpus-derived first, second and third-order word affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (Euralex'94)*, pages 279-290.

(Grefenstette, 1994b) Gregory Grefenstette. 1994b. Explorations in automatic thesaurus discovery. *Kluwer Academic Publisher*, Boston, MA, USA.

(Haghighi *et al*., 2008) Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for computational Linguistics (ACL'08)*, pp. 771-779.

(Hazem & Morin, 2012) Amir Hazem and Emmanuel 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, pages 83-96.

(Heo, 2010) Cheol Heo. 2010. Examination how many using compound of Chinese character words and investigate the frequency of use by using analysis of modern Korean words 1, 2. Journal of *Society for Korean Classical Chinese Education*, volume 34, pages 221-244.

(Ismail & Manandhar, 2010) Azniah Ismail and Suresh Manandhar. 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.

(Júnior *et al*., 2013) Everton Luiz de Almeida Gago Júnior, Gean Davis Breda, Eduardo Zanoni Marques, and Leonardo de Souza Mendes. 2013. Knowledge discovery: Data mining by self-organizing maps. *Web Information Systems and Technologies*, Lecture Notes in Business

Information Processing, volume 140, pages 185-200.

(Katz & Giesbrecht, 2006) Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Association for Computational Linguistic (ACL'06) Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12-19.

(Kay & Röscheisen, 1993) Martin Kay and Martin Röscheisen. 1993. Text-Translation alignment. *Computational Linguistics - Special issue on using large corpora: I*, volume 19, issue 1, pages 121-142.

(Klami & Lagus, 2006) Mikaela Klami and Krista Lagus. 2006. Unsupervised word categorization using self-organizing maps and automatically extracted morphs. *Lecture Notes in Computer Science*, volume 4224, pages 912-919.

(Koehn & Knight, 2002) Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the Association for Computational Linguistic (ACL'02) Workshop on Unsupervised Lexical Acquisition*, pages 9-16.

(Koehn *et al*., 2003) Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*, volume 1, pages 127-133.

(Koehn, 2005) Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceeding of the 10th Conference on Machine Translation Summit (MT Summit X)*, pages 79-86.

(Kohonen, 1988) Teuvo Kohonen. 1988. Self-organized formation of topologically correct feature maps. *Neurocomputing: foundations of research*, pages 509-521.

(Kohonen, 1995) Teuvo Kohonen. 1995. Self-organizing Maps. *Springer-verlag*.

(Kunchukuttan, 2007) Anoop Kunchukuttan. 2007. Multiword Expression Recognition. *Indian Institute of Technology*, Bombay.

(Kupiec, 1993) Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL'93)*, pages 17-22.

(Kwon *et al*., 2014) Hongseok Kwon, Hyeong-Won Seo, and Jae-Hoon Kim. 2014. Iterative bilingual lexicon extraction from comparable corpora using a modified perceptron algorithm. Journal of *Contemporary Engineering Sciences*, volume 7, issue 24, pages 1335-1343.

(Li *et al*., 2006) Cuiling Li, Hao Zhang, Jian Wang, and Rongyong Zhao. 2006. A new pattern recognition model based on heuristic SOM network and rough set theory. *Vehicular Electronics and Safety*, ICVES 2006, IEEE International Conference on, pages 45-48.

(Lin, 1999) Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 317-324.

(Lu *et al*., 2009) Bin Lu and Benjamin K. Tsou. 2009. Towards Bilingual Term Extraction in Comparable Patents. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'23)*, pages 755-762.

(Manning & Schütze, 1999) Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

(McKeown *et al*., 1996) Kathleen McKeown, Frank Smadja, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, volume 22, pages 1-38.

(Moirón & Tiedemann, 2006) Begoña V. Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06) Workshop on Multiword Expressions in a Multilingual Context*, pages 33-40.

