Minimally Supervised Techniques for Bilingual Lexicon Extraction

Azniah Binti Ismail

PhD

The University of York
Department of Computer Science

September 2012

Abstract

Normally, word translations are extracted from non-parallel, bilingual corpora, and initial bilingual lexicon, i.e., a list of known translations, is typically used to aid the learning process. This thesis highlights the study of a series of novel techniques that utilized scarce resources. To make the study even more challenging, only minimal use of resources was allowed and important major linguistic tools were not employed. Thus, this study introduces some novel techniques for learning a translation lexicon based on a minimally-supervised, context-based approach. The performance of each technique was measured by comparing the extracted lexicon to a reference lexicon based on the F_1 score, which is a weighted average of the precision and the recall. The scores may range from 0 (worst) to 100% (best). Analysis performed on the proposed techniques showed that these techniques had recorded promising F_1 scores, ranging from 57.1% to 80.9%, which indicate moderate and best performances. Overall, the findings of this study further reinforce the use of techniques in exploiting words from small corpora, suggesting that words that are contextually-relevant and occurring in a similar domain are potentially useful. This thesis also presents a technique to deploy extra (i.e., additional) data, which are harvested from the web, and a novel method for measuring similarity of features between two words of different languages without involving the use of initial bilingual lexicon.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I wish to thank my supervisor, Dr. Suresh Manandhar for his invaluable expertise and support rendered throughout the course of this study. I also would like to thank the other members of The Department of Computer Sciences, especially Ioannis Klapaftis, Burcu Can, Shailesh Pandey, Shuguang Li, Matthew Naylor and Michael J.B. for sharing with me their knowledge and technical resources. Last but not least, my heartfelt gratitude to my husband Azhar and our lovely daughters Dina and Nadiah for their love, patience and support during the challenging years.

Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. Some of the research work published during my doctorate has contributed to form a few chapters of this thesis, and they are as follows:

- Chapters 3 and 4
  Ismail, A. and Manandhar, S. (2009). Utilizing contextually relevant terms in bilingual lexicon extraction. *In the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, NAACL-HLT 2009*, Colorado.
- Chapter 5
  Ismail, A. and Manandhar, S. (2010). Bilingual lexicon extraction from comparable corpora using in-domain terms. *In Proceedings of International Conference on Computational Linguistics (COLING) 2010*, Beijing.
- Chapter 6
  Ismail, A. and Manandhar, S. (2011). Multi-word level context features: An approach towards context feature improvement. *In Proceedings of the Artificial Intelligence and Simulation Behaviour (AISB) 2011 Annual Convention*, York.

The thesis work was conducted under the supervision of Dr. Suresh Manandhar in The Department of Computer Science, University of York.

Chapter 1

Introduction

So many languages, so few resources. How to bridge the gap?

Mike Maxwell
Linguistic Data Consortium

To preserve any form of languages in the world, the languages not only needed to be learned and constantly used, but also be documented. Such documentations preserve the languages and become linguistic resources to help describe the embodied structure of the languages. The linguistic resources are essential for further linguistic learning purposes. However, linguistic resources are scarce, which causes problems to many linguistic processing efforts.

This introductory chapter provides a brief explanation of the linguistic resources acquisition problem and its impact on computational linguistic efforts, followed by a brief section about building the linguistic resources automatically. The thesis continues with an introduction to bilingual lexicon extraction that is vital in building bilingual lexicons automatically to preserve important linguistic resources. automatically. The chapter ends with a summary that outlines the thesis structure.

1.1 Linguistic resources acquisition problem and its impact on computational linguistic efforts

In computational linguistic field, linguistic resources especially parallel texts, which are language translations of one another sources, are commonly the stimulus for many areas such as machine translation and information retrieval. For example, parallel texts are used to extract translation probabilities for a machine translation system. According to Al-Onaizan et al. (1999), "the larger the parallel texts available for the training the more improved the performance of a translation system will be". Nonetheless, other linguistic resources such as bilingual lexicons are also important. A bilingual lexicon provides lexical data in form of word-to-word translation pairs. The bilingual data can be used, for example, to assist the translation process in a machine translation system.

For a data-driven system, inadequate amount of linguistic resources in hand could pose a threat to the system. A serious shortage of resources could cause bottleneck. Therefore, acquiring adequate linguistic resources is a crucial task. Though more resources are slowly becoming more available, the acquisition problem can be more serious with low-density languages such as Malay and Hindi compared to mainstream languages like English, French and Arabic.

Both parallel texts and bilingual lexicons are invaluable linguistic resources and they are essential for linguistic processing efforts. However, the process of creating one is arduous, labour intensive and time consuming, especially when linguistic resources are often manually created. Though field linguists have started documenting studied languages aided by tools, the robustness of the process has not been well established thus far. Ultimately, this lack of a refined process has caused linguistic resources to remain scarce despite today's massive information excess. Taking cognizance of this situation (i.e., linguistic resources acquisition problem), computational linguists have started putting efforts in building linguistic resources through automatic systems.

1.2 Automatic building of linguistic resources and its limitation

Building linguistic resources automatically has become a common research topic in natural language processing field. Normally, the task of automatic building requires learning from other linguistic resources. Hence, the task itself may face problems in acquiring linguistic resources.

The process of automatic building may encounter problems as its effectiveness is highly dependent on the quality of the linguistic resources that are being used. Al-Onaizan et al. (1999) and Fung and Cheung (2004b) emphasize that good quality outcomes may only be achieved if good quality linguistic resources with sufficient amount are present. Otherwise, the outcomes of the automated process may not be good. However, although good quality linguistic resources may be available, getting sufficient amount of resources for the automatic building process may be another area of serious concern.

1.3 Introduction to bilingual lexicon extraction

A bilingual lexicon can be defined as "a list of word pairs deemed to be word-level translations" (Haghighi et al., 2008). The word lexicon can be considered as a kind of formatted expanded dictionary that can be read by a computer (Manning and Schütze, 2002). In this thesis, a bilingual lexicon is defined as an expandable machine-readable bilingual dictionary consisting of bilingual word pairs. Each pair of the words generally consists of a word in one language paired with its translated equivalent in another language.

Past studies have shown that bilingual lexicons can be learned automatically and effectively from bilingual, parallel text corpora by applying statistical methods to machine translation (Brown et al., 1990; Koehn and Knight, 2000).

Correlations between lexical types in two different languages are captured using statistical translation models.
The translation models are essentially of word-level, in which, word-level alignment is used to estimate the translation probabilities. The limitation of this approach is that it requires texts to be well-aligned at sentence level before the word-level alignment takes place. Furthermore, the effectiveness of the methods seems to be highly dependent on the availability of sizeable, quality parallel corpora. Otherwise, the approach may not succeed.

From parallel corpora to comparable corpora

Clean parallel corpora consisting of well-aligned texts are essential for bilingual lexicon extraction to be successful. Unfortunately, these crucial linguistic resources are hard to come by for many language pairs. The findings of previous studies (Fung and Cheung (2004b); Munteanu and Marcu (2006)suggest that bilingual, parallel texts should be automatically acquired first before letting a bilingual word pair mapping process to take place. These researchers' methods include an initial step, which will recognize the parallel texts to form clean parallel corpora for further use. However, their methods require a heavy premium in terms of vast linguistic resources and tools for successful implementation.

Some other researchers, including Fung (1995); Koehn and Knight (2001); Rapp (1995), adapted less-supervised approaches by employing comparable texts instead of parallel texts. These approaches would pave the way to eliminating high dependency on parallel texts - a common approach employed by many current efforts.

Monolingual texts may form non-parallel yet comparable corpora. They are far more widely available than parallel corpora because comparable texts can be gathered easily from the Internet. For instance, data from online news articles, encyclopaedia (e.g. Wikipedia), web pages and blogs can be collected to form the corpora.

According to Fung (1995), the characteristics of non-parallel corpora cause lexicon extraction accuracy to be lower than the ones obtained from parallel corpora of similar sizes; hence, the quality of a bilingual lexicon learnt from non-parallel, comparable corpora has become a major concern until today. However, the corpora sources for monolingual texts are very much extensive and much more accessible that have led many researchers attempting to exploit these abundant resources.

Since the quality of the extracted bilingual lexicons is relatively sensitive to the type of corpora used in the extraction process, one would be naturally sceptical with the ideas of building a bilingual lexicon automatically using comparable corpora. For this reason, the use of comparable corpora remains important and valid in bilingual lexicon extraction, which includes the followings:

- Parallel corpora are too scarce and the existing ones does not cover most domains. Because of these limitations, many domains of interest - both generic and specific fields are not easily, readily accessed.
- Most parallel corpora are catered for major languages such as English, Spanish and Chinese. For under-resourced languages such as Malay, Czech and Tagalog, the acquisition problems of parallel corpora are more severe.
- Parallel corpora are not mandatory requirements in bilingual lexicon extraction. Good quality comparable corpora and other extra knowledge resources can be as effective as (or almost on par with) the smaller parallel corpora. This contention is aptly emphasized by Koehn and Knight (2000) Knight (2000) who strongly argue for the learning of good quality comparable corpora that can be as effective as using parallel corpora when the initial bilingual lexicon is sufficiently high to assist the extraction process.
- Less-supervised learning means less extraction time and minimal resource consumption. An effective, fully automated bilingual lexicon

extraction, which is capable of compensating the shortage of linguistic resources, can certainly advance the development of NLP in both specific and general linguistic processing.

Finding comparable corpora of average quality is quite easy as these resources are readily available. On the other hand, finding good quality comparable corpora might not be easy in view of the scarcity of high quality resources. In addition, initial bilingual lexicon that serves as an extra knowledge resource to assist the extraction task is very useful. However, obtaining the initial lexicon of sufficient size (the size here refers to not less than 20,000 entries in each task) is clearly a chicken-and-egg problem. Despite these difficulties, some researchers such as Koehn and Knight (2002) and Haghighi et al. (2008) believe that the extraction of bilingual lexicons from comparable corpora as the best technique will remain debatable.

Figure 1.1 shows a few examples of bilingual word pairs that have been successfully extracted from comparable texts of different language pairs, namely the English-Spanish, English-French, and English-Chinese pairs (Haghighi et al., 2008). Each word pair consists of two words, i.e., a word in one language and its translated equivalent in another language. In the example given, word pairs such as `education`-*`educacion`* and `tourism`-*`turismo`* in the extracted English-Spanish bilingual lexicons mean that the words `education` and `tourism` are in English and, respectively, the words *`educacion`* and *`turismo`* are the Spanish equivalent, respectively. These examples show that comparable corpora are potential resources that could be well exploited in extracting high accuracy bilingual lexicon involving different languages.

An effective bilingual lexicon extraction approach could provide further information on the theoretical framework needed, particularly in assisting the bilingual lexicon acquisition for under-resourced languages. Moreover, techniques that could minimise dependency on heavy resources by using minimal resources, while maintaining high precision of collected word pairs, would

(a) English-Spanish

Rank	Source	Target	Correct
1.	education	educación	Y
2.	pacto	pact	Y
3.	stability	estabilidad	Y
6.	corruption	corrupción	Y
7.	tourism	turismo	Y
9.	organisation	organización	Y
10.	convenience	conveniencia	Y
11.	syria	siria	Y
12.	cooperation	cooperación	Y
14.	culture	cultura	Y
21.	protocol	protocolo	Y
23.	north	norte	Y
24.	health	salud	Y
25.	action	reacción	N

(b) English-French

Rank	Source	Target	Correct
3.	xenophobia	xénophobie	Y
4.	corruption	corruption	Y
5.	subsidiarity	subsidiarité	Y
6.	programme	programme-cadre	N
8.	traceability	traçabilité	Y

(c) English-Chinese

Rank	Source	Target	Correct
1.	prices	价格	Y
2.	network	网络	Y
3.	population	人口	Y
4.	reporter	孙	N
5.	oil	石油	Y

Figure 1.1: Sample outputs for (a) English−Spanish (b) English−French (c) English−Chinese language pairs

Source: Haghighi et al. (2008)

highly benefit the under-resourced languages. However, learning from minimal resources (e.g., limited comparable corpora) as proposed by these methods in best possible ways would be desirable but their feasibility remains unexplored thus far, warranting a detailed, focused study.

1.4 Research scope and objectives

The main goal of the study is to conceptualize and develop novel techniques that could help extract bilingual lexicons with higher precision from minimal resources automatically. The focus is on using non-parallel corpora (which are abundantly available), whilst keeping the amount of these resources to be used to a minimum. Effectively, this condition highlights a worst-case scenario of a scarce problem. In light of this setting, a series of new techniques has to be tested in order to evaluate their performance under extreme conditions. The results of this evaluation could further improve current linguistic extraction practice. In addition, the term *learning* used in this thesis refers to the task performed under minimally supervised way.

Premised on the issues mentioned earlier, several research objectives were formulated to guide the study as follows:

- *To survey previous studies in bilingual lexicon extraction*
  The first objective is to search available previous studies and to gather relevant information on past and current state-of-the-art approaches.

- *To implement the most basic method using minimal resources as a baseline*
  The second objective is aimed toward the development of a baseline for comparison purposes among techniques proposed in this thesis. A basic context-based model is to be built and tested with a slight different setting. One of the systems will be chosen as the baseline in this study based on justifiable criteria of selection.

- *To propose, develop and implement minimally supervised techniques for bilingual lexicon extraction*
  The third objective is mainly to search relevant techniques that would generate higher precision bilingual lexicons from minimal resources compared to the baseline. Each technique would be measured to determine its performance under certain conditions. In addition, this objective is also to demonstrate an appropriate technique that would be able to utilize web data for a bilingual lexicon extraction task under extreme setting. Under this circumstance, the technique would be thoroughly tested to examine its capability in yielding a high precision bilingual lexicon, if possible.

- *To identify problems that may occur in bilingual lexicon extraction task*
  This forth objective is to identify potential limiting factors that would compromise the quality of the extracted bilingual lexicons.

1.5 Contributions

Bilingual lexicons are important linguistic resources to NLP research community, specifically, and to the linguistic society, generally. More importantly, the successful development of techniques capable of automatic extraction of bilingual text of high quality would provide immense contribution to the research community. In this thesis, we conceptualized and developed several new, novel techniques using minimal resources to build the bilingual lexicon. Moreover, we deployed these techniques to exploit relevant words, which were embedded in the corpus. The techniques developed have been tested under extreme conditions, thus reinforcing its robustness, flexibility and endurance.

1.6 Chapters summary

The thesis is structured into seven (7) main chapters, including In overall, the thesis is divided into seven chapters, including the introductory chapter

as detailed in the previous page. The remaining chapters are organized as follows:

- **Chapter 2** provides a brief history of earlier work in bilingual lexicon extraction, followed by sub section pertaining to text corpora to acknowledge their importance in the research. Basic concepts that are applicable to this research field are introduced. This chapter also discusses a number of previous work with greater emphasis on context-based methods for learning high precision bilingual lexicon from non-parallel corpora.

- **Chapter 3** discusses the general framework for a basic bilingual lexicon extraction, built from components that have been identified in the literature chapter. The general approaches for each component are also elaborated. This chapter also discusses the series of experiments that were conducted to decide the best setting for a baseline to be used in this study. In particular, the experiments that compared several systems using high dimensional data and lower dimensional data are also discussed in terms of the practical values of the evaluation results.

- **Chapter 4** presents a novel technique in a context-based algorithm that exploits *contextually-relevant words*. The idea behind this technique is to carefully select the source word and the target word in the initial step. Contextual terms of a word that seem quite relevant to the word are taken into account for the extraction purpose. Evaluation results of the experiments using the model to learn bilingual lexicon extraction from small, comparable corpora are discussed. In addition, an automatic approach to build small initial bilingual lexicon that has been implemented in this study is highlighted.

- **Chapter 5** discusses another novel technique based on a context-based algorithm that exploits in-domain terms to match a source word and its translated equivalent. This chapter elaborates the evaluation results of a number of experiments, which were also performed on small scale English-Spanish comparable corpora.

- **Chapter 6** discusses the methods to acquire small scale, comparable corpora and to harvest more data from the web for unrelated language pairs. In addition, another technique to improve context term lists by using multi-word feature to replace conventional, single word context term is also proposed. This chapter ends with detailed discussion on the evaluations results of experiments using the technique and acquired resources.

- **Chapter 7** discusses the related literature and findings from the current work. Moreover, this chapter highlights the findings learned from the study to provide better understanding of the practical implications based on the implementation of a series of new, novel techniques. This chapter concludes with the recommendations of the researcher for further research in an effort to enhance and improve the current techniques, thus enriching the body of knowledge in the bilingual extraction of text corpus.

Chapter 2

Literature Review

Fellow citizens, we cannot escape history.

Abraham Lincoln (1809-1865)
President of the United States

2.1 Introduction

Bilingual lexicon extraction has been a topic in NLP research for over a decade. A number of methodologies have been developed and many promising contributions have been realised. Research in bilingual lexicon extraction is quite extensive, encompassing a wide spectrum of contexts: learning based on simple frequencies to advance statistics, identical spellings to cognates, and learning from context words to dependency syntaxes. In general, all these techniques are based on certain clues, such as spelling and context of words.

In essence, bilingual lexicon extraction tasks may range from supervised through less-supervised to unsupervised learning. Learning from labelled data such as parallel texts and bilingual lexicons is commonly considered as supervised learning problem. Less supervised learning problems usually use fewer annotated resources but incorporates more unlabelled data such as comparable corpora. On the other extreme, unsupervised learning involves learning from

unlabelled data using appropriate algorithms.

In this regard, learning algorithms may differ for different linguistic resources (Koehn and Knight, 2001). For example, the learning method for parallel texts is usually different from the one for unrelated texts.

This chapter presents a survey on the bilingual lexicon extraction currently being used by many practitioners. Furthermore, the findings of previous work are also described. Basic methodologies and important fundamental concepts are also introduced to highlight aspects deemed critical for this study.

2.2 Research background

Starting from early 1990s, the number of studies on bilingual lexicon extraction has increased. During that period, the niche of the studies was in learning bilingual word pairs from large volumes of parallel corpora through sentence and word alignments.