(Monti *et al*., 2011) Johanna Monti, Anabela Barreiro, Annibale Elia, Federica Marano, and Antonella Napoli. 2011. Taking on new challenges in multi-word unit processing for machine translation. In *Proceedings of the 2nd International Workshop on Free/Open-Source Rule-Based Machine Translation*, (URI: http://hdl.handle.net/10609/5646).

(Nag *et al*., 2005) Ashok K. Nag, Amit Mitra and Sharmishtha Mitra. 2005. Multiple Outlier Detection in Multivariate Data Using Self-Organizing Maps Title. Journal of *Computational Statistics*, volume 20, issue 2, pages 245-264.

(Nazar, 2008) Rogelio Nazar. 2008. Bilingual terminology acquisition from unrelated corpora. In

Proceeding of 13th EURALEX Conference (European Association for Lexicography), pages 1023-1029.

(Nunberg *et al*., 1994) Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. In Stephen Everson (Ed.), *Language*, volume 70, issue 3, pages 491-539.

(Paik *et al*., 2001) Kyonghee Paik, Francis Bond, and Shirai Satoshi. 2001. Using multiple pivots to align Korean and Japanese lexical resources. In *Proceedings of the Workshop on Language Resources in Asia Natural Language Processing Pacific Rim Symposium 2001(NLPRS'01)*, pages 63-70.

(Pecina, 2008) Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'08) Workshop: Towards a Shared Task for Multiword Expressions*, pages 54-57.

(Piao & McEnery, 2001) Scott S. Piao and Tony McEnery. 2001. Multi-word unit alignment in English–Chinese parallel corpora. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja (Ed.), In *Proceedings of the Corpus Linguistics 2001 Conference: Lancaster University Cetre for Computer Corpus Research on Language*, technical papers, volume 13 - Special issue, pages 466-474.

(Piao *et al*., 2003) Scott S. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the Association for Computational Linguistic (ACL'03) Workshop on Multiword expressions: Analysis, Acquisition and Treatment*, volume 18, pages 49-56.

(Piao *et al*., 2005) Scott S. Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction, *Computer Speech and Language*, volume 19, issue 4, pages 378-397.

(Picchi & Peters, 1998) Eugenio Picchi and Carol Peters. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Gregory Grefenstette (Ed.), *Crosslanguage Information Retrieval*, Kluwer Academic Publishers, pages 81-90.

(Prochasson *et al*., 2009) Emmanuel Prochasson, Emmanuel Morin, and Kyo 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)*, pages 284-291.

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

(Rapp, 1999) Reinhard 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)*, pages 519-526.

(Rayson *et al*., 2009) Paul Rayson, Scott S. Piao, Serge Sharoff, Stefan Evert, and Begoña V. Moirón. 2009. Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, volume 44, issue 1-2, pages 1-5.

(Ramisch *et al*., 2008) Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008) Workshop: Towards a Shared Task for Multiword Expressions*, pages 50-53.

(Sag *et al*., 2002) Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conferences on Intelligent Text Processing and Computational Linguistics*, 1-15.

(Schafer & Yarowsky, 2002) Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL'02)*, volume 20, pages 1-7.

(Seo *et al*., 2006) Hyung-Won Seo, Hyung-Chul Kim, Hee-Young Cho, Jae-Hoon Kim, and Sung-Il Yang. 2006. Automatically constructing English–Korean parallel corpus from Web documents. In *Proceedings of the 26th Conference on Korea Information Processing Society (KIPS) Fall Conference*, volume 13, issue 2, pages 161-164 (in Korean).

(Seo *et al*., 2013a) Hyeong-Won Seo, Hongseok Kwon, and Jae-Hoon Kim, 2013a. Context-based lexicon extraction via a pivot language. In *Proceedings of the 13th Conference on Pacific Association for Computational Linguistics (PACLING 2013)*.

(Seo *et al*., 2013b) Hyeong-Won Seo, Hongseok Kwon, and Jae-Hoon Kim, 2013b. Rated Recall: Evaluation method for constructing bilingual lexicons. In *Proceedings of the 25th Annual Conference on Human and Cognitive Language Technology (HCLT)*, pages 146-151 (in Korean).