Earlier research examples include a tool system that was built to help users in identifying technical terms and in supporting translation process (Dagan and Church, 1994). The semi-automatic system used in the previous studies was known as *Termight*. Another example is a translation project work that was been developed at the IBM T.J. Watson Research Center. The project was called *Candide* (Pietra and Pietra, 1994). The project had gained great success; and thus it served as a referential benchmark for other researchers of early studies in bilingual lexicon extraction involving parallel corpora.

According to Fung and Yee (1998), bilingual lexicon extraction is initially revolutionized by automatic term translation using statistical information of word features from clean, parallel corpora. Pascale Fung at Columbia University, New York has developed K-vec, a statistical method, which could be

used to help bilingual lexicon task (Fung and Church, 1994). However, Fung and McKeown (1994) suggest that this method is more suitable for aligning sentence pairs in noisy parallel corpora.

Another previous project to address linguistic extraction needs was carried out by Melamed (1995). This project principally focussed on method that employs sentence-aligned texts together with several filters, including part-of-speech filters, bilingual dictionary filters, cognate filters and word alignment filters. The technique worked well for extracting bilingual word pairs from parallel texts rather than depending solely on the alignments (Melamed, 1995).

According to Rapp (1999), algorithms for aligning words in translated texts have already been well-established by late 1990s. Similarly, the problem of extracting word translations from parallel corpora is also well studied (Melamed, 2000). To support this claim, a number of successful implementations of the methods to extract bilingual lexicon from parallel corpora has appeared prominently in the literature (Melamed (2000); Callison-Burch et al. (2004)).

Studies have now been progressing with other range of corpora. In 1994, Fung and McKeown discovered some word pairs could serve as anchor points for rough alignment of noisy parallel corpora (Fung and McKeown, 1994). An algorithm based on anchor points seems to be applicable for both types of corpora, though learning from parallel corpora is altogether a different problem compared to learning from non-parallel corpora. To address the latter challenge, Fung proposed her first model for non-parallel corpora, which was among the earliest models the following year.

The methods for bilingual lexicon extraction will be presented in detail in this chapter, but the following sub section is presented first to discuss the pivotal impact of the text corpora in the research of many important linguistic resources.

2.3 Main linguistic resources: the text corpora

First, catch your corpus.

Somers (2001)

Obtaining text corpora is typically the first requirement for any knowledge extraction task involving texts. Text corpora are employed in bilingual lexicon extraction to provide lexical and statistical data such as the followings:

- *Lists of vocabularies*: these include the source and target words that form test words, which are initially selected to test an extraction model.
- *Context information*: the information of occurrence frequency of the words surrounding a word pair. In other words, they are the words co-occurring in the context of a source word or a target word in their respective corpora. The translations for the source word should hold similar information with the original source word, i.e., context words co-occurring frequently with the original source word should have translations that co-occur frequently with the translations of the source word.

Generally, text corpora comprise huge collections of texts. Different types of corpora have been used for learning bilingual lexicons. Parallel corpora are common collections of texts that are translations of several linguistic sources. For example, the European parliament proceedings (Europarl) are one of the parallel corpora available. The proceedings were written in many EU languages such as English, Spanish, German and French. Meanwhile, comparable corpora are collections of texts (which are not parallel translations)but they could be related by certain characteristics such as topic, title, event, domain or date. These corpora are also known as non-parallel but comparable corpora. The types of corpora can be categorized in more detail based on the relatedness of the texts. Fung and Cheung (2004a) defined the types of corpora as follows:

- Parallel corpora,
- Noisy parallel corpora,
- Non-parallel but comparable corpora,
- Very non-parallel corpora.

Somers (2001) defined parallel corpus as a text available in two (or more) languages: it may be an original text and its translation, or it may be texts written by a large group of authors in a variety of languages or through co-ordinated international efforts and published in various languages. Parallel corpora serve ideally for various bilingual computational linguistic purposes. Parallel corpora are considered as rich linguistic resources as they are used as the crucial basis for constructing robust bilingual linguistic knowledge resources, including translation model and thesauri (Chen et al., 2004). These corpora also form the basis for techniques such as tokenizing, part-of-speech tagging, morphological and syntactic analysis (Somers, 2001).

On the other hand, noisy parallel corpora contain non-aligned sentences, but they are mostly bilingual translations of the same document. For example, most of these corpora contain documents that are mere rough translations of one another, but the thematic topics (with some insertions and deletions of paragraphs) would be preserved – the focus of extraction is on themes, rather than pure translation per se (Fung and Cheung, 2004b). Therefore, the corpora are considered comparable. Other comparable corpora include those containing texts aligned only by topic. For example, newspaper articles that are collected from two sources of different languages but within the same window of publication date. This type of corpora is better known as non-parallel but comparable corpora.

In contract, very-non-parallel corpora contain far more disparate, very-non-parallel bilingual documents that could be either on the same topic (in-topic)

or not (off-topic). Fung and Cheung (2004b) refer these documents as quasi-comparable corpora, which means the ones that contain non-aligned and non-translated bilingual documents, which could be either on the same topic or not.

Table 2.1: Characteristics of parallel and non-parallel corpora

Parallel Corpora	**Non-parallel Corpora**
Words have one sense per corpus	Words have multiple senses per corpus
Words have single translation per corpus	Words have multiple translations per corpus
No missing translations in the target document	Translations might not exist in the target document
Frequencies of bilingual word occurrences are comparable	Frequencies of occurrences not comparable
Positions of bilingual word occurrences are comparable	Positions of occurrence not comparable

Fung and Yee (1998) distinguish the characteristics of parallel and non-parallel corpora with reference to bilingual lexicon extraction as summarized in Table 2.1. According to Somers (2001), the first characteristic of parallel corpora, i.e., "*words have one sense per corpus*" is often true, especially for words which have terminological status. The second characteristic, i.e., "*words have single translation per corpus* " is a much less safe assumption. The assumption of 1:1 word correspondence is too naive since polysemy, homonym and inflectional problems do occur much or less.

The third characteristic of the parallel corpora, i.e., "*no missing translations in the target document*" is possible; but it is likely to find some parts of the texts to have been omitted from the text of the other language. The fourth characteristic, i.e., "*frequencies of bilingual word occurrences are comparable*" seems to share similar problem with the second characteristic not only due to the grammatical inflection, but other discrepancies such as capitalization

of words functions differently in English and German because all nouns in German need to be capitalized (Somers, 2001).

Finally, according to Somers (2001), the fifth characteristic of the parallel corpora, i.e., "*positions of bilingual word occurrences are comparable*", is the most fundamental assumption for alignment. In contrast, most of the characteristics of non-parallel corpora may be true, especially their fifth characteristic, i.e., about "*positions of bilingual word occurrences are not comparable*". Hence, this contrasting element makes learning from non-parallel corpora becomes harder compared to parallel corpora.
Text corpora can be classified into three main categories: monolingual, bilingual and multi-lingual corpora. Monolingual corpora contain collections of texts of a single language. On the other hand, bilingual corpora contain collections of texts of two languages that can be parallel or comparable. In this context, monolingual corpora may form bilingual corpora, having several features: they are non-parallel, and may be comparable; or they may not be comparable at all. For the third category, multilingual corpora consist of texts of more than two languages.

Thus, this diverse amount of linguistic resources poses increasing challenges to linguistic extraction efforts. Several initiatives and efforts have been carried out to build parallel corpora automatically. For example, Resnik (1999) used a web crawler to find parallel documents from the World Wide Web. Later, Diab and Finch (2000) built text collections by taking outputs from existing machine translation systems, whilst Koehn and Knight (2000) mapped sentence pairs from parliament proceedings. Thereafter, parallel corpora have gradually become available to a sizeable number of mainstream languages such as English, Chinese and Arabic. In some cases, the parallel corpora contain parliament proceedings, such as the Canadian Hansard and Europarl. In addition, the parallel corpora also contain other types of documentary text involving law and government materials. Despite this positive trend, the number of parallel corpora available as well as the types of languages they offer is still

quite sparse given the sheer amount of resources that have yet to be tapped on.

On the other hand, monolingual corpora can be easily built for most written languages. Apparently, a monolingual corpus that comprises text in a single language can be collected at any time. However, many researchers in NLP including Hwa et al. (2006) assert that the quality of the mapping of two words of different languages would depend on the degree of relatedness between the texts in used. Hence, based on this assertion, bilingual extraction of monolingual corpora would yield results of lesser quality compared to similar effort using parallel corpora.

On a positive note, some researchers contend that the quality of bilingual word pairs derived from monolingual corpora can be improved through better techniques. For example, Fung and Cheung (2004b) and Munteanu and Marcu (2006) have proposed models to improve the quality of bilingual word pairs based on an innovative extraction procedure: firstly, extract the parallel texts from the monolingual corpora; and then, use the parallel texts to extract bilingual word pairs. For the method to work effectively, they suggest the use of huge bilingual lexicon to guide the mapping process. (The details regarding the methods in building parallel corpora, as well as comparable corpora, automatically are presented in Chapter 3 and 6).

Other linguistic resources

Bilingual lexicons or dictionaries are actually the second important resources in providing lexical information to corpora (Manning and Schütze, 2002). Other resources that are likely to enrich the corpora are thesauri and encyclopaedias. Alternatively, syntactic and semantic information may also be used in learning bilingual lexicons. In this type of learning, certain linguistic resources, such as part-of-speech (POS) information and syntactic constituent, provide the required syntactic information. Likewise, disambiguation of information that are derived from the WordNet can be used to provide semantic information.

2.4 Different learning tasks

Linguistic resources especially parallel corpora remain scarce for most languages. This scarcity has led researchers to look into different methods of extracting bilingual lexicon using different types of linguistic resources, namely unlabelled data such as comparable corpora. To handle the many varieties of methodologies used, a classification has been adopted to categorise the systems based on the algorithms learned by them. More specifically, each learning algorithm is organized based on the desired taxonomy of the desired outcome (Manning and Schütze, 2002) as follows:

- Supervised learning:
    1. Availability of annotated data/ input-output examples.
    2. Generation of a function that maps the information of the words surrounding a word pair; inputs to desired outputs (by referencing the input-output examples of the function).
- Unsupervised learning:
    1. Unavailability of annotated data (i.e., examples).
    2. Collection of a data set of input objects, which is treated as a set of random variables.
    3. Construction of a model of observations (i.e., joint density) on the data set.

Supervised learning as described by Manning and Schütze (2002) is the actual status known where each piece of data used in training is labelled with corresponding correct outputs. In other words, training data containing examples annotated with some sort of labels are required. These labelled data are usually coded manually by humans. Therefore, this manual production of data is labour-intensive and costly. Nonetheless, supervised learning is highly preferred given the relatively high accuracy results as evidenced from previous work. Examples of supervised learning are those work involving parallel text

with or without bilingual lexicons.

In contrast to the supervised method, unsupervised method involves unlabelled instances such as non-parallel texts. Researchers have a good reason to believe that the information technology revolution will bring forward massive monolingual text resources. About 10 years ago, Koehn and Knight (2002) reported that the World Wide Web alone consisted of over one billion documents and according to Google search engine at the time, the word `directory` occurred more than 42 million times, `empathy` 180,000 times and `reflex` 372,000 times. Today, the number of these documents has increased (and is continually expanding) exponentially. For example, a search for the word `directory` will give 3 billion hits, the word `empathy` 7 million hits, and the word `reflex` 113 million times hits (at the time of this thesis writing).

There has also been a considerable interest in using a combination of minimum annotated data from parallel corpora and a large amount of unlabelled data taken from monolingual corpora. One may want to use this approach for either one of these two conditions: labelling task of high volume data is not affordable, or labelling of available labelled data is precluded due to even higher volume of unlabelled data. This type of approach is known as semi-supervised learning. On the other hand, a weakly-supervised approach refers to learning of lesser annotated data. In this context, the goal is to reduce the cost of creating new annotated corpora by (semi-) automating the process (Fung and Yee, 1998).

Identifying distinctive types of supervision to be used in an algorithm has been seriously discussed as an important methodological issues of the NLP. According to Manning and Schütze (2002), this methodological issue has received special attention in the context of word sense disambiguation. This issue has also received the same attention in other important areas area such as bilingual lexicon extraction and machine translation research. Despite the initial difficulty in this identification task, learning methodologies have been

successfully categorised based on certain characteristics of the corpora that could be exploited by the algorithms (Fung and Yee, 1998). In essence, learning methodologies can be simply divided into the learning methodologies could be simply divided into categories based on the main linguistic resources involved in the learning as follows:

- Learning from parallel corpora
- Learning from parallel corpora and an initial bilingual lexicon
- Learning from monolingual corpora
- Learning from monolingual corpora and an initial bilingual lexicon
- Learning from a mixture of parallel, monolingual and an initial bilingual lexicon

Generally, the main resources are either parallel or monolingual corpora, with or without a bilingual lexicon. Similar categories could be found in Koehn and Knight (2001).

2.4.1 Learning from parallel corpora

Parallel corpora have been used extensively in bilingual lexicon extraction. The reason for the adoption of these corpora is best summed up by Koehn and Knight (2001) who note the use of parallel texts in the word-level translation model as follows:

> ...a word-level translation model is a core element of machine translation; 95% of nouns can be translated within a conventional bilingual lexicon. These models are trained on parallel texts.

In addition, according to Rapp (1995), a word pair that co-occurs more often than expected by chance in the aligned sentence pair is the most likely translations of each other. Furthermore, Rapp also assumes that the co-occurrence patterns in original texts are similar to those in translated texts. As algorithms

for the alignment of words between translated texts are well-established (Rapp, 1999), lexicon acquisition from parallel texts may output a one-to-one word alignment with high accuracy scores.

Earlier Work

In 1994, a tool system was built to help user in identifying technical terms, and also to support a translation process (Dagan and Church, 1994). The system is known as *termight*.

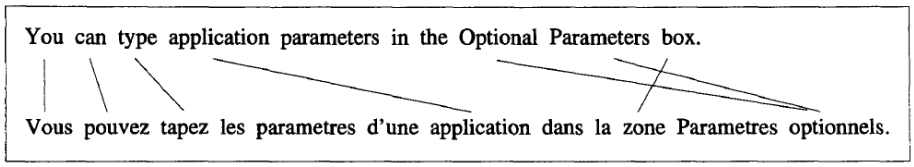

Figure 2.1: An example of word alignment output for the English − French versions of the Microsoft Windows manual. The alignment of `Parameters` to *`optionnels`* is an error.

Source: Dagan and Church (1994)

The termight system obtains a list of source terms and bilingual corpora that are aligned at the word level as input. The system identifies a translation candidate for each occurrence of a source term based on the alignment of the source term. The translation candidate is defined as "*the sequence of words between the first and the last target positions that are aligned with any of the words of the source term*" (Dagan and Church, 1994). For example, the translation candidate for `Optional Parameters box` is *`zone Parametres optionnels`*. See Figure 2.1. Though the word `Parameters` specifically is aligned incorrectly, *`Zone`* and *`optionnels`* are the first and last French words that are aligned with the words of the English terms, respectively.

The system employs a series of systematic steps in succession. First, it collects the translation candidates from all occurrences of a source term; it sorts these

translation candidates according to their frequencies; and finally, it presents the outputs to the user (see Figure 2.2). To assist the user, other information such as corresponding concordances is also provided by the system.

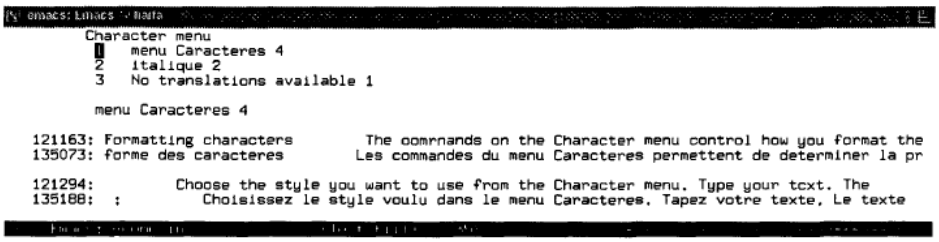

Figure 2.2: A snapshot of Termight screen containing the current term, candidate translations with their frequencies and a bilingual concordance for each translation candidate

Source: Dagan and Church (1994)

In general, the conventional methods currently used would require two alignment steps: (i) a sentence alignment, and (ii) a word alignment (Dagan and Church, 1994; Kaji et al., 1992). Without clean, parallel corpora, a good sentence alignment is typically required to locate the sentence pairs. One of the early approaches for aligning parallel texts was based exclusively on sentence lengths (Gale and Church, 1991). This method assumes the followings:

- paragraph alignment is known, and
- sentence alignment is not known.

Another approach is to "anchor" sentence-to-sentence correspondences using similar spelling word pairs or cognates (Melamed, 1999; Simard et al., 1992). However, Simard et al. (1992) caution this implementation wholesale by recommending that it should use cognates "only in situations where the length-based method alone runs into trouble". Furthermore, using cognates alone may not work, but the cognates help locate potential errors to improve the length-based alignment (Simard et al., 1992). The details of the algorithm

could be found in Simard et al. (1992).

A word alignment is generally used to analyse word correspondence in the sentence pairs, hence this approach can be used to help locate word pairs. However, the alignment process at sentence level can be difficult because well-aligned parallel texts are not extensively available. Without adequate aligned-sentence pairs derived from the first alignment step, adapting the word alignment in the second step might be impossible and, consequently the learning might fail.

The advancement in automatic term translation using statistical information of word features derived from clean, parallel corpora may have revolutionized learning methods of bilingual lexicon extraction (Fung and Yee, 1998). For example, a method using pattern matching, which is known as $K - vec$, has been proposed by Fung and Church (1994). In this model, the source and target word candidates have to be first determined. Using the basis of similar distribution, parallel corpora are split into K-text pieces of equal-sized. K-dimensional binary vectors are created for each of the source and the target word candidates. The distributions of each word are recorded in binary vectors $1 \ldots K$. The corresponding flag in the vector for the source word is set whenever a specific text piece in the source language contains the source word. The process is repeated to each source and target word candidates. Finally, a statistical method is used to find the similarities of any two distributions.

The $K - vec$ model can generally be represented by the followings: Let SS_i and TS_j denotes segments that are translations of each other and S and T are word tokens in the source and the target languages, respectively. S and T are translation equivalents if $S \in SS_i$ and $T \in TS_i$. Similar to other models such as introduced by (Gale and Church, 1991) and (Melamed, 1996), the technique used will divide each half of a bilingual corpus into a number of segments, whence each segment is aligned. Two word tokens are deemed as translations if each has occurred in aligned segment pairs. The differences

among the methods may lies in the way used to divide the segments.