(Seo *et al*., 2014) Hyeong-Won Seo, Hongseok Kwon, Min-Ah Cheon, and Jae-Hoon Kim. 2014. Bilingual multi-word lexicon construction via a pivot language. *Contemporary Engineering Sciences*, volume 7, issue 23, pages 1225-1233.

(Schmid, 1995) Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, UK.

(Shahzad *et al*., 1999) Iram Shahzad, Kiyonori Ohtake, Shigeru Masuyama, and Kazuhide Yamamoto. 1999. Identifying translations of compound nouns using non-aligned corpora. *Workshop Mal Proceedings*, pages 108-113.

(Shin *et al*., 2012) Joon-Choul Shin and Cheol-Young Ock. 2012. A Stage Transition Model for Korean Part-of-Speech and Homograph Tagging. Journal of Korean Institute of Information Scientists and Engineers (KIISE): Software and Applications, volume 39, issue 11, pages 889-901 (in Korean).

(Shirai & Yamamoto, 2001) Satoshi Shirai and Kazuhide Yamamoto. 2001. Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proceedings of the 19th International Conference on Computer Processing of Oriental Language (ICCPOL'01)*, pages 174-179.

(Smadja *et al*., 1996) Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, volume 22, issue 1, pages 1-38.

(Stigler, 1987) Stephen M. Stigler. 1987. The history of statistics: the measurement of uncertainty before 1900. Journal of the Royal Statistical Society Series A (General), volume 150, issue 4, pages 404-405.

(Tanaka & Iwasaki, 1996) Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th Conference on Computational linguistics (COLING'96)*, volume 2, pages 580-585.

(Tanaka & Umemura, 1994) Kumiko Tanaka and Kyoji 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.

(Tony *et al*., 1997) McEnery Tony, Langé Jean-Marc, Oakes Michael, and Véronis Jean. 1997. The exploitation of multilingual annotated corpora for term extraction. In Garside Roger, Leech Geoffrey, and McEnery Anthony (Ed.), *Corpus Annotation - Linguistic Information from Computer Text Corpora, Longman*, London & New York, pages 220-230.

(Tsunakawa *et al*., 2008) Takashi Tsunakawa, Naoaki Okazaki, and Jun'ichi Tsujii. 2008. Building a bilingual lexicon using phrase-based statistical machine translation via a pivot language. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, Posters Proceedings, pages 18-22.

(Uchiyama *et al*., 2005) Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language*, volume 19, issue 4, pages 497-512.

(Venkatapathy & Joshi, 2006) Sriram Venkatapathy and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Association for Computational Linguistic (ACL'06) Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20-27.

(Venkatsubramanyan & Perez-Carballo, 2004) Shailaja Venkatsubramanyan and Jose Perez-Carballo. 2004. Multiword expression filtering for building knowledge maps. In *Proceedings of the Association for Computational Linguistic (ACL'04) Workshop on Multiword Expressions: Integrating Processing*, pages 40-47.

(Villavicencio *et al*., 2007) Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, pages 1034-1043.

(Voorhees, 1999) Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, pages 77-82.

(Vulić *et al*., 2011) Ivan Vulić, Wim De Smet, Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 479-484.

(Wakuya *et al*., 2007) Hiroshi Wakuya, Hiroyuki Harada, and Katsunori Shida. 2007. An architecture of self-organizing map for temporal signal processing and its application to a braille recognition task. *Systems and Computers in Japan*, volume 38, issue 3, pages 62-71. Translated from Denshi Joho Tsushin Gakkai Ronbunshi, volume J87-D-II, issue 3, 2004, pages 884-892.

(Wermter & Hahn, 2004) Joachim Wermter and Udo Hahn. 2004. Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 980-986.

(Wu & Xia, 1994) Dekai Wu and Xuanyin Xia. 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*, pages 206-213.

(Wu, 1997) Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, volume 23, issue 3, pages 377-403.