Likewise, Fung and Church (1994) used the $K - vec$ system to align noisy French-English parallel text that revealed encouraging results, but the results for extracting translation pairs were not available. On the other hand, Fung and McKeown (1994) reported poor performance of this method when it was used for Japanese-English and Chinese-English parallel corpora. They contend that the problem rests with the K value. More specifically, the measurement becomes unreliable when K is set too small; when the value of K is too high, the signal is lost. Hence, Fung and McKeown (1994) suggest that this method is more suitable for aligning sentence pairs in noisy parallel corpora because it will not perform bilingual lexicon extraction task very well. Likewise, Fung and Church (1994)'s work provides evidence that this method for bilingual extraction based on noisy parallel text should be further improved for greater use. In this regard, Fung and McKeown (1994) proposed another model known as $DK - Vec$ for aligning pairs of Asian/Indo-European noisy parallel texts without sentence boundaries. Essentially, $DK - Vec$ uses frequency, position and recency information as features for pattern matching. In addition, these researchers used Dynamic Time Warping as the matching technique between word pairs. In their experiment, this algorithm produces a small bilingual lexicon, which provides anchor points for the alignment (see Figure 2.3).

Another earlier work was performed in 1995 by Melamed (1995). The work utilised a method employing sentence-aligned texts together with several filters, namely part-of-speech filter, bilingual dictionary filter, cognate filter, and word alignment filter. The first filter removes every translation pair candidate with different part-of-speech tags. The second filter uses machine readable bilingual lexicon to remove all sentence pairs that include "only one part of a lexicon entry pair but not both of the pair" (Melamed, 1995). The cognate filter searches similarities between the translation pairs and ranks them according to their *level of cognateness* using the Longest Common Subsequence Ratio (refer to Sub Section 2.5 regarding word spelling for the detail of the

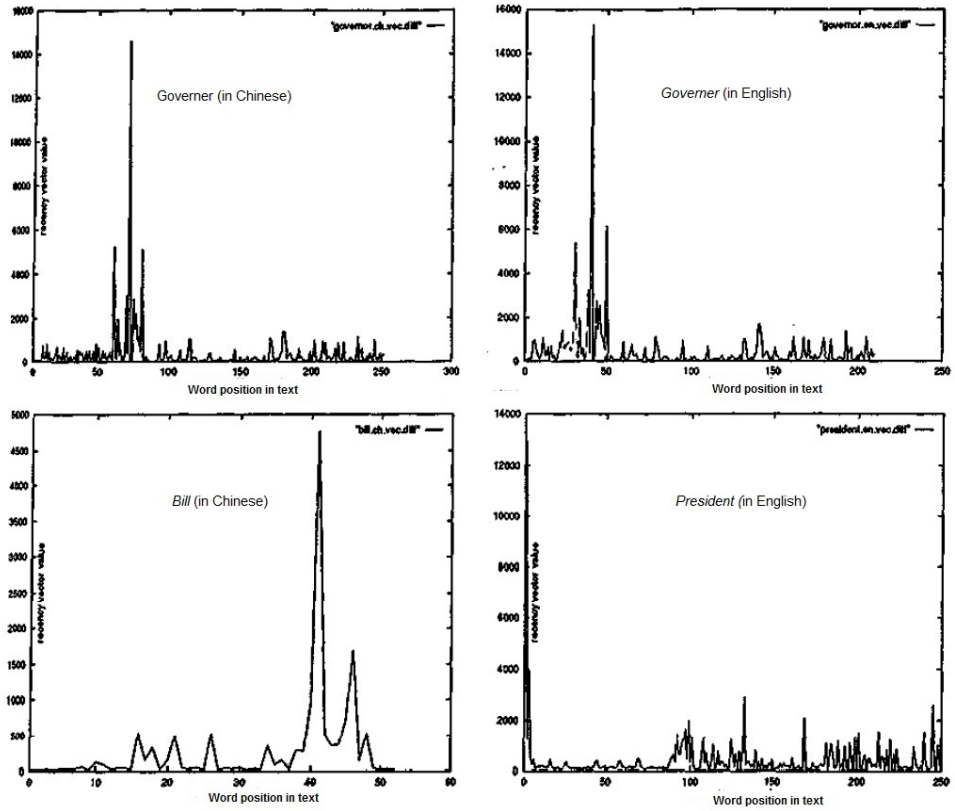

Figure 2.3: High similarity between `Governor` in English and Chinese is shown by DK-vec signals compared to `Bill` (in Chinese) and `President` (in English)

Source: Fung and McKeown (1994)

measurement). Finally, the fourth filter assumes that sentence pairs of related language pairs share similar word order.

The method by Melamed (1995) finds word pairs that satisfy some matching predicate and performs the extraction of bilingual word pairs from parallel texts immensely. However, this method requires huge linguistic resources, i.e., a POS-tagged corpus and a bilingual lexicon for the first filter and the second filter, respectively. Moreover, different tag sets for different languages have introduced problems in the work; hence a common tag set is suggested instead.

Later on, in 1997, Melamed designed an MT method that included a fast algorithm for estimating word translations from parallel corpora (Melamed, 1997). The model incorporated hidden parameters λ^+ and λ^-, and likelihood ratios denoted by $L(a, b)$. $L(a, b)$ represents how likely a and b are translations of each other (Dunning, 1993). λ^+ and λ^- are the probabilities of links for co-occurrence of mutual translations and not mutual translations, respectively. Alternatively, these probabilities can also be interpreted as the percentage of true and false positives (Melamed, 1997). For each co-occurring pair of a and b, the $L(a, b)$ is re-estimated together with hidden parameters λ^+ and λ^-.

In the model, all the parameters are initialized, including λ^+ and λ^-, and likelihood ratios. Likelihoods are set in proportion to co-occurrence frequency $n(a, b)$ and in inverse proportion to their marginal frequencies $n(a)$ and $n(b)$. The likelihood ratios and a *competitive linking algorithm* are used to find a set of "links" among word tokens in parallel, bilingual corpora. The links are used to re-estimate λ^+ and λ^-, and likelihood ratios. The steps that find the sets of links are repeated until the model converges to the desired output.

In the competitive linking algorithm, all likelihood scores $L(a, b) < 1$ are discarded, and a and b are sorted to find those with the highest likelihoods. Token pairs a and b are linked, and the number of links between the tokens, $k(a, b)$, are obtained. Then, all linked word tokens are removed from bilingual

corpora. If there are more a and b in the sorted likelihoods, the process iterates.

The ratio $\frac{k(a,b)}{n(a,b)}$ tends to be high if a and b are mutual translations, causing λ^+ to be called. Otherwise, λ^- will be called if the ratio is quite low. If the translations in the bilingual texts are consistent and the model is accurate, the λ^+ should be near 1 and λ^- should be near 0. Note that λ^+ and λ^- do not need to sum to 1 because they are conditioned on different events.

According to Melamed (1997), the word-to-word model can derive a bilingual lexicon comprising 13 million words from the Canadian Hansards with precision topping at 99% accuracy. The range for (λ^+, λ^-) is from (.43, .000094) to (.78, .00016). (See details in Melamed (1997)).

There are many other efforts for extracting bilingual lexicon from the parallel corpora that can be found, such as in Melamed (2000) and Callison-Burch et al. (2004). Most of the methods used were typically based on word co-occurrence frequencies in parallel texts. The range of accuracy scores was from 77% to nearly 100% (Melamed, 1997; Koehn and Knight, 2001). According to Koehn and Knight (2001), most of the studies were based on parallel corpora indicative of the influence of the success of Candide translation project carried out in the early 1990s.

The Candide translation project was an experimental machine translation system developed at IBM T.J. Watson Research Center, New York. The objectives of the project were as follows: 1. to develop a fully-automatic, large vocabulary, French-to-English translation system, and 2. to develop an interactive translator workstation that will increase the speed and productivity of a human translator (Pietra and Pietra, 1994). The experiment combined both statistical information acquired automatically from bilingual, parallel corpora and linguistic knowledge provided by human experts within a probabilistic framework.

In many respect, many work that have been carried out were based on the noisy channel model which took the view that: "the target sentences are just distorted text of the source language, caused by a translation process" (Brown et al., 1990). To overcome this challenge, three components that can be treated individually are determined in each translation task, namely language modelling, translation modelling and decoding.

For example, the source language is French and the target language is English, both are denoted by f and e, respectively. The translation problem is to translate a sequence of French f to English e. A $p(e \mid f)$ is a model that estimates the conditional probability of any English sentence e from a given French sentence f. The problem here is to find the translation for French that maximises $p(e \mid f)$ as follows:

$$p(e \mid f) = \frac{P(e, f)}{P(f)}$$

Given English input and an English language model $p(e)$, Bayes rule can be used to decode French sentences (Brown et al., 1990) in the following form:

$$\frac{P(e, f)}{P(f)} = \frac{P(e)P(f \mid e)}{\sum_e P(e)P(f \mid e)}$$

The problem is decoded into the prior $p(e)$ and a conditional distribution that models the noise of the channel, $p(f \mid e)$. Since the denominator $p(f)$ is constant over all possible English strings, the above equation is reformulated as follows:

$$\underset{e}{\operatorname{argmax}}\, p(e \mid f) = \underset{e}{\operatorname{argmax}}\, p(e)p(f \mid e)$$

Obviously, the model does not directly perform statistical analysis on the training corpus on how likely an English translation of French input will be. Instead, the word-level translation probabilities $p(f \mid e)$ is used, and these probabilities are learnt from the parallel corpus. First, the most likely word alignments are determined, which are then used to be the base of the word level probability to estimate the $p(f \mid e)$. To simplify their estimation, a count

is performed on the number of occurrences of the English word e and the number of times it is aligned to the French word f.

An alignment a identifies the English word that has originated from the corresponding French word.

$$a = a_1, \ldots, a_m$$

where each $a_j \in 0, 1$

If the English sentence e has l words $e_1, \ldots, e_l$ and the French sentence f has m words $f_1, \ldots, f_m$, there will be $(l+1)^m$ possible alignments. Models can be defined for $P(a \mid e)$ and $P(f \mid a, e)$ as follows:

$$P(f, a \mid e) = P(a \mid e)P(f \mid a, e)$$

and

$$P(f \mid e) = \sum_{a \in A} P(a \mid e)(f \mid a, e)$$

where A is the set of all possible alignments.

In the Candide system, a French sentence is converted into an intermediate structure in which various linguistic components are identified. A structure obtained from the previous step is then transferred to a corresponding structure in English. An English sentence is then synthesized from the intermediate English.

Figure 2.4 shows the overall architecture that is used in the Candide system. This system relies on several important components, but the transfer component plays the most critical role. This emphasis is strongly stressed by Pietra and Pietra (1994) who state that "the heart of the system is the transfer component".

Figure 2.4: The architecture of the Candide system

The Candide's transfer component incorporates the following constituents:

1. *A language model*
   The model that estimates the probability of an intermediate English structure,

2. *A translation model*
   The model that estimates the conditional probability of a French structure from a given English structure, and

3. *A decoder*
   The decoder searches the English structure of a given French structure, which maximizes the product of the language model and translation model probabilities.

Further details of the translation project can be referred to Brown et al. (1990). Some of their project tools have been made freely available, including a word alignment tool system named Giza tool kit, to help other researchers to carry out further studies. Koehn and Knight (2001) reported that they had used Giza tool kit together with a stack decoder to align the German-English word pair. Encouragingly, their effort resulted in 80% accuracy for the English nouns that they had translated from the German. More critically, they observed that at least about 50 ideal occurrences with no ambiguity cases were required for each word pair to have them matched perfectly. Moreover, they concluded that "it is still hard to find perfect word alignments if the process is not limited by a bilingual lexicon especially for rare words in the corpus" (Koehn and Knight, 2001).

Having parallel corpora together with a bilingual dictionary allow the word-level translation to be restricted by context. In this regard, Koehn and Knight (2001) have extracted German-English word pairs using the context feature. For each noun occurrence, they discovered the following context features:

- Up to three words of local context around the target word.
- Any open-class word in the same sentence.
- Any open-class word in the same document.

Open-class words include nouns, verbs, adjectives and adverbs. To find the overall best translation for a word, they used the features to train a decision list (Yarowsky, 1994). With the extra knowledge, almost 90% of accuracy could be achieved (Koehn and Knight, 2001). Comparing the methods, Koehn and Knight (2001) assert that better results could be achieved using parallel corpora together with an initial bilingual lexicon compared to using parallel corpora alone. Koehn and Knight (2001) suggest, for higher accuracy, employing supervised word sense disambiguation technique in the method could be the best answer.

2.4.2 Learning from monolingual corpora

Apparently, robust methods could be offered with the use of parallel corpora. However, the lack of resources of well-aligned and noisy parallel corpora limits the implementation of these methods. This contention is further highlighted by Koehn and Knight (2002) who emphasize that parallel corpora will always be limited resources, especially in different domains. Monolingual texts are–though with some reservations–the most easily available linguistic resources. Fortunately, monolingual corpora can be an alternative to parallel corpora especially if the required extra knowledge is provided, which is an initial bilingual dictionary of sufficient size (Koehn and Knight, 2002).

Using monolingual lexicon alone means purely unsupervised learning. Similar to parallel texts, monolingual texts also provide vast lexical and statistical data. However, being non-parallel (though with large sizes), it is unlikely for two monolingual corpora of different languages to be able to provide a perfect set of context words for both the source word and the target word. In essence, the problem will increase in magnitude when the amounts of comparable corpora decrease. Another problem with using comparable corpora to find translation equivalents is that there is no obvious bridge between the two languages Sharoff et al. (2006).

To make the unsupervised learning feasible, most studies have relied on the assumptions that relate a word with its translation equivalent (Koehn and Knight, 2002; Diab and Finch, 2000; Rapp, 1995). Thus, the most obvious approach is in finding word pairs that are spelled identically or similarly across the languages (Koehn and Knight, 2002). For example, Figure 2.5 highlights a series of word pairs collected by Koehn and Knight (2002) in their experiments. However, this approach that is based on word spelling similarity would not help extend bilingual lexicon so much. The reason for this is because a pair of languages does not have many words of similar spelling across them unless both languages are historically and culturally related, such as loanwords.

Previously, work using an initial bilingual lexicon were of context-based approach. In this regards, Rapp (1999) insists that an initial bilingual lexicon is required to improve accuracy (see Sub Section 2.5 for details). To address this requirement, Rapp developed a model that bridged two monolingual texts using *seed words*. Seed words are known bilingual translations in an initial bilingual lexicon: one side is used to represent the context of the source word, and the other side is used to represent the context of the target word, with respect to the languages. Both sides can be used to bridge the two monolingual texts and map out the word pairs. Essentially, Rapp's work is based on the notion that "*words that co-occur frequently in one language have translations that also co-occur frequently in another language*" (Rapp, 1995; 1999). He

German	English	Score	
Organisation	organization	0.92	correct
Präsident	president	0.90	correct
Industrie	industries	0.90	correct
Parlament	parliament	0.90	correct
Interesse	interests	0.89	correct
Institut	institute	0.89	correct
Satellit	satellite	0.89	correct
Dividende	dividend	0.89	correct
Maschine	machine	0.88	correct
Magazin	magazine	0.88	correct
Februar	february	0.88	correct
Programm	program	0.88	correct
Gremium	premium	0.86	wrong
Branche	branch	0.86	wrong
Volumen	volume	0.86	correct
Januar	january	0.86	correct
Warnung	warning	0.86	correct
Partie	parties	0.86	correct
Debatte	debate	0.86	correct
Experte	expert	0.86	correct
Investition	investigation	0.85	wrong
Mutter	matter	0.83	wrong
Bruder	border	0.83	wrong
Nummer	number	0.83	correct

Figure 2.5: An example of word pairs learnt from monolingual corpora using spelling-based approach

Source: Koehn and Knight (2002)

used such properties to map bilingual word pairs and to fill gaps in an existing lexicon. Likewise, similar efforts carried out by Fung (1995); Fung and Yee (1998) are based on the same principle, which allowed them to add novel word pairs to a lexicon.

In another related work, Koehn and Knight (2000) also proposed the use of a lexicon, together with a corpus in the target language, and a comparable corpus in the source language. However, their approach is similar to an approach that views the corpus in the source language as being the distorted target corpus corrupted by a noisy channel. Based on word-level translation probabilities and a language model, the most likely target word can be determined

for each source word. Furthermore, given parallel corpora, the word-level translation probabilities can easily be estimated. The chosen approach, however, is not a straight forward route–the word-level translation probabilities are needed to estimate the best target word matches without the availability of bilingual word pairs and, at the same time, the bilingual word pairs should be established without the word-level translation probabilities. Hence, Koehn and Knight (2000) used the Expectation Maximization (EM) algorithm to deal with the problem. The algorithm alternatively scores the possible target words for each source word in the expectation step. In the maximization step, it estimates translation probabilities based on that until convergence.

Later, Koehn and Knight (2001) conducted several experiments based on the same models for monolingual corpora. The models assume the availability of many linguistic tools including POS taggers and morphological analyser. In these experiments, they compared the models that used an initial bilingual lexicon with those that did not use any lexicon. They took only nouns into account and found that the first model (those using initial bilingual lexicon) had registered higher accuracy compared to the second model (those without initial bilingual lexicon). The experimental results ranged from 75% to 79% and 11% to 39% for the first and second model, respectively. In summary, Koehn and Knight (2001) contend that a parallel corpus can be replaced with monolingual corpora and a bilingual lexicon. (A survey on previous work that proposed models for monolingual corpora, with or without a bilingual lexicon, is presented in Sub Section 2.6.)

2.4.3 Learning from parallel corpora and monolingual corpora

Another approach in bilingual lexicon extraction is the one involving both parallel and monolingual corpora. More interestingly, some methods used an approach that does not require any external lexical resources; and one of the method was proposed by Otero and Campos (2005). The model developed adopted an approach that would capitalize on the positive aspects of

both parallel corpora and non-parallel corpora, which are high accuracy and high coverage, respectively. Their strategy was to extract a representative set of bilingual correspondences between unambiguous *lexico-syntactic* templates from small parallel corpora. The pairs of bilingual templates were then used as local contexts to extract word translations from comparable, non-parallel corpora. This approach adopted in Otero and Campos (2005)'s model resulted in a better performance by achieving 89% accuracy, which was close to the score reached by the extraction approaches from clean, parallel corpora. (See details of the method in Sub Section 2.6).

In another development, Koehn and Knight (2002) developed a model that had been adopted from an earlier model proposed by Mann and Yarowski. As a starting point, this model utilizes a bootstrapping technique by using an initial decision list trained on supervised data. By labelling new word occurrences in a monolingual corpus, it allows more evidence to be collected and enable the construction of a superior decision list. Koehn and Knight (2002) attempted to replicate the technique, but their effort was unsuccessful because nearly all ambiguous German words used in the experiment had strong majority translations. Essentially, the algorithm of the model quickly converged to a decision list that would always predict the majority case, resulting in incorrect translations in most instances. To address this situation, they recommend a larger parallel corpus made up of 50,000 sentence pairs of transcripts and their parallel translations of German news report (DE-NEWS) from 1996-2000 to be used on top of the original training data set.

2.5 Basic concepts of bilingual lexicon extraction

"Showing the word bread *before the word* butter
will speed up the recognition of butter*".*

Nick Milton (1994)
Knowledge Engineer

Understanding the challenges in bilingual linguistic extraction entails a firm grasp on the extraction concepts, which can be used in extracting bilingual word pairs from corpora.

According to the early theory of word recognition in cognitive study in 1969, common words are usually recognised more quickly than uncommon ones due to the word frequency effect. Furthermore, showing some related words before the target word can accelerate the recognition process due to the context effect. For example, the Cognitive System (also known as Context System) computes associative and sentential context of words. The information provided by the system mediates the recognition process. Hence, words semantically associated with the current context are recognised by people more quickly than non-contextual words of the same frequency. However, the theory in the cognitive studies is not entirely rigorous because other parameters also have an impact on people's abilities, such as speed measure in responding to certain stimuli.

Similar theory can also be used in extracting bilingual word pairs from corpora as it also involves the process of recognising some words, i.e., the bilingual translation pairs, which are then computed by machines. The importance of this process of recognition is strongly stressed by many researchers, such as Al-Onaizan et al. (2000); Callison-Burch and Osborne (2003), who say:

> "If a program sees a particular word or phrase one thousand times during training, it is more likely to learn a correct translation than if sees it ten times, or once, or never. Increasing the amount of training material therefore leads to improved quality."

To affirm the above contention, Callison-Burch and Osborne (2003)'s work is very assuring, indicating that one of the ways to measure the differential

effects of the varying amount of data on accuracy. In this regard, Figure 2.6 summarizes their finding, which highlights the relation between accuracy and data size–the former increases when the latter increases. This interesting revelation concerning the theory has shed important clues that help build the foundations of certain approaches or methodologies in research.

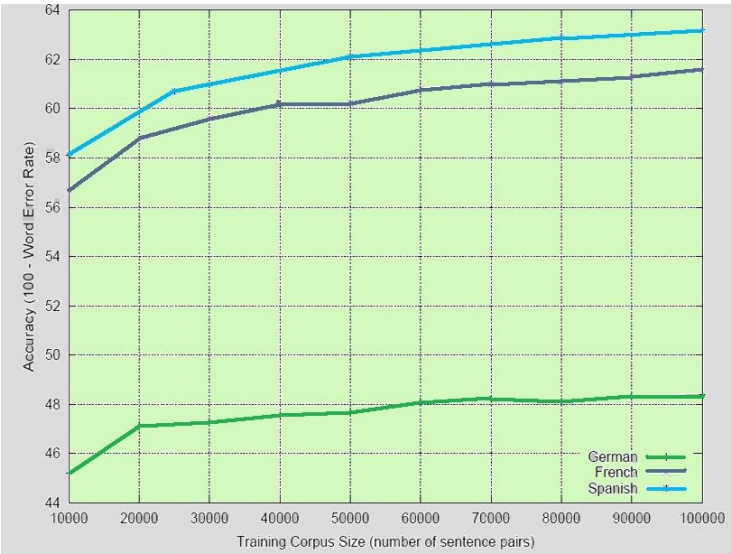

Figure 2.6: Accuracy increases as amount of data increases

Source: Callison-Burch and Osborne (2003)

2.5.1 Extraction clues

Understanding clues that helps the extraction process underscores the imperative of the underlying the mechanism of the process itself in the first place. Bilingual lexicon extraction involves a mapping process of a word in the source language to its translation equivalent in the target language, known as the source word and the target word, respectively. Since the source word and the target word are equivalent, they are expected to share certain mutual characteristics. In a more concrete sense, the extraction process has been eloquently defined by Hwa et al. (2006) as the mapping between two disjoint sets

of symbols.

More importantly, certain clues can help define the characteristics or properties of the words, both the source word and the target word, that are used during learning. Nonetheless, certain clues may be applicable to some types of corpora only. Table 2.2 presents some of the examples of the clues that are generally used in the bilingual lexicon extraction, especially when parallel corpora are involved as correspondence of word and sentence order, correlation between word frequencies and similar spelling word pairs. The table also shows the comparison of the types of corpora, including the monolingual corpora.

For parallel corpora, the correspondence of word and sentence order is usually the strongest, but this is not the case for monolingual corpora as the clue is not applicable (Rapp, 1999). Correlation between word frequencies is not strong compared to the first clue because many words are ambiguous in natural languages, even in parallel texts. For comparable corpora, the clue is still applicable, though with a low reliability; however, the same clue is not useful for unrelated texts.

The third clue, which is similar spelling word pairs, is generally limited to the clear identification of the pairs only. For all other majority pairs, this clue needs to be combined with the first clue to be useful. Similarly, for monolingual texts, the third clue is not useful for identification of the majority of the pairs because the first clue would not work. Hence, the task to extract a bilingual lexicon from monolingual corpora is more difficult because "most statistical clues useful in the processing of parallel texts cannot be applied to non-parallel texts" (Rapp, 1999).

To overcome the above shortcoming,Koehn and Knight (2002) have identified five clues that can be used for the extraction purposes when monolingual texts are involved, which include the following: identical word, similar spelling,

Table 2.2: Extraction clues: their usefulness vs. type of corpora

Statistical clue example	**For parallel corpora**	**For monolingual corpora**
1. Correspondence of word and sentence order	Usually by far the strongest clues.	Not applicable.
2. Correlation between word frequencies	Generally less powerful than the first clue because most of words are ambiguous in natural languages and many ambiguities are different across languages.	For comparable texts the clue is applicable but with a much lower reliability than for parallel texts. For unrelated texts the usefulness may be near zero.
3. Similar spelling word pairs	Generally limited only to the identification of word pairs with similar spelling. For all other (majority) pairs, the clue is usually used in combination with the first clue.	Not useful for identification of the majority of the pairs for both comparable and unrelated texts

contexts, similar words and word frequency. For the purpose of this thesis, the clues have been divided into three major properties, i.e., word spelling, word frequency and word context because the remaining elements are likely to be derived from these three major properties. The descriptions of the three clues are as follows:

- **Word frequency**

  Word frequency is one of the clues shown in Table 2.2. The clue of the word frequency is applicable as long as the texts are comparable. However, the accuracy score may decline greatly due to low reliability of the comparable texts, compared to the parallel texts.

Frequency of words can be useful to help extract bilingual word pairs from ideal parallel texts. The assumption held is: "*the frequencies of word pairs of parallel corpora, especially the most frequent ones, are parallel*". For comparable corpora, frequent words in one corpus should also have their translation equivalents that are also frequent in the other corpus. For example, in English-Malay news corpora, English word `government` is more frequent than `flower`. Respectively, Malay word ***kerajaan*** is more frequent than ***bunga***. While the most frequent word in the target corpus is not necessarily the translation of the most frequent word in the English corpus, the former should also be frequent as the latter. Inevitably, some of the translations might occur less frequent in the other corpus of a target language. Hence, m-th frequent target word cannot be simply aligned with the n-th frequent source word. For most of word pairs, there is a considerable correlation between the frequency of a word and its translation. The frequency is usually redefined as a ratio of the word frequencies normalized by the corpus sizes (Koehn and Knight, 2002).

- **Word spelling**

  Two different languages may contain a number of identical words, especially when both are related. More importantly, both words may originate from the same root, or one of the words may have originated from one of the languages that is later adopted by the target language. This type of words may be adopted exactly; or these words are changed slightly according to some rules or without rules. Nevertheless, this technique may not be able to build a huge repository of word pairs, unless the languages to be paired are closely related with one another, such as English and Spanish. Likewise, the same technique is also applicable if one of the languages has a reasonable number of loanwords taken from the other language. Detail descriptions of the characteristics are as follows:

1. *Identical words*

   Certain number of identical or exact words with the same meaning can be found in two or more languages. Usually, the word is adopted completely (with no translation or modification) into another language; for instance, the English words `hospital` and `pen` are used in Malay in their entirety without any changes in spelling. Another example of words that is adopted exactly is the word `internet`. Thus, the identical words are based on the assumption that the identically spelled words are translations of one another.

2. *Similar spelling, or cognates*

   Some words may have very similar written translations due to their common language roots (e.g., *freund* and `friend`). These words are known as cognates, or adopted words (e.g., *bajet* and `budget`) where the adopted words are derived from another taken into one language from another little translations or minor modifications. Moreover, these words may differ in spelling (even by a very few letters), but these words still maintain similar meaning. As an example, Koehn and Knight (2002) provides a computation that works as follows:

   For a given word pair (`friend, `*freund*), these words share five letters (`fr-e-nd`), and each of them has a word length of 6. Thus, the spelling similarity between them is 5/6, or 0.83. This measurement is called longest common subsequence ratio (LCSR), which has been proposed by Melamed (1995) as follows:

$$LCSR(A, B) = \frac{length(LCS(A, B))}{max(length(A), length(B))}$$

   where

   A and B are the words to be measured, and
   LCS is the longest common subsequence not necessarily continuous in A and B.

From the example, the LCS is equivalent to the five letters (`f`,`r`,`e`,`n`,`d`).

Another measure that can be used to find similarity in spelling is the string edit distance or Levenshtein distance. Compared to LCSR, which only allows addition and deletion operations, Levenshtein distance allows substitution operation on top of the other two operations. However, Haghighi et al. (2008) caution one disadvantage of using edit distance operation precision quickly degrades with higher recall. Instead, they recommend assigning a feature to each sub string of length of three or less for each word and use the set of features to be elements of a word vector, which is ready to be matched with other word vectors in a vector space.

3. *Transliteration*
   Invariably, some English words would appear in foreign language text, especially in science reports or journals. Word pairs may be derived simply by looking for collections of documents in the foreign language containing English words. Most frequent words in the foreign text corpora are likely to be the translation of the corresponding English words. Such approach is language-independent and domain-independent.

The spelling approach may not be suitable if majority of the word pairs to be processed have spelling with little resemblance. (See example of the output in Figure 2.5).

- **Word context**
  Context is defined by the frequencies of context words in the surrounding positions. Words that co-occur in a certain context should also have their translations co-occur in a similar context in the target corpus. Hence, the clue is based on the co-occurrence patterns of words in certain window of words. Rapp (1995) indicates that co-occurrence clue is based

on the assumption that there is a correlation between co-occurrence patterns in different languages.

A context of occurrence for each word j is approximated by bag-of-words that occurs within a window of n-word length or n-word distance. If $n = 2$ the window size is five by considering a neighbourhood of +/- 2 words around the current test word sums up to five words in the window. A context window of a sentence can also be used. Some related examples are discussed in Chapter 3 (see Sub Sub Section 3.3.6 for details).

A context vector of a word j is initially the vector of all words in the bag-of-words. Each word i in this vector is assigned a weight that represents its number of occurrences in that bag-of-words, which is also the number of co-occurrences of word i and j in the same context windows.

The following sentence examples are taken from Rapp (1999):

> "Economy nearer recession after weak growth data."
> "Economy growth is the increase in value of the goods and services produced by an economy."
> "Report shows US economy growth weak if not in recession."
> "How can we increase economy growth in the future?"

Words tend to co-occur frequently in the context of the word `economy` are all underlined in the sentences. Using the above example by Rapp (1999), the English word `economy` co-occurs frequently with `growth` as the German word *`Wirtschaft`* does with *`Wachstum`*. In the English and German context words, Rapp (1999) discovered that the English words `teacher` and `school` co-occur more than expected by chance in the English corpus, which was in sync with their translations in German, i.e, *`Lehrer`* (`teacher`) and *`Schule`* (`school`).

> Interestingly, the clue not only holds for parallel texts but also holds for monolingual texts. The hypothesis is that a pair of words in two separate corpora is more likely to be translation of each other when the distributions of their context words are similar. An initial bilingual lexicon is required to provide translations for the context words. For each word in the corpus, a context vector of co-occurrence statistics pattern between the word and all words in the initial bilingual lexicon, or within certain specified context, is built.

To determine which context words that strongly correspond to a source word or a target word, a measure of association can be used. To compute the similarity between two distributions of context words, a similarity measure should be considered. The most popular concept used in bilingual lexicon extraction is the *vector space model*.

2.5.2 Vector space model

Research based on context similarity usually takes a vector space model into account. This consideration entails firm understanding of important concepts in information retrieval (IR) discipline, which are applicable to bilingual lexicon extraction.

A bilingual lexicon extraction system computes the best word-to-word matching. The aim is to locate a word in one language and its translation in another language. Typically, a context-based bilingual lexicon extraction model consists of three components as follows:

- a source word representation,
- a target word representation, and
- a matching algorithm.

In a vector space, words are modelled as points or elements. The space is generated by a set of basis vectors of context terms of a language. A source word is represented by a vector, which can also be represented as a linear combination of the context term vectors. Each of the context term vectors represents a weighted value for a term indicating its degree of association with the source word. Likewise, a set of target word vector is obtained in the similar form. Using the vectors, each target word in the set is matched against the source word. However, their vectors cannot be compared in a word space since they are consisting of different languages. Hence, one of the vectors of one of the language pair has to be transformed, or translated into the other language of the language pair. For example, the source word vector is translated into the target language. This translation or transformation is very important in helping with the matching process in an initial bilingual lexicon. Once this step is accomplished, the vectors can be compared in the word space.

To measure the matching a simple similarity algorithm based on basic linear algebra can be used on both vectors. The best match, represented by the closest target word vector to the translated source word vector in the target language, is proposed as a translation pair.

According to Manning and Schütze (2002), the vector space model is "one the most used models for ad hoc retrieval, mainly because of its conceptual simplicity and the appeal of the underlying metaphor of using spatial proximity for semantic proximity". Additionally, this model is also the widely used model for bilingual lexicon extraction, especially for tasks using co-occurrence counts collected from a fixed word-window as explained in the previous section.

2.5.3 Similarity measure

Similarity is an important concept in many areas of research, including IR and NLP. A similarity measure is used to assess pairs of objects comparability. The basic notion underlying similarity measures is objects that are

structurally similar are likely to have similar properties usually learnt from corpora. In bilingual lexicon extraction, the similarity between words forms the underlying principle. A similarity between words in the research means two ways: first, it consists of similar spelling, it represents synonyms, and it exhibits similar behaviour such as cat and dog as in groups of animals; secondly, it uses the same context. The latter is the focus in this section, though the former measures may also be applicable as well. Given the complexity of this approach, selecting the most appropriate measure is very challenging as "there is no clear way of deciding the best measure" (Weeds and Weir, 2003), though a few attempts have been explored for a specific task (Weeds and Weir, 2003; Andrade et al., 2010). The following examples highlight the mathematical modelling approaches used for the similarity measure.

Let a set of weighted term t_s that represents a source word s is denoted with X. A set of weighted term t_t that represents a target word t is denoted with Y. The counting measure of $|\ .\ |$ gives the size of the set.

Dot Product

The simplest vector similarity metric is the Dot Product, which is also known as *scalar product*. This metric does not take into account the sizes of vector X (also written as $\vec{x}$) and vector Y (also written as $\vec{y}$) but it considers the inner product between the vectors. The metric can be interpreted over sets simply as $X \cap Y$. Given vector $X = (x_i, \cdots, x_n)$ and vector $Y = (y_i, \cdots, y_n)$:

$$sim(\vec{x}, \vec{y}) = \sum_{i=1}^{n} (x_i \cdot y_i)$$

where

$sim(\vec{x}, \vec{y}) = x \cdot y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$,

x_i and y_i are values of i-th element of X and Y.

The dot product of X and Y may favour frequent words because words with many and large co-occurrence counts usually end up being very similar to most other words. Hence, normalized versions of dot product is highly preferable.

The metric also computes the size of shared terms between x and y over binary vectors, which is the number of $t$rue positives where both values are 1. Hence, the metric may also be interpreted over sets as:

$$\mid X \cap Y \mid$$

Thus, the similarity over binary vectors $\vec{x}$ and $\vec{y}$ can be computed as:

$$sim(\vec{x}, \vec{y}) = \sum_{i=1}^{n} \mid x_i \cdot y_i \mid$$

Cosine Measure

Two vectors that are pointing in a similar direction can be determined by measuring their cosine similarity. When the angle between two vectors is 0, the cosine value is 1. The lowest value of the cosine of another angle is -1.

According to Sahlgren (2006), the Cosine Measure is one of the most used metrics in word space research because it is efficient. In addition, this metric provides a fixed measure of similarity, ranging from 1, 0, and -1 for identical vectors, orthogonal vectors, and dissimilar vectors, respectively. The measure can be interpreted over sets as follows:

$$\frac{X \cap Y}{\mid\mid X \mid\mid \;\; \mid\mid Y \mid\mid}$$

The metric performs the dot product of the vectors and then divide the product by their norms. Normalized vector (or norm) can be achieved by factoring out the effects of vector length:

$$\mid\mid X \mid\mid = \sqrt{X \cdot X} = \sqrt{X^2}$$

Again, for given vectors $X = (x_i, \ldots, x_n)$ and $Y = (y_i, \ldots, y_n)$:

$$sim(\vec{x}, \vec{y}) = cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \sqrt{\sum_{i=1}^{n} (y_i)^2}}$$

where x_i and y_i are values of i-th element of x and y.

Dice Measure

Dice calculates the intersection of two vectors, and the intersection ranges from 0 to 1. Dice measure can be interpreted as:

$$\frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid}$$

In terms of vector operations over binary vectors $\vec{x}$ and $\vec{y}$, the set operations can also be expressed as follows:

$$sim(\vec{x}, \vec{y}) = \frac{2 \mid x \cdot y \mid}{\mid x \mid^2 + \mid y \mid^2}$$

where $\mid x \cdot y \mid$ is the number of true positives.

which actually gives:

- a similar outcome between binary and non-binary vectors, as well as
- a more general similarity metric over vectors in general terms.

Jaccard Measure

The form of Jaccard Measure is quite similar to Dice measure. However, the former has slightly different characteristics. Jaccard computes the similarity between two words represented by term sets X and Y, respectively, by comparing terms that are shared and not shared by the sets. This measure is commonly used to compute similarities between binary vectors.

$$\frac{\mid X \cap Y \mid}{\mid X \cup Y \mid} = \frac{\mid X \cap Y \mid}{\mid X \mid + \mid Y \mid - \mid X \cap Y \mid}$$

Given vector $X = (x_i, \ldots, x_n)$ and vector $Y = (y_i, \ldots, y_n)$:

$$sim(\vec{x}, \vec{y}) = \frac{\mid x \cdot y \mid}{\mid x \mid + \mid y \mid - \mid x \cdot y \mid}$$

where $\mid x \cdot y \mid$ is the number of true positives.

2.5.4 Distance Measure

Distance values can also play an important part in computing the similarity between vectors. The metric can be viewed as the inverse of a similarity measure (Sahlgren, 2006) because of the followings:

- The more similar objects the higher similarity score.
- The nearer distance of objects the lower distance score.

Thus, a distance measure $dist(x, y)$ can be transformed into a similarity measure $sim(x, y)$ with the following transformation formula:

$$sim(x, y) = \frac{1}{dist(x, y)}$$

Euclidean Distance

One of the simple distance metric is the Euclidean Distance that is a linear distance between two points. The metric is measured as:

$$dist(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where x_i and y_i are values of i-th element of x and y.

City-block Metric

Another example of distance metrics, which is even simpler than the Euclidean distance, is the City-block (or Manhattan) metric, which is:

$$dist(\vec{x}, \vec{y}) = \sum_{i=1}^{n} \mid x_i - y_i \mid$$

where x_i and y_i are values of i-th element of x and y.

The metric computes the similarity between vector X and vector Y as "the sum of the absolute differences of corresponding vector positions" (Rapp, 1999). The measure leads to:

- a value of zero for identical matrices,
- a large value in the case when an entry with a non-zero in one matrix always corresponds to an entry with a zero value in the other matrix.

However, frequent words usually end up being too far off from the other words (Sahlgren, 2006). Thus, solving the problem entails the effects of the vector length to be factored out by normalizing the vectors by their norms.

2.5.5 Association measure

An association value indicates the degree of relationship or association between two objects. In bilingual lexicon extraction, the metric is used to weigh a term to see whether it highly co-occurs with a test word in a corpus or not. A group of high co-occurrence terms is usually a good context indicator for a word. There are many methods for measuring the degree of association. The followings are some of the examples:

Simple Co-occurrence Metric (or **Term Frequency**)

Let a_{ij} represents the degree of relationship between context term i and a source word j. The co-occurrence metric is about how many times context term i occur simultaneously with word j. It can be simply determined by their co-occurrence frequency in the corpus within a certain window size. If the frequency is denoted by $freq(i\&j)$, the formula for the weighting will be as follows:

$$a_{ij} = freq(i\&j)$$

This measure is similar to an IR measure known as term frequency, which was used in Fung and Yee (1998)'s study. Fung and Yee used the metric by collecting all words $i$in the context of j and counted their occurrence frequencies. The formula for term frequency tf_{ij} is as follows:

$$a_{ij} = tf_{ij}$$

The metric is informative, but it focuses only on local word co-occurrences. A context term word that highly co-occurs with a source word within a window size is usually a good context indicator for the source word. However, the term is less valuable when it also occurs frequently with most source words in a global collection because it may refer to a general situation. For example, terms such as `virus` and `infection` are highly related to `flu` compared to terms like `discuss` and `fall`, where the latter are commonly premised in a general context (Fung and Yee, 1998).

Inverse Document Frequency

The significance of terms can be emphasized by using the inverse document frequency (IDF). The IDF, as proposed by Fung (1998) is an association metric that is also an original IR method. It takes into account the word occurrences globally, which means it also considers the overall occurrence frequency of a term throughout a corpus. IDF is denoted by idf_{ij}, and a variant formula for IDF is given as follows:

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

where

$freq_{max}$ is the maximum frequency of any words in the corpus,

$freq(i)$ is the total number of occurrences of term i in the corpus.

Term Frequency Inverse Document Frequent

Another term weighting is known as the term frequency inverse document frequent (TFIDF). TFIDF balances the local and global word co-occurrences in the corpora by taking into account the number of times term i co-occurs with word j and the total number of occurrences of term i in a collection. The formula for the TFIDF is as follows:

$$a_{ij} = tf_{ij}. \log \frac{N}{idf_i}$$

where

tf_{ij} is the term frequency of word i in context of word j,

idf_i is the overall occurrence frequency of term i throughout a corpus,

N denotes the total number of words in the corpus.

The $\log \frac{N}{idf_i}$ focuses on global word occurrences. When a word occurs in all documents in a collection, the value of idf_i is equal to the value of N. Hence, the word will not be considered in the computation. (See details in Fung (1998)). Likewise, there are other measures that can be used to define the degree of association between two objects such as *pointwise mutual information* (PMI) and *log-likelihood ratio* (LLR). (See pages 85 and 169 for the descriptions on PMI and LLR, respectively.)

2.6 Previous work

As the focus of this research is addressing the needs for parallel texts, a review of previous work that has proposed a model for non-parallel is essential. There has been a surge in interest to study the issues pertaining to the needs for parallel texts. Moreover, some of these studies have adopted the context-based approach that is relevant to the undertaken study.

Translation pairs that have very similar occurrence frequencies may be widely observed between two parallel corpora. However, this observation is not reliable to be used for learning bilingual lexicons from non-parallel corpora. Hence, to overcome this shortcoming, the context-based approach is introduced, which involves deriving information learned from the context of the source word and the target word. In this regard, many studies that have been carried out thus far are quite diverse in contexts, making inferences more challenging. Moreover, other subtle differences in the details of the studies, such as the resources and measures involved, heightens the intricacies of the approaches used by many of the researchers. The followings are the diverse contexts that are very important in this study.

Context heterogeneity

Fung

Fung (1995) assumes that, "the context heterogeneity of a given domain specific word is more similar to the context heterogeneity of its translation in another language than to any of unrelated word". According to Fung, occurrence frequencies between word pairs across languages in non-parallel corpora are not likely to correspond to each other significantly. Using Fung's text sample as an example, the word `air` in English text occurs 176 times. In contrast, its translation in Chinese text only occurs 37 times. On the other hand, both word pairs are content words in specific domains, indicating that they are used mostly in similar contexts. These two words are not randomly paired with other words and their word bi-grams are limited. This limitation has prompted Fung to take into account the number of unique bi-grams to indicate a degree of heterogeneity between a word and its neighbours in a text.

Fung defines the context heterogeneity vector of a word to be an ordered pair of left heterogeneity LH and right heterogeneity RH in the form of (LH, RH). LH for a word w is given by:

$$LH_w = \frac{a}{c}$$

where

a is the number of different types of tokens immediately preceding w in the text,

c is the number of occurrences of w in the text.

The RH is defined similarly except that its numerator is taken from the number of different types of tokens immediately following w in the text as follows:

$$RH_w = \frac{b}{c}$$

where

b is the the number of different types of tokens immediately following w in the text.

Fung used the Euclidean distance to compute the similarity between context heterogeneity between words across languages. Additionally, Fung removed function words such as `the` and `by` from the texts to increase the context heterogeneity values of many nouns.

Fung tested a method for non-parallel corpora, which were derived from HKUST English-Chinese Bilingual Corpora. The evaluation was performed on 58 English words against 58 Chinese words, deriving from hand-compiled English-Chinese word pairs. In the test, 12 words were correctly mapped to their translations from the top 5 candidates. Additionally, over 50% of the words were correctly mapped to produce a correct translation when Top 10 candidates were tested. To emphasize this approach, Fung suggests adding more linguistic information, such as word order, larger context window and larger non-parallel corpora, as some of the means to improve the measure and to allow the compilation of bilingual lexicons.

Nonetheless, the approach that has been explained earlier is not very practical, given the prevailing condition: many new words are continually being introduced almost on a daily basis. This continual introduction of new words would result in constant changes in corpora from time to time, which inherently alter the number of unique bi-grams of the test words in the related words collection. This is further compounded when the pace of development of one language is different from other languages. More precisely, the probability of having concurrent developments among diverse languages is very low. Furthermore, the results of Fung's study (despite the large comparable corpora and hand-compiled word pairs used) may not be applicable in today's rapidly changing context. Currently, there is no new study based on this context that has been reported in the literature.

Word association

Rapp

Rapp assumes that co-occurrence patterns of words in corpora of different languages are correlated such that "*if two words co-occur frequently in a text of one language then their translations should also co-occur frequently in text of another language*" (Rapp, 1995). He proposed two models based on the assumption for German-English non-parallel corpora.

His first model did not use any linguistic tools such as lemmatizer, POS taggers or an initial bilingual lexicon. In this model, two co-occurrence matrices were constructed, each consisting of equivalent number of English and German vocabularies. The German vocabulary contains selected translations of the English words. Rapp collected the co-occurrence frequencies within 11-word window for each vocabulary from the corpus. In, addition, Rapp also recommended the use of association between words instead of taking the co-occurrence counts directly in order to reduce the effect of word frequency on the co-occurrence counts and to emphasize significant word pairs. The computation used by Rapp for the co-occurrence matrices was performed by modifying each entry using the following formula:

$$A_{ij} = \frac{(freq(i\&j))^2}{freq(i)freq(j)}$$

where

$freq(i\&j)$ is the frequency of co-occurrence of the two words i and j in the corpus,

$freq(i)$ is the corpus frequency of word i, and

$freq(j)$ is the corpus frequency of word j.

To observe the co-occurrence patterns for the English and German vocabularies, Rapp randomly permuted the word order of the German matrix. The number of words that had been shifted to different positions from those in the original German matrix c was identified; however, when $chadreachedavalueof11$,

it was not considered due to the impossibility imposed by the rule. Finally, the similarity between the new German matrix and the English matrix was computed. The process was repeated until a set of 1000 similarity values was available for each value of c. To compute the similarity between the co-occurrence matrices, Rapp used a simple city-block metric.

(a)

		1	2	3	4	5	6
blue	1		•			•	
green	2	•		•			
plant	3		•				
school	4						•
sky	5	•					
teacher	6				•		

(b)

		1	2	3	4	5	6
blau	1		•	•			
grün	2	•				•	
Himmel	3	•					
Lehrer	4						•
Pflanze	5		•				
Schule	6				•		

(c)

		1	2	5	6	3	4
blue	1		•	•			
green	2	•				•	
sky	5	•					
teacher	6						•
plant	3		•				
school	4				•		

Figure 2.7: Dot patterns of the English and German matrices are identical when the word orders in the matrices correspond with one another

Source:Rapp (1995)

Based on the model, Rapp conducted simulation experiments using exactly 100 words of English and 100 words of German vocabularies. To compute the word co-occurrence frequencies, he employed non-parallel texts from: 1. an English corpus of 33 million words that includes texts from the *Brown Corpus*, *Wall Street Journal* and scientific abstracts from different fields, and 2. a

German corpus of 46 million words containing mainly a compilation of newspaper texts such as *Frankfurter Rundschau* and *Mannheimer Morgen*. From Rapp's experimental study, the finding indicates that two matrices will contain identical patterns when the word orders in the matrices correspond with one another (see Figure 2.7).

Following his first study, Rapp (1999) developed a second model, which he found the model had performed remarkably well by achieving 72% accuracy. Based on this finding, he proposes the use of an initial bilingual lexicon containing translations that are known beforehand as anchor points to limit the search space. Though the second model used the same assumption and similarity metric as that of his previous work, Rapp, nonetheless, introduced some other changes to the system, namely linguistic tools such as lemmatizer and morphological analyser. In addition, he also removed all function words from the texts.

The algorithm used by Rapp was based on the vector approach. However, the co-occurrence counting approach was modified from a single co-occurrence vector for each word to several that was exactly one for each position within a window. For example, in a 2-word window, there should be four co-occurrence vectors between word i and j for the following positions: two words ahead, one word ahead, one word behind, and two words behind. The four vectors of length n was combined into a single vector of length $4n$ to keep some information of the word order. Rapp used LLR to compute the association between words in order to obtain information for the vectors.

In his method, an association vector was computed for each source word. Entries from the vector was deleted if their translations were not found in the initial bilingual lexicon. An association matrix was computed for the target words whose: 1. rows were all word types highly occurring in the corpus of the target language, 2. columns are all target words appearing as first translations of the source word in the initial bilingual lexicon. Source word vectors

German test word	expected translation	and rank	top five translations as automatically generated				
Baby	baby	1	baby	child	mother	daughter	father
Brot	bread	1	bread	cheese	meat	food	butter
Frau	woman	2	man	woman	boy	friend	wife
gelb	yellow	1	yellow	blue	red	pink	green
Häuschen	cottage	2	bungalow	cottage	house	hut	village
Kind	child	1	child	daughter	son	father	mother
Kohl	cabbage	17074	Major	Kohl	Thatcher	Gorbachev	Bush
Krankheit	sickness	86	disease	illness	Aids	patient	doctor
Mädchen	girl	1	girl	boy	man	brother	lady
Musik	music	1	music	theatre	musical	dance	song
Ofen	stove	3	heat	oven	stove	house	burn
pfeifen	whistle	3	linesman	referee	whistle	blow	offside
Religion	religion	1	religion	culture	faith	religious	belief
Schaf	sheep	1	sheep	cattle	cow	pig	goat
Soldat	soldier	1	soldier	army	troop	force	civilian
Straße	street	2	road	street	city	town	walk
süß	sweet	1	sweet	smell	delicious	taste	love
Tabak	tobacco	1	tobacco	cigarette	consumption	nicotine	drink
weiß	white	46	know	say	thought	see	think
Whisky	whiskey	11	whisky	beer	Scotch	bottle	wine

Figure 2.8: Results for 20 test words in the German-English translations using a context-based model

Source: Rapp (1999)

were compared to all vectors of the target word association matrix by using a similarity metric. For each source word, the target word was ranked according to the similarity value.

In the experiment, he used a German-English bilingual dictionary that contains 16,000 entries and larger German-English corpora than before, i.e., 135 million words and 164 million words of English and German non-parallel corpora, respectively. (See Figure 2.8 for an example of the resulting output of Rapp (1999)'s study).

To conclude, the first model proposed by Rapp may involve a prohibitively expensive computational effort because he assumed there was no bilingual lexicon available. However, the model managed to simulate the patterns of

word associations quite well. The simulation result strongly supports the significance of word associations in bilingual lexicon extraction. For the second model, Rapp employed similar assumption, but on this occasion he presumed the size of the initial bilingual lexicon be reasonably large. Interestingly, the performance of the second model was better than his first model, making the former a reference to other context-based studies that follow. In other word, Rapp's effort in using word association has set a standard of the context-based approach in bilingual lexicon extraction.

Fung

Similar to Rapp's work, Fung (1998)also proposed a model using word association information. She posits the following characteristics of comparable corpora:

- Words having the same topics across languages will have comparable contexts.
- Words existing in the same domain and time period will have comparable usage patterns, e.g., Zipf's Law.

Using an IR approach in her algorithm, Fung used the model known as Convec, which is quite similar to Rapp (1999). In this model, the context of an unknown word in the source language is extracted and treated as a *query*. Likewise, the contexts of all candidate translations in the target language are treated as the *documents*. Translations are earmarked in the document that best matches the *query*. Fung suggests building context vectors for each unknown source word s in the source language and repeat the process to each target word t in the target language. Then, the similarity of both vectors is computed. The output is ranked according the similarity score. N highest ranking t is chosen as the translation candidate for s.

To find the context of unknown words, Fung initially used the IDF. Based on this approach, she used the term frequency to search the relevant context

words, which revealed a very insightful clue: for the test to be effective, the term frequency does not only emphasise on relevant content-specific words but also stress on generic words. The term frequency takes all words with a high occurrence frequency in the context of a test word because it only considers local word co-occurrence. Hence, Fung suggests using the IDF to de-emphasize general usage words.

For the similarity metric, Fung suggests a variant of Cosine measure as follows:

$$sim(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{t}(x_i y_i)}{\sqrt{\sum_{i=1}^{t} {x_i}^2 \sum_{i=1}^{t} {y_i}^2}}$$

where

$x_i = TF_{xi}IDF_i$,

$y_i = TF_{yi}IDF_i$,

The i-th dimension of a vector, x_i or $y_i = TF_iIDF_i$,

The i corresponds to an element in a vector.

Essentially, Fung (1998)'s model takes into account the reliability of the initial bilingual lexicon using a measure known as Confidence Weighting. The mathematical operation involves dividing the sum of the dot product between x and y in the similarity metric by the rank of candidate t proposed for the source word s. In other words, if a word t is the k-th candidate for word s then the sum of dot product is divided by k. Transforming the similarity score will result in the formulation as follows:

$$sim(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{t}(x_i y_i)/k_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \sum_{i=1}^{t} y_i^2}}$$

Fung tested the method on comparable corpora containing various English and Chinese newspaper articles. She used two sets of evaluation: the first evaluation involved 118 English unknown content words against frequent Chinese unknown words, and the second evaluation involved 40 known English words (randomly selected) against 900 known Chinese words. Based on the

score	English	Chinese
0.008421	Teng-hui	登輝 (Teng-hui)
0.007895	SAR	特區 (SAR)
0.007669	flu	流感 (flu)
0.007588	Lei	鴨 (Lei)
0.007283	poultry	家禽 (Poultry)
0.006812	SAR	建華 (Chee-hwa)
0.006430	hijack	登輝 (Teng-hui)
0.006218	poultry	特區 (SAR)
0.005921	Tung	建華 (Chee-hwa)
0.005527	Diaoyu	登輝 (Teng-hui)
0.005335	PrimeMinister	登輝 (Teng-hui)
0.005335	President	登輝 (Teng-hui)
0.005221	China	林 (Lam)
0.004731	Lien	登輝 (Teng-hui)
0.004470	poultry	建華 (Chee-hwa)
0.004275	China	登輝 (Teng-hui)
0.003878	flu	鴨 (Lei)
0.003859	PrimeMinister	建華 (Chee-hwa)
0.003859	President	建華 (Chee-hwa)
0.003784	poultry	梁 (Leung)
0.003686	Kalkanov	珠海 (Zhuhai)
0.003550	poultry	鴨 (Lei)
0.003519	SAR	葉利欽 (Yeltsin)
0.003481	Zhuhai	建華 (Chee-hwa)
0.003407	PrimeMinister	林 (Lam)
0.003407	President	林 (Lam)
0.003338	flu	家禽 (Poultry)
0.003324	apologise	登輝 (Teng-hui)
0.003250	DPP	登輝 (Teng-hui)
0.003206	Tang	唐 (Tang)
0.003202	Tung	梁 (Leung)
0.003040	Leung	梁 (Leung)
0.003033	China	特區 (SAR)
0.002888	Zhuhai	農曆 (Lunar)
0.002886	Tung	董 (Tung)

Figure 2.9: A part of CONVEC output from the mapping of unknown English words unto unknown Chinese words

Source: Fung (1998)

first evaluation setting, most of the unknown English words were correctly matched with their corresponding Chinese words. Figure 2.9 shows some of the matched words of the CONVEC output.

For the second evaluation, the system achieved about 30% of the accuracy when only top 1 candidate (i.e., the first candidate proposed for a source word by the system) was taken into account. In contrast, the accuracy of the translations improved substantially, reaching 70% when top 20 candidates were considered (see the details of the model in Fung and Yee (1998)'s report). Fung and Yee also developed several variants of the cosine similarity to compute similarity between word vectors. They tested the model using English-Chinese non-parallel newspaper texts (downloaded automatically by the two researchers). The corpora consisted of the English newspaper Hong Kong Standard and the Chinese newspaper Mingpao, which were published from December 12, 1997 to December 31, 1997. From the finding of the test, they conclude that the approach will work reasonable well for comparable corpora. In addition, they also suggest bootstrapping as a technique to be used, which is essentially a method of adding high-ranking bilingual word pairs from the output and iterating the extraction process, to yield more bilingual word pairs.

The most striking difference between this model and other models is the confidence weighting. According to Fung, the initial bilingual lexicon is not reliable in establishing "bridges" between non-parallel texts; thus its quality naturally affects the system output. Ambiguity due to different properties, such as a single word having multiple word translation among words across languages is one of the problems. In essence, Fung associates the problem to the task of rearranging and cleaning-up a translation.

Chiao and Zweigenbaum

Chiao and Zweigenbaum (2002) used a context-based technique that aimed

to translate "simple" words which have been listed in a multilingual lexicon, especially for cross-lingual retrieval of medical information. They used the extraction algorithm to translate input queries into target language queries. However, their model would require a large initial bilingual medical lexicon when learning word translations from non-parallel, comparable corpora.

Chiao and Zweigenbaum defined a 3-word window for the word context occurrence counts. The occurrence is approximated by the bag-of-words that occur within the window. They used the frequency, the TFIDF, and mutual information for weighting words that represent the context vectors. They also used two different similarity metrics based on Jaccard or Cosine measures.

abarognosie	abarognosis
abarthrose	abarthrosis
abarticulaire	abarticular
abasie	abasia
abattement	prostration
abaxial	abaxial
abcédé	abscessed
abcès	abscess
abdomen	abdomen, belly
abdominal	abdominal
abdomino-génital	abdominogenital
abdomino-thoracique	abdominothoracic
abdomino-vésical	abdominovesical
abducteur	abducens, abducent

Figure 2.10: An example of 'simple' words of specialized medical term

Source: Chiao and Zweigenbaum (2002)

The algorithm consists of several steps: 1. Each corpus at the non-alphanumeric characters is segmented, stop words are removed, and lemmatization is performed. 2. A vector for each word s and t in their respective languages is created. 3. A context vector $\vec{x}$ for a source word s is transferred into a target language context vector $tr(\vec{x})$ using an existing English-French bilingual

lexicon. If several translations are listed, only the first one is taken into consideration. 4. A similarity score $sim(tr(\vec{x}), \vec{y})$ is computed for each target language context vector $\vec{y}$ based on the 'transformed' version of the source language context vector $tr(\vec{x})$ to rank target words, and 5. A target word t with the highest ranked k is assumed as the best potential translation for the source word s.

Meas.	Weight	Fr word	En word	R	Top 5 ranked candidate translations
Cos.	*cooc*	anxiété	anxiety	1	anxiety .55, depression .45, medication .36, insomnia .36, memory .34
Cos.	*MI*	anxiété	anxiety	1	anxiety .52, depression .44, insomnia .43, medication .41, term .40
Cos.	*tf.idf*	anxiété	anxiety	1	anxiety .54, depression .41, eclipse .33, medication .29, psychiatrist .29
Jac.	*cooc*	anxiété	anxiety	2	memory .21, anxiety .21, insomnia .19, confusion .19, psychiatrist .18
Jac.	*MI*	anxiété	anxiety	12	insomnia .24, memory .23, confusion .23, thought .20, psychotic .20
Jac.	*tf.idf*	anxiété	anxiety	1	anxiety .21, psychiatrist .17, confusion .15, memory .14, phobia .14
Cos.	*cooc*	infection	infection	2	infected .55, infection .52, neurotropic .47, homosexual .43, virus .43
Cos.	*MI*	infection	infection	1	infection .51, infected .48, virus .44, neurotropic .43, std .42
Cos.	*tf.idf*	infection	infection	3	infected .56, neurotropic .49, infection .48, aids .45, homosexual .41
Jac.	*cooc*	infection	infection	1	infection .33, aids .21, tract .17, positive .16, prevention .15
Jac.	*MI*	infection	infection	1	infection .31, aids .19, prevention .17, hiv .17, positive .17
Jac.	*tf.idf*	infection	infection	1	infection .27, aids .24, positive .17, hiv .15, virus .15

Figure 2.11: An example of Top 5 ranked candidate translations for French words *anxiété* and `infection` with methods using different weighting and similarity measures

Source: Chiao and Zweigenbaum (2002)

Chiao and Zweigenbaum tested their model on two medical corpora selected from the Web through the consultation of MeSH-indexed Internet catalogues of medical websites, including the CISMeF, a French language medical website, and CliniWeb, an English language medical website. To obtain comparable corpora, they chose the sub tree domain under the MeSH concept of "Pathological Conditions, Signs and Symptoms" as the best representation in the CISMeF (Chiao and Zweigenbaum, 2002). The selected web pages from CISMeF contained 591,594 word corpus, which yielded 39,875 unique words after lemmatization. They obtained 608,320 words from CliniWeb, which yielded 32,914 unique words after lemmatization. Additionally, they also compiled a French-English lexicon base containing 'simple' words from several sources including:

- an online French medical dictionary, i.e., Dictionnaire Medical Mason, which includes English translations in most of its entries.
- a set of English-French biomedical terminologies from the UMLS metathesaurus (MeSH, WHOART and ICPC).

The resulting lexicon contained 18,437 'simple' word entries, which were mainly specialized medical terms (see an excerpt in Figure 2.10 showing terms with several translations).

The size of the context vector in their experiments was 4,963. The source and target words were among the many words in the context vectors, which means they were all known translations. Their aim was to test whether the expected translation could be differentiated from other context words of the chosen domain.

In this experiment, their method attained 23% accuracy, where the French test words contained the expected translation as the highest ranked words using the MI weighting, computed with either Cosine or Jaccard. In contrast, with a simple term frequency and Jaccard, the results they achieved were just about 20%. Figure 2.11 illustrates the examples of top 5 ranked words for words *anxiété* and `infection`. Accordingly, Chiao and Zweigenbaum (2002) contend that the LLR measure did not prove to be effective in their work; and likewise, the results of City-block measure were too poor to be practical.

Another model proposed by Chiao, Zweigenbaum and Sta was aimed to prune translation alternatives (Chiao et al., 2004). They re-scored the translation candidates in the target language by applying the same translation algorithm in the reverse direction and re-ranked them according to the harmonic mean (HM) score as follows:

$$HM(r_1, r_2) = \frac{2r_1 + r_2}{r_1 + r_2}$$

where

r_1 is the original rank of a target word i given a source word j.

r_2 is the rank of word j for i obtained by the reverse translation module.

For example, given the French word ***nez***, the top 3 translation candidates were: first, `respiration` with its similarity score 0.20155; second, `ear` with its similarity score 0.19018; and third, `nose` with score 0.18652. The rank computed by the reverse method was ∞ for `respiration`, #4 for `ear` and #1 for `nose`. Using harmonic mean score, it gave revised scores of 2, 2.667 and 1.5 respectively; and eventually led to the correct translation `nose` to be ranked first.

Chiao and Zweigenbaum's models are not much different to other work using word associations. The first work compared several models to find the best settings for the second model. Re-scoring translation candidates is the difference in the second model. They assumed large resources especially in medical domain. It would be more interesting to see results using general domain, or any other domain.

Koehn and Knight

A work by Koehn and Knight (2002) deserves special mention because their effort presents a comprehensive work in extracting bilingual lexicons from unrelated monolingual corpora. In their study, they combined the results of several approaches based on the linguistic clues, such as cognates, word frequency, similar context, and preservation of word similarity to find translations of nouns.

They replicated Rapp (1999)'s model for their context-based approach, but instead of using the 2-word window, they collected context occurrences within the 2-noun window, i.e., they used a 2-preceding and 2-succeeding word positions. The approach taken by the two researchers is called positional context

window.

In addition, Koehn and Knight also demonstrated a method to learn some high-quality lexical entries using cognates. After examining their German vocabulary, they found that about 1300 words were similar or identical to the English words. Based on a check over a reference lexicon and, they observed that the mapping was 88% correct. More revealing, they observed that the correctness of identical word mappings is highly dependent on the word length. Thus, the assumption that identically spelled words are translations of each other is not always true.

Table 2.3 shows the accuracy of word pairs versus word length reported by Koehn and Knight. The table shows that the translation accuracy for the identical 3-letter words was 60%; in contrast, for the identical 10-letter words the translation accuracy was 98%. Clearly, for shorter words, the accidental existence of an identically-spelled word in the other language is much higher. For English-German translation, the words include `fee`, `ton`, `art`, and `tag`. Hence, word length can be used to increase the accuracy of the collected word pairs. For example, only words having length 6, or more, should be considered.

To test this method, they acquired two monolingual corpora from news resources as follows: 1. an English corpus derived from the Wall Street Journal published in 1990 till 1992, and 2. a German corpus German News Wire published in 1995 till 1996. Both of these were fairly comparable in terms of their general use based on the orientations and time periods.

They also used a bilingual lexicon generated using their previous spelling-based system containing 1000 German-English word pairs. Koehn and Knight took 9,206 distinct German nouns and 10,645 distinct English nouns from a general German-English bilingual lexicon of 19,782 lexicon entries to be their test words. The matching pairs were checked against the existing bilingual

Table 2.3: Accuracy of identical word pairs vs. length of the words

Length	**Number**	**of words**	**Accuracy**
3	33	22	60%
4	127	48	69%
5	129	22	85%
6	162	13	93%
7	131	4	97%
8	86	4	96%
9	80	4	95%
10	57	1	98%
11	50	3	94%

Source: Koehn and Knight (2002)

lexicon, yielding about 100 correct word pairs.

This work is interesting and applicable because it combines all useful clues that characterize each test word. More importantly, their suggestion to extract identical and similar spelling words from bilingual corpora helped the researchers to implement a system that could obtain an automatic initial bilingual lexicon for a minimal supervised learning approach.

Word coherence

Kikui

Kikui (1998) proposed another model based on the co-occurrence frequencies for monolingual corpora. Essentially, the method includes a disambiguation algorithm that was based on Kikui's suggestion on using *coherence score*. This score is a measure that captures associative relations between two words, which do not co-occur in the corpora.

Kikui assumes that "*two vectors with high proximity are coherent with respect to their associative properties*". He extends the notion to m-words, which

col. no. *word* (Eng.)	...	89 ***shikin*** (fund)	...	468 ***hashi*** (bridge)
ginko (bank:money)	...	483		31
teibo (bank:river)	...	8	...	120

Figure 2.12: An example of co-occurrence frequencies
Source: Kikui (1998)

means if a group of vectors are concentrated, the corresponding words are coherent. If the vectors are scattered, the corresponding words are in-coherent. The concentration of vectors is measured by the average proximity from their centroid vector. Proximity $prox(\vec{x}, \vec{y})$ is given as cosine between the vectors such as follows:

$$prox(\vec{x}, \vec{y}) = \frac{\vec{x}.\vec{y}}{\mid \vec{x} \mid\mid \vec{y} \mid}$$

For a given word set W, its coherence $coh(W)$ is defined as the followings:

$$coh(W) = \frac{1}{\mid W \mid} \sum_{w \in W} prox(\vec{v}(w), \vec{c}(W))$$

where

$\mid W \mid$ is the number of words in W,
$\vec{c}(W)$ is a group of vectors, which can be defined as: $\vec{c}(W) = \sum_{w \in W} \vec{v}(w)$,
$\vec{v}(w)$ is a corresponding word to be tested.

To disambiguate the combination of translation alternatives, the largest $coh(W)$ value is simply selected.

An example of co-occurrence statistics is illustrated in Figure 2.12 that shows the word ***ginko*** (*bank*; a place to deposit money) co-occurring with the ***shikin***

(*fund*) 483 times and with *hashi* (*bridge*) 31 times. The co-occurrence vector of *ginko* contains 483 as its 89th element and 31 as its 468th element. Each row vector represents a word with its relationship values that correspond to each term in the column vectors. This word representation forms a co-occurrence table (or a matrix).

Table 2.4: An example of translation alternatives

Source	**Translations**
bank	ginko (bank:money)
	teibo (bank:river))
interest	risoku (interest:money)
	kyoumi (interest:feeling)

Source: Kikui (1998)

Kikui improved his model by extending the word representation to m-words. As an example, the English word `bank` and `interest` are both related, and both of them have two translation alternatives (see the examples in Table 2.4). From this example, the English word `bank` can be translated into different Japanese words of different meaning, namely *ginko* and *teibo*.

The translation alternatives can be combined to form four translation candidates, namely *ginko, risoku*, *ginko, kyoumi*, *teibo, risoku* and *teibo, kyoumi*. The coherence scores for these four combinations are shown in Table 2.5.

Clearly, any two words occurring in a similar context would be coherent if their corresponding vectors have high proximity. This assertion, however, has problems of high-dimensionality and sparseness of data. In order to solve these problems, Kikui converted the original co-occurrence vector space into a condensed low dimensional real-valued matrix by using Singular Value Decomposition (SVD). For example, in his experiment, he reduced a 20,000-by-1,000

Table 2.5: An example of coherence scores and ranks for combinations of translation alternatives

rank	Candidate	Coherence Score
1	(ginko, risoku)	0.930
2	(teibo, kyoumi)	0.897
3	(ginko, kyoumi)	0.839
4	(teibo, risoku)	0.821

Source: Kikui (1998)

matrix to a 20,000-by-100 matrix.

To obtain translation alternatives, an algorithm called re-translation algorithm can be used. In the re-translation algorithm, a source word is first mapped to a target language by using a source-target (forward) bilingual lexicon. Each translated word is then mapped back to the original language by using a target-source (backward) bilingual lexicon. The unions of the translations from the backward dictionary are the translation alternatives to be disambiguated.

In his experiments, Kikui created a reference lexicon, in which its correctness was manually judged. The reference lexicon was used to compare it against translation output during evaluation stage. He used co-occurrence data extracted from newspaper articles, namely 1994's New York Times article (in English) and 1990's Nikkei Shinbun article (in Japanese). He took a number of the topmost words ranked by their TFIDF scores to be the test words. Using this setting, Kikui's method managed to achieve about 80% of the translation accuracy.

Starkly, there are two major differences in this approach compared to others: 1. the m-word representation is used, and 2. the SVD is used to convert statistic data. The advantage of SVD is that it represents the co-occurrence relationship between two words sharing similar contexts but do not co-occur

in the same text. Unfortunately, according to Kikui, the SVD conversion also weakens the co-occurrence relations. Moreover, Kikui's model relies on the availability of two external bilingual lexicons. Without these bilingual lexicons, it would be difficult to have m-word representations. Under the prevailing constraint, we based our work on minimal resources as implementing similar work would be beyond the scope of the study.

Co-occurrence information of collocate tokens

Diab and Finch

Diab and Finch (2000) proposed a model similar to Rapp (1999)'s model, where both of these models do not require any linguistic tools. Furthermore, Diab and Finch's model assumes that the linguistic resources are not available. Their model's underlying assumption is that "*if two terms have close distributional profiles, their corresponding translation's distributional profiles should be close in a comparable corpora*".

Slightly similar to Rapp's model that uses pattern of word co-occurrences, they used pattern of word relationships, which makes the model dependable on co-occurrence information of collocate tokens. Moreover, Diab and Finch's model also uses a 2-word window, keeping the four collocation positions information separately in a vector. Another important assumption of their model is that punctuation characteristics across the languages are similar.

In their model, a set of vectors with elements of the topmost S tokens from focal N tokens for four collocation positions is built for N most frequent words in the source corpus. Another N x $4S$ contingency table for words is built in the target language. The distance between row vectors in each individual table is measured to find the correlation of a pair of focal token vectors. Then, the paired focal token vectors are mapped between corpora.

Focal token	*P2*				*P1*				*M1*				*M2*			
	pr_1	pr_2	...	pr_S	pr_1	pr_2	...	pr_S	pr_1	pr_2	...	pr_S	pr_1	pr_2	...	pr_S
$focal_1$	f_{11}	f_{12}	...	f_{1S}	f_{11}	f_{11}	...	f_{1S}	f_{11}	f_{12}	...	f_{1S}	f_{11}	f_{12}	...	f_{1S}
$focal_2$	f_{21}	f_{22}	...	f_{2S}	f_{21}	f_{22}	...	f_{2S}	f_{21}	f_{22}	...	f_{2S}	f_{21}	f_{22}	...	f_{2S}
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
$focal_S$	f_{S1}	f_{S2}	...	f_{SS}	f_{S1}	f_{S2}		f_{SS}	f_{S1}	f_{S2}	...	f_{SS}	f_{S1}	f_{S2}	...	f_{SS}
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
$focal_N$	f_{N1}	f_{N2}	...	f_{NS}	f_{N1}	f_{N2}	...	f_{NS}	f_{N1}	f_{N2}	...	f_{NS}	f_{N1}	f_{N2}	...	f_{NS}

Figure 2.13: A contingency table

Source: Diab and Finch (2000)

Diab and Finch's model uses the Spearman rank order correlation (R) to measure the distance between focal token vectors from the contingency table. By using a non-parametric measure of rank correlation, the words are respectively ranked with one another, rather than taking frequencies for the distance measurement.

They conducted a preliminary investigation based on the model. The statistical procedures were applied to within the respective source and target language corpora, and not across them, meaning that they were of the same language. They employed an English corpus of economic genre and divided this corpus into two separate corpora of 40 million words each. Each corpus went through a pre-processing such as tokenizing, token counting and sorting. They used N = 2000 tokens and S = 150 tokens of most frequent words. Hence, the size of the contingency table was 2000 x 600. Figure 2.13 illustrates the size of the contingency table.

Ultimately, unsupervised learning in bilingual lexicon extraction would be the main aim of many studies as Diab and Finch's study has shown some promising results using in this approach. However, their research has used the same particular languages for the source and target language pair in their experiments. Thus, studies across language pairs are urgently needed to shed more light in this field of research.

Lexico-syntactic information

Otero and Campos

Otero and Campos (2005) proposed a model relying on lexicosyntactic templates, which are derived from parallel corpora, to extract word translations from non-parallel corpora. Their model assumes that "*a word s in the source language can be the translation of a word t in the target language if s occurs in local contexts that are translations of the local contexts containing t*".

A lexico-syntactic template is a representation of the local context that contains a set of syntactically-related words. The template, also known as seed context, is provided by a binary dependency. For example, given a seed expression with a binary dependency:

of (import, sugar)

The seed expression provides two templates: the first template of <*import of [NOUN]*> represents the set of nouns that can appear after the word `import of` in the texts, for example, `sugar` and `oil`. The second template is <*[NOUN] of sugar*>, which represents the set of nouns that can appear before `of sugar`.

Table 2.6 shows the examples of binary dependencies and their corresponding templates. Note that *lobj* represents the relationship between a verb and the noun immediately appearing at its left; *robj* is the relationship between a verb and the noun appearing immediately at its right. *ModAdj* is the relationship between a noun and its adjective modifier and *modN* is the relationship between two nouns: the head and the modifier.

Given lexicosyntactic templates, bilingual correspondences between templates from small parallel corpora are extracted. Table 2.7 shows some of the examples of bilingual correlations between templates extracted from the Europarl. Otero and Campos's model seeks to take advantage of the high accuracy statistical information from parallel corpora and high coverage information from

Table 2.6: Examples of binary dependencies and their corresponding templates

Binary dependencies	**Templates**
of (import, sugar)	<import of [NOUN]> <[NOUN] of sugar>
robj (approve, law)	<approve [NOUN]> <[VERB] law>
lobj (approve, president)	<president [VERB]> <[NOUN] approve>
modAdj(legal, document)	<legal [NOUN]> <[ADJ] document>
modN(area, protection)	<protection [NOUN]> <[NOUN] area>

Source: Otero and Campos (2005)

Table 2.7: Examples of bilingual correlations between templates derived from noun-verb dependencies

English	**Spanish**	**Dice**
of <Africa [VERB]>	<frica [VERB]>	0.66
of <agreement [VERB] >	<acuerdo [VERB]>	0.45
of <agriculture [VERB]>	<agricultura [VERB]>	0.50
of <aid [VERB]>	<ayuda [VERB]>	0.57
of <alcohol [VERB]>	<alcohol [VERB]>	0.76

Source: Otero and Campos (2005)

monolingual corpora. Essentially, their model performs two main steps: 1. extract a representative set of bilingual correspondences between unambiguous lexico-syntactic templates from small parallel texts, and 2. use these template pairs as local contexts to extract word translations from comparable, non-parallel corpora.

In their algorithm, text processing is performed on both the parallel and non-parallel corpora, namely tokenizing, lemmatization and POS-tagging. Seed expressions are extracted from parallel texts to create a list of templates t for each language S and T. These templates are used to collect bilingual corre-

spondences using similarity score $sim(t_s, t_t)$.

Given a bilingual list of lexicosyntactic templates, the number of times it instantiates each template of the bilingual list is considered for every word s in the source language, and a vector $\vec{x}$ is built with that information. Similarly, the number of times t instantiates each template of the bilingual list is considered, and another vector $\vec{y}$ is built with this information. Then, the similarity score $sim(\vec{x}, \vec{y})$ is computed using the template pairs to find the word translation pairs.

With syntactic contexts, Otero and Campos's method does not require any association measure, but it uses two similarity measures: 1. the dice coefficient to compute similarity between the template pairs, and 2. the weighted version of Jaccard's coefficient to compute the degree of similarity between two words.

To test their method, Otero and Campos (2005) used two English-Spanish parallel corpora, each containing one million words (which they considered small). Otero and Campos gained a significant achievement when the method had yielded 89% of translation accuracy of the words. Figure 2.14 shows some examples of the vector positions of the feature vector that defines the English word `president` and the corresponding vector positions that defines the corresponding word *`presidente`* in Spanish.

Otero and Campos's method is an example of a syntactic context approach that assumes the availability of linguistic tools for it to be feasible. Moreover, the method is not solely for non-parallel corpora as parallel corpora of the same language pairs of the non-parallel corpora must also exist. Though the parallel corpora used by Otero and Campos only contain one million words, which are considered rather small, but the same could not be said when dealing with under-resourced languages (that is invariably quite voluminous in contents).

president		**presidente**	
00034 <conference of [NOUN]>	323	00034 <Conferencia de [NOUN]>	509
00176 <[NOUN] of council>	218	00176 <[NOUN] de consejo>	1013
00182 <[NOUN] of court>	69	00182 <[NOUN] of tribunal>	54
00234 <[NOUN] of republic>	35	00234 <[NOUN] of republica>	134
00701 <former [NOUN]>	69	00701 <antiguo [NOUN]>	36

Figure 2.14: Excerpts of vector associated with English-Spanish words of `president` and *presidente*

Source: Otero and Campos (2005)

High priority context words

Prochasson et. al

Prochasson et al. (2009) proposed a technique for the context-based approach, in which, context words should be highlighted based on their priority. To address this assertion, they introduced several concepts, namely *anchor points* and specialised vocabularies with three properties: 1. they must be easily identified to allow an automatic process, 2. they must be relevant regarding the corpora topics, and 3. they should be unambiguous (no polysemy) to ensure an efficient characterization.

In their model, they used the anchor point words to improve the discriminative strength of the context vector, thus enhancing the quality of the results. If a pair of anchor points is found between two compared vectors, its similarity score will increase; otherwise, if an anchor point is found only in one of the two compared vectors, their similarity score will decrease.

Prochasson et al.'s model employs depth in flat context vectors by "dispatching association scores of non-highlighted terms on highlighted terms", which means the model lowers the score for the non-highlighted terms and passes this score back to the highlighted elements, keeping a balanced overall score among

context vectors. In essence, their model uses an offset equation as follows:

$$offset_l = \frac{\mid AP \mid_1}{\mid \neg AP \mid_1} X\beta$$

where

AP is the number of anchor points found in the context vector l,

$\neg AP$ is the number of other elements.

β is used to calibrate the importance given to the highlighted elements.

The overall weight is the sum of all association scores for all items of a given vector. The new association measure of element j in the context vector l, $assoc_w eight_j^l$, is given as follows:

$$assoc_w eight_j^l = assoc_l^j + \beta, if j \in AP$$

or,

$$assoc_w eight_j^l = assoc_l^j - offset_l, if j \notin AP$$

The window size is set to 25 words before and after the source word (or the target word in the target language). Computing the similarity between context vectors entails the application of the cosine measure.

In their experiment, two classes of vocabulary were used for specific language pairs, namely Japanese transliteration for Japanese-English language pairs, and scientific compounds for English-French language pairs. For Japanese transliteration, they looked for a loan term from the English language that had been adapted to fit the Japanese language speech sound and scripts. For example, the Japanese term *i-n-su-ri-n* is related to the English word `insulin`.

Scientific compounds for the English-French are words, in both languages, built with specific roots. They can easily be identified from their morphology. For example, the English word `psychology` corresponds to the French word

psychologie with the -y suffix in the English word corresponding to *-ie* suffix in the French word.

To test the method, Prochasson et al. used a French-Japanese dictionary, which had been compiled from four online dictionaries that are freely available such as from http://kanji.free.fr and http://quebec-japon.com. For comparable corpora, they manually compiled several collections of web documents written in English, French and Japanese using the search engine and specialized discourse of scientific topics such as diabetes and nutrition using the PubMed search engine. They found that the results were slightly improved using scientific compounds compared to using transliterations, owing to the low quality transliterations that had been obtained automatically. An average improvement was recorded for the correct translation ranking especially for badly-ranked translations of the top 50 to top 100 candidates. Additionally, many new translations are expected to be discovered and well-ranked translations are less likely to change much.

Prochasson et al.'s finding provides strong evidence in paying particular importance to trusted vocabulary for better performance in translation. Although, the improvement may not be very substantial, but the use of trusted vocabulary would lead to better results when the anchor points are improved. Despite being language dependent, their method used by Prochasson et al. could be extended by replacing the anchor point detection mechanisms with similar-spelled word pairs. However, this replacement may introduce general words rather than specialized words. Nonetheless, the potential benefits of such an approach look promising.

Positively associated context words

Andrade et al.

Andrade et al. (2010) suggest that in a bilingual lexicon extraction model the degree of associations should not be assumed to be very similar across

languages and should not be compared without much pre-processing. Essentially, they recommend a method that involves two main steps as follows: 1. determine the sets of context words that are all positively and significantly associated with the source language and the target language, respectively, and 2. compare these sets in the source language with the sets in the target language. These positively associated context words are called pivot words.

To find the pivot words, Andrade et al. used the smoothed PMI. Given the unreliable estimates of the probabilities for PMI using relative frequency (especially for low frequency words), they determined the uncertainty by defining confidence intervals over PMI values (see the details on page 85). They sampled $p(x \mid y)$ and $p(x)$ independently, and then calculated the ratio of the number of times $PMI > 0$ to determine $P(PMI > 0)$. For measuring the similarity, they suggest the degree of pointwise entropy that has an estimate of m matches to be the basis, which is represented as follows:

$$Information(m, q, c) = -log(P(m))$$

where q and c are the pivot words for the source word and the target word, respectively.

The above formulation lowers the score of candidates with larger feature sets by finding the most likely target candidate word with c pivots that has m matches with a source word, with the latter having q pivots. The process as described is based on the probability computation as follows:

$$P(m) = \frac{\binom{q}{m}\binom{w-q}{c-m}}{\binom{w}{c}}$$

where

w is the total number of pivot words.

$\binom{q}{m}$ is the number of possible combinations of pivots, which the candidate has in common with the query.

The probability $P(m)$ indicates "the number of possible different features sets that the candidate can have such that it shares m common pivots with the query" (Andrade et al., 2010). The smaller the m is the less likely the best m is achieved. In other words, it means that m matching could occur more or less than expected.

Andrade et al. used two sets of evaluation for the English-Japanese translations. The first set used the data consisted of cars complaints taken from the Japanese Ministry Land, Infrastructure, Transport and Tourism (MLIT) and the USA National Highway Traffic Safety Administration (NHTSA). For comparison, they also used less comparable but much larger data from the Japanese newspaper Mainichi Shinbun (published in 1995) and English articles (published in 1997) from Reuters. For the initial bilingual lexicon, they used a large general bilingual lexicons of the Japanese-English dictionary JMDic. Their experiments also involved major pre-processing tasks using POS-tagger and lemmatizer. For the test words, only 100 noun pairs were considered in each experiment, which is considered as an ideal amount, because these test words represent not only as nouns but they also occur frequently in the corpora. Moreover, these nouns are technical terms rather than general terms. In addition, the experiment carried out by Andrade et al. did not include any synonyms.

Their experiment revealed interesting results, especially when considering the improved accuracy of over 10% for the top 1 candidates when compared the system to the accuracies of the best baseline of TFIDF and Cosine. In contrast, the baseline of LLR and Manhattan performed the worst, achieving less than 40% of the accuracy than their proposed model for the top 20 candidates. Interestingly, the results would improve quite substantially when only positively associated pivots are taken into account. They also analysed the association measure separately, showing that their version of PMI had performed slightly better than LLR. Moreover, they found the pointwise entropy had improved

the translation accuracy over the other standard similarity measures.

Word signature

Shezaf and Rappoport

Shezaf and Rappoport (2010) generated bilingual lexicons using pivot language lexicons by relying on bilingual data with a third language, which is called the pivot language, for each of the source language and the target language. According to Shezaf and Rappoport, the pivot language–more often than not–is English. Thus, Shezaf and Rappoport assume the availability of pivot lexicons based on the English language. Prior to their work, this similar approach had been used by Tanaka and Iwasaki (see details in Tanaka and Iwasaki (1996)'s study).

In their study, they introduced the *non-aligned signatures* (NAS), which is a cross-lingual word context similarity score to eliminate incorrect translations from the generated lexicon. Using this technique, an initial noisy lexicon *iLex* containing translation candidates for each source word was generated from two pivot-language lexicons. In their study, the translation in English for the French word *printemps* was `spring`, which is one of the four seasons. In translating the same English word into Spanish, they obtained both the correct and incorrect translations, namely *primavera* and *resorte*, respectively, where the latter refers to an elastic object.

Then, they computed the signatures for each source word and its translation candidates. This computation utilizes monolingual corpora to discover words that are most strongly related to the source word and its translation candidates. In other words, the term signature they used is referring to the context word. The difference is that they did not take co-occurrence counts into account.

The computation for the signatures is based on the Pointwise Mutual Information (PMI) as follows:

$$PMI(w_1, w_2) = log\frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)}$$

where

$Pr(w_1, w_2)$ is the co-occurrence count,

$Pr(w_i)$ is the total number of appearance of w_i in the corpus.

The PMI probabilities can be simply estimated using relative frequencies as follows:

$$PMI(w_1, w_2) = log\frac{\frac{f(w_1,w_2)}{n}}{\frac{f(w_1)}{n}\frac{f(w_2)}{n}}$$

where

$f(w_i)$ is the occurrence frequency of word w_i in the corpus.

$Pr(w_1, w_2)$ is the co-occurrence frequency,

The similarity between each translation candidate's signature G_t and source word's signature G_s is computed by the following NAS measure:

$$NAS_{blex(s,t)} = \frac{\mid w \in G_s \mid blex_w \cap G_t \neq 0 \mid}{N}$$

The similarity is measured by the numbers of context words of the source word s that may be translated to words in the context words of the target word t, normalized by the total number of context words N for the source word. A source-target language initial bilingual lexicon $blex$ is required in this step. Finally, the translation candidates are ranked according to the NAS similarity score. For each headword, the translations with the highest NAS scores are selected as the correct translations.

The main aim of Shezaf and Rappoport's effort was to improve the quality of noisy lexicons generated using pivot lexicons. They tested their model based on two pivot English lexicons (i.e., Hebrew-English and English-Spanish lexicons) that had been compiled by a professional publishing house to generate

a Spanish-Hebrew lexicon. In this test, they used 510 million tokens and 560 million tokens of Hebrew and Spanish corpus data, respectively. Interestingly, their test yielded the correct translations amounting to 82% precision and over 87% for the Hebrew-Spanish lexicon and Spanish-Hebrew lexicon, respectively. They also compared their results against the other baseline systems (i.e., cosine and city block measures) that also use a revised NAS score algorithm. Under similar setting, their algorithm performed relatively better than these baseline systems, attaining increased precision scores–ranging from–about 30% to 40%.

Shezaf and Rappoport's model involves a simple, straightforward approach of context words counting compared to other similarity measures that use cosine or city-block. In many cases, most researchers have translated a word vector of one language and compared the translated vector to a word vector of another language. In contrast, Shezaf and Rappoport's model only counts the context words of the source word if, and only if, their translations occur in the context words of the target word.

In addition to the use of the pivot data and NAS score, the other important difference is the utilization of the multiple alignment for the many-to-many relations in the initial bilingual lexicon of their model. They generated M random possible alignments and took the average distance metric across these alignments. Their model's performance is better than the cosine and the city block measures as the latter models will under-perform as the approach adopted may introduce serious noise. Thus, they suggest taking NAS as a general measure for word similarity among languages. Furthermore, the setting that they have used seems 'ideal' in providing some benefits to the score as not all of the target lexicon words have appeared as translation candidates. However, their approach has resulted in all translation equivalents for each source word appearing in the translation candidate list, underscoring the importance of pivot lexicons for under-resourced languages.

Latent information

Gaussier et. al

Gaussier et al. (2004) proposed three methods inspired by the latent semantic analysis (LSA) to treat problems that are induced by the use of a bilingual lexicon. Inherently, there are two main potential problems in any language translations: 1. coverage of the lexicon, and 2. polysemy and synonymy. If the context of the test word (being either the source word or the target word) is large enough and binding with many general words, a general bilingual dictionary should be used. Nonetheless, a problem may arise if too few words of a corpus are covered by the initial bilingual lexicon. Similarly, other problems may also arise when several lexicon entries share the same meaning (synonymy), or multiple meaning (polysemy).

The first type of lexicon entries (i.e., synonymy entries) poses a serious problem to the standard approach used as this method is incapable to process the similarities of synonyms as one shared term. More poignantly, a context vector for a word may not necessarily be similar to its synonyms as the projections of two synonyms may not be correlated. Similarly, the second type of lexicon entries (i.e., polysemous entries) also poses a problem as these entries may not be present, but the standard approach used will treat them as a single vector in the corpus. Invariably, the context vectors for synonyms are likely to be similar, while the context vectors for polysemys are likely to be dissimilar for most translation pairs of different contexts. In view of these contrasting vectors, Gaussier et al. developed several new methods based on this critical information.

Using the first method, Gaussier et al. computed a vector space $\vec{x}$ where the synonymous lexicon entries were close to each other and performed the translations of all the synonyms to find the target word vector $\vec{y}$. Instead of projecting $\vec{x}$ on a sub space formed by $(s_1, \ldots, s_p)$, Gaussier et al. suggest an

extended method that maps $\vec{x}$ into a new sub space generated by $(\vec{s_1}, \ldots, \vec{s_p})$. In the standard approach, the similarity may be rewritten as follows:

$$sim(x, y) = <\vec{x}, tr(\vec{y})>$$

Whilst, for the extended method, the similarity becomes:

$$sim(x, y) = <SQ_s\vec{x}, TQ_t tr(\vec{y}) >$$

where

S (respectively T) is a translation mapping, which encodes the relations between the source word (respectively target) and the lexicon entries in the source language (respectively target language),
Q_s (respectively Q_t) is the extended mapping in the source side (respectively target side).

The second method proposed by Gaussier et al. uses the canonical correlation analysis (CCA) to identify directions in the source and target views that are maximally correlated or "behave [in] the same way w.r.t. translation pairs". The directions capture the implicit relations between translation pairs via latent semantic axes. Once the first two directions are identified, the process is repeated in the new sub space, which is defined by context vectors of the translation pairs.

In addition, Gaussier et al.'s third third method aims to cluster translation pairs with synonymous words together, while putting translation pairs with polysemous words in different clusters. They employed probabilistic latent semantic analysis (PLSA) to identify bilingual latent classes and used the cosine measure to compute similarities between translation pairs. They also used the Fisher Kernels to derive a similarity measure from a probabilistic model because a direct similarity between observed features is quite difficult to define or to quantify. They used a context window of size 5 or a 2-word window and

the mutual information to measure the association.

They tested their models using an English-French corpus (containing 35,000 English token words and 21,000 French token words), which was provided by the CLEF03 that stored news-wire of the Los Angeles Times (published from May to December, 1994) and *Le Monde*. They also utilized the ELRA multilingual dictionary containing 13,500 entries. In their experiment, only lexical words of noun, verb, adverb and adjective word types were considered.

In their test, the first method (i.e., extended method) yielded high F_1 score, achieving an average precision score at 44% compared to 35% of the standard approach. This improved precision reinforces the use of smaller context windows to reduce the vulnerabilities of the extended approach to serious noise. In other words, the negative effects of serious noise are reduced by using an appropriate window sizes. The positive results of the first test was not replicated in the second and the third methods as no significant improvements were observed. The lack of improvements for the last two methods could be attributed to CCA, which might have introduced serious noise problem because each canonical direction is defined by a linear combination of many different vocabulary words.

Studies to solve synonymy and polysemy problems are rarely attempted by researchers. Thus, Gaussier et al.'s study serves a critical role in learning the many intricacies of dealing with the prevailing problems. More importantly, they have demonstrated that the extended approach would perform better by using high dimensional data than lower dimensional data. This approach, however, might not work with minimal resources due to incomplete data, which is likely to introduce more noise that leads to spurious translations. Moreover, it has been shown that the standard approach for bilingual lexicon extraction is reasonably useful and easy to work with.

Haghighi et. al

Haghighi et al. (2008) proposed three models based on CCA to learn bilingual lexicon from monolingual corpora alone. Their models are based on the following approaches: 1. a context-based approach, 2. a spelling-based approach, and 3. a combined approach (i.e., using context and spelling-based approaches). Interestingly, they employed vectors not only to compute similarity for their context-based model but also for the spelling-based model.

In their study, the algorithm for each of the source words s was paired with each of the target words t to get a set of matching m, where $m \in M$. Unmatched word types were allowed (but not for multiple translations for the source words) to generate a simple model that made it easier for them to make comparison with previous work. In addition, not all of the words were required to participate in the matching. With the matching M the EM steps were performed to back-trace the best matching m and to generate the final valid outputs of bilingual word pairs.

Essentially, the approach in using the EM as the learning algorithm, which has been used by Haghighi et al., is based on a general form consisting of the followings: 1. E-step, which finds the maximum weighted m of all $m \in M$, and 2. M-step, which finds the best parameters θ by performing CCA. In the E-step word vectors are generated and matched. For a starting point hard EM algorithm is used where the best matching is computed under the following model:

$$m = \underset{m'}{\operatorname{argmax}} \, log\, p(m', s, t; \theta)$$

Given a matching m, the M-step optimizes the likelihood of the observed data $log\, p(m, s, t; \theta)$, or in other word, it finds maximum likelihood of the parameters by using CCA. The maximum likelihood can be rewritten as:

$$\max_{\theta} \sum_{(i,j) \in m} \quad log\, p(s_i, t_j; \theta)$$

The E-step is again performed but this time a *maximum weighted bipartite matching* is used. It can be loosely viewed as a PMI measure. The edge weight between source word i and target word t is defined by:

$$w_{i,j} = log\, p(s_i, t_j; \theta) - log\, p(s_i; \theta) - log\, p(t_i; \theta)$$

The objective $log\, p(m, s, t; \theta)$ should be equal to the weight of a matching plus some constant. Edges with wi, j <0 are set to zero. If zero edge is present, the involved words are removed from the matching. The matching pairs are expected to be true translation pairs, hence, the best matching set should give the maximum weight.

Figure **??** shows an illustration of the model proposed by Haghighi et al., which called as the *matching canonical correlation analysis* (MCCA) model. According to the model, if two words, s and t, are truly translations, then it is possible to observe the relation through the latent space. To alleviate noise, Haghighi et al suggest explicit control on the number of edges involved in a matching. They emphasise "a bootstrapping-style approach that only permits edges with high confidence at first, and then slowly permits more over time". They contend on retaining the highest weighted edges in the optimal full matching such that the number of edges will be gradually increased to be retained before executing the EM. In testing this recommendation, Haghighi et al. used several sets of different resources. Their study resulted in precision accuracy of 89% for 33% recall using their best feature set, which was based on English-Spanish corpora containing topically similar but non-parallel sources (i.e., the Europarl).

Based on the premise discussed thus far, Haghighi et al. (2008) and Gaussier et al. (2004) represent a few researchers who have employed CCA in their models. Their studies have revealed many useful findings–in terms of the positive and negative outcomes–that will benefit other researchers and practitioners in lexicon translations field. For example, Gaussier et al. were quite

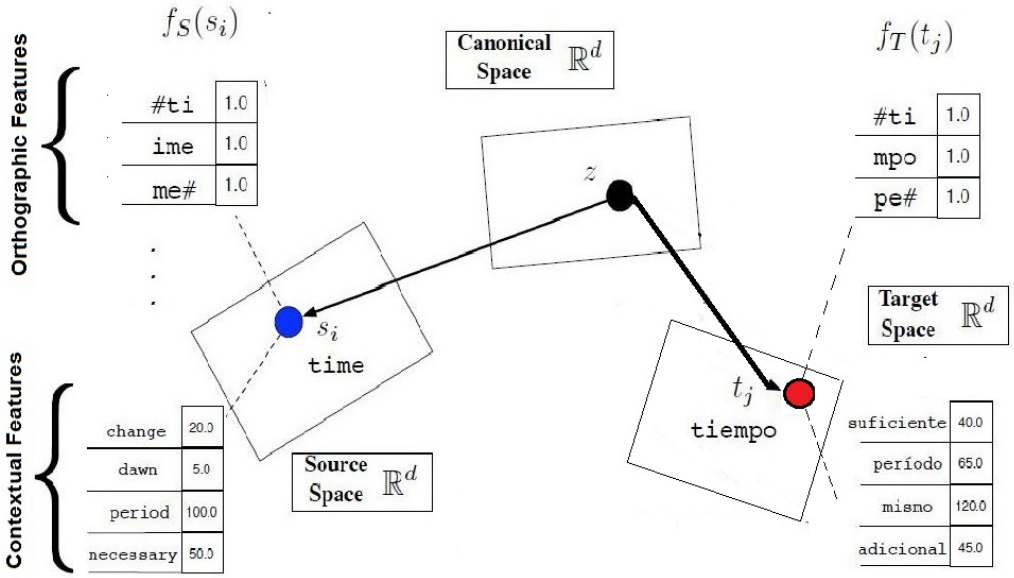

Figure 2.15: An illustration of a combination model of context-based and spelling-based, shown in a canonical space

Source: Haghighi et al. (2008)

concerned with the results yielded with their CCA-based model. To investigate the problem of the CCA-based model, Cancedda et al. (2004) proposed a similar model to Gaussier et al. Unfortunately, Cancedda et al.'s study only focused on other CCA-based systems, leaving out the bilingual lexicon extraction. Hence, their results cannot be compared to Gaussier et al.'s results. In view of this shortcoming, the study undertaken would serve as further, focused study to examine the performance of CCA based on a well-planned setting.

On the other hand, Haghighi et al. recorded a significant success with their CCA-based models. Their finding indicates that by taking 'extremely careful' steps–via back-tracing all potential matchings in the latent space and controlling the number of matching edges–very high precision bilingual lexicon can be generated. Their finding also suggest that in difficult cases, such as the unavailability of seed lexicon, different domains of the monolingual corpora,

and strongly divergent languages, a sizeable set of high precision translation can still be extracted.

Moreover, Haghighi et al were the first to use a probabilistic model, which has produced interesting results across a variety of language pairs and data conditions. Unfortunately, they did not compare their context-based model with the standard approach in the same particular setting. Instead, they compared their model with a basic spelling-based model that uses edit distance. In another setting, they compared their combination model with the standard approach. Given the diverse settings existing in today's prevailing lexicon landscape, further studies are warranted to examine the impacts of translation systems based on comparable, specific settings (i.e., comparing 'an apple' with 'an apple') to ensure findings could be meaningfully interpreted. Thus, the lack of focus on this particular setting has paved the way for the researcher to conduct several experiments based on the CCA approach.

2.7 Summary and conclusion

This chapter discusses the brief history of bilingual lexicon extraction, in particular by focusing on issues related to the most used linguistic resources in the field. Different types of corpora have been elaborated that deal with the use of the resources for different learning tasks. This chapter also discusses the basic concepts that are relevant to the focus of the undertaken study. In addition, a survey on a number of previous work available in this research field is also presented to highlight prevailing conditions which present the many challenges in this research area.

In the early years, learning from parallel corpora was the primary task in bilingual lexicon extraction. Work relying on parallel texts have already recorded significant success. The parallel corpora are mostly constructed manually or semi-automatically, especially by field linguists. These laborious tasks carried

out by the researchers are inevitable as people do not naturally produce the same text in multiple languages, thus rendering parallel corpora to be not readily available. Fortunately, non-parallel corpora or monolingual corpora are easily accessible, which make them a viable alternative to parallel corpora. However, the performance of a system using monolingual corpora is generally lower than a system using parallel corpora in cases where the information available is not extensive. Alternatively, extra information such as external bilingual lexicons may improve the former's performance, but this required additional information can be quite scarce on many occasions.

Other methods with good performance could be identified, but they too have limitation namely being less effective in dealing with the orthographic representation of a word. Overcoming this limitation would entail seeking a more natural means that could find word pairs that are spelled identically across the languages. However, such approach would not help extend the lexicon to a greater height as a pair of languages does not have many words of similar spelling between them, unless both languages share a similar history and posses a similar cultural perspective, notably found in loanwords.

To address these problems, many researchers have attempted to improve the learning from non-parallel corpora by using a context-based approach. This approach could represent a method that is very feasible by exploiting the common patterns observed between word pairs of different languages involving co-occurrence information. In general, the three main steps used in the context-based approaches are as follows:

- Define the context in which a word occurs in the source and target languages.
- Translate as many source context words into the target language using an external bilingual lexicon.

- Compute the translation for each unknown source word using the similarity metric, which uses the target word with the most similar context.

Many of the models that have been proposed thus far are different in many ways, but their main difference lies in the co-occurrence information they acquire from the context. For example, Fung (1995) suggests using the context heterogeneity, the information about the number of different tokens preceding and following a word in a context, whereas many others have suggested using highly associated context words. In addition, some researchers recommend using different existing association measures to find highly associated context words (Rapp, 1999; Chiao and Zweigenbaum, 2002; Koehn and Knight, 2002; Shezaf and Rappoport, 2010). Likewise, there are other researchers who suggest several improved versions of the measures (Andrade et al., 2010), or advocate several new measures (Prochasson et al., 2009).

Likewise, there are several researchers who have used other context-based techniques, which have achieved encouraging results, namely using confidence weighting for each context term (Fung, 1998), re-scoring translation candidates for a source word (Chiao and Zweigenbaum, 2002), and eliminating incorrect translations (Shezaf and Rappoport, 2010). More interestingly, a purely unsupervised work has also been explored by a few researchers (Diab and Finch, 2000).

All the studies carried out by these researchers have their own specific goal. Some of these studies aimed to translate general words (Rapp, 1999; Koehn and Knight, 2002), while some others focused on domain specific terms (Chiao and Zweigenbaum, 2002). In addition, many of these studies tried to translate single-word terms (Haghighi et al., 2008; Diab and Finch, 2000; Otero and Campos, 2005) and others attempted to handle multi-word terms.

As a conclusion, many techniques are available in bilingual lexicon extraction. Some of the existing ideas are interestingly simple, whilst others require rather

complicated techniques. Therefore, which one is the most effective technique is left open to debate –underscoring the imperative in furthering research on exploring new, novel methods. Currently, the context-based approach remains popular in extracting a bilingual lexicon from non-parallel corpora, with most of the methods resorting to the use of an external bilingual lexicon to improve performance.

In this chapter, the main discussion is mainly focused on the methods that mainly weigh the co-occurrences based on context windows. In addition, there is another different way to weigh the co-occurrence by replacing the traditional window-based with a syntax-based co-occurrence counting approach. An example of this approach based on Otero and Campos (2008)'s work is elaborated by a survey carried by the researcher. Another syntax-based work that warrant analyses are the dependency-based counting (Garera et al., 2009), and the positional-based counting approach using POS equivalents (Tanaka and Matsuo, 1999; Tanaka, 2002) among others. These models are effective and efficient to produce precise lexicon translations as Otero and Campos (2008) argue that "syntactic contexts are considered to be less ambiguous and more sense sensitive than contexts defined as windows of size N". Therefore, higher precision should be expected with a syntax-based method compared to a window-based method. However, this contention will be hard to defend as studies that have been performed were mainly based on the availability of extensive external resources and tools, such as lemmatizers and POS taggers. In actual applications, other additional resources such as pivot lexicons or thesaurus have to be made available. Currently, the literature is quite replete with studies that are mostly based on word co-occurrence in certain window, bag-of-work vectors and/or minimal resources requirement. However, research that focuses on minimally supervised work is quite wanting, where robustness and versatility in learning texts from minimal resources are expected to be the forefront of future systems.

Based on the preceding discussion pertaining to previous studies, the author observes the following important aspects that entail further investigation. First, extracting bilingual lexicon from non-parallel corpora represents a new focus in future research as existing lexicon landscape is expected to change based on the fast changing communications taking place in the world. Second, conditions of minimal resources do exist, but they have not been dealt with thoroughly thus far. Third, further compounding these two situations, advanced linguistic tools, such as lemmatizers and POS taggers, which have features prominently in many studies, may not always be available to other researchers. Based on these prevailing constraints, this doctoral study aims to examine the impacts of minimally supervised learning using minimal resources, including latent data, which will represent an important breakthrough by using new, innovative approach in lexicon translation efforts.

In the next chapter, we present our initial experiments' results using that approaches, but first, we review some related approaches or decisions that have been taken by other researchers in their work in order to perform each stage in a general bilingual lexicon work. The sub tasks include, for instance, the resource acquisition and the selection of the lists of vocabularies.

Chapter 3

A General Framework, Related Approaches and Initial Experiments

3.1 Introduction

Chapter 2 described several previous studies that are quite notable in this research field of bilingual lexicon extraction. Most previous studies were based on the context vector approach. Hence, their methods have not so much differences apart from different measures they choose or different external resources they acquire, although there were previous studies that provide interesting techniques in their work; such as giving higher priority to some specialised words in the context of a word (Prochasson et al., 2009) and back-tracing potential matching (Haghighi et al., 2008); among others. From descriptions provided in the previous chapter, we have identified some important components in a bilingual lexicon extraction task to be presented in this chapter.

This chapter presents a general framework, which was formed by several important components; including, corpora, initial bilingual lexicons, context windows and association measures. For each component, related approaches that have been taken by previous studies are discussed in this chapter. Based on the

framework introduced in this study, basic bilingual lexicon extraction systems were built and several initial experiments were conducted, with each experiment used slightly different settings to one another. The results were then compared and discussed to justify the best settings to be chosen as the baseline systems, which would be used in experiments in this study for comparison purposes. Additional experiments involving systems using lower dimensional data were also conducted to compare the systems to the basic systems using high dimensional data. This chapter ends with a summary and a conclusion.

3.2 A general framework

This section introduces a general framework to provide the fundamental requirements for a context-based bilingual lexicon extraction task (see Figure 3.1). This section describes each of components that become the requirements.

To conduct any bilingual lexicon extraction experiment, text corpora are the most critical requirements. The type of corpora used would (indirectly) define the type of learning used in the extraction process, and vice versa. The extraction process usually required two monolingual corpora or a bilingual corpus; these corpora usually composed of two different language pairs.

Corpora require common pre-processing jobs to ensure that the texts can be easily and automatically processed. These pre-processing jobs include tokenizing, removing unimportant formattings and punctuations. Basic pre-processing tools are available for sentence boundary detection and tokenizing, while more advanced tools are required to do major pre-processing jobs such as lemmatization, segmentation or POS-tagging. Many researchers prefer to build their own codes in order to do simple pre-processing jobs such as cleaning the texts from some unimportant structures or formattings, and some of the jobs could be done using regular expressions.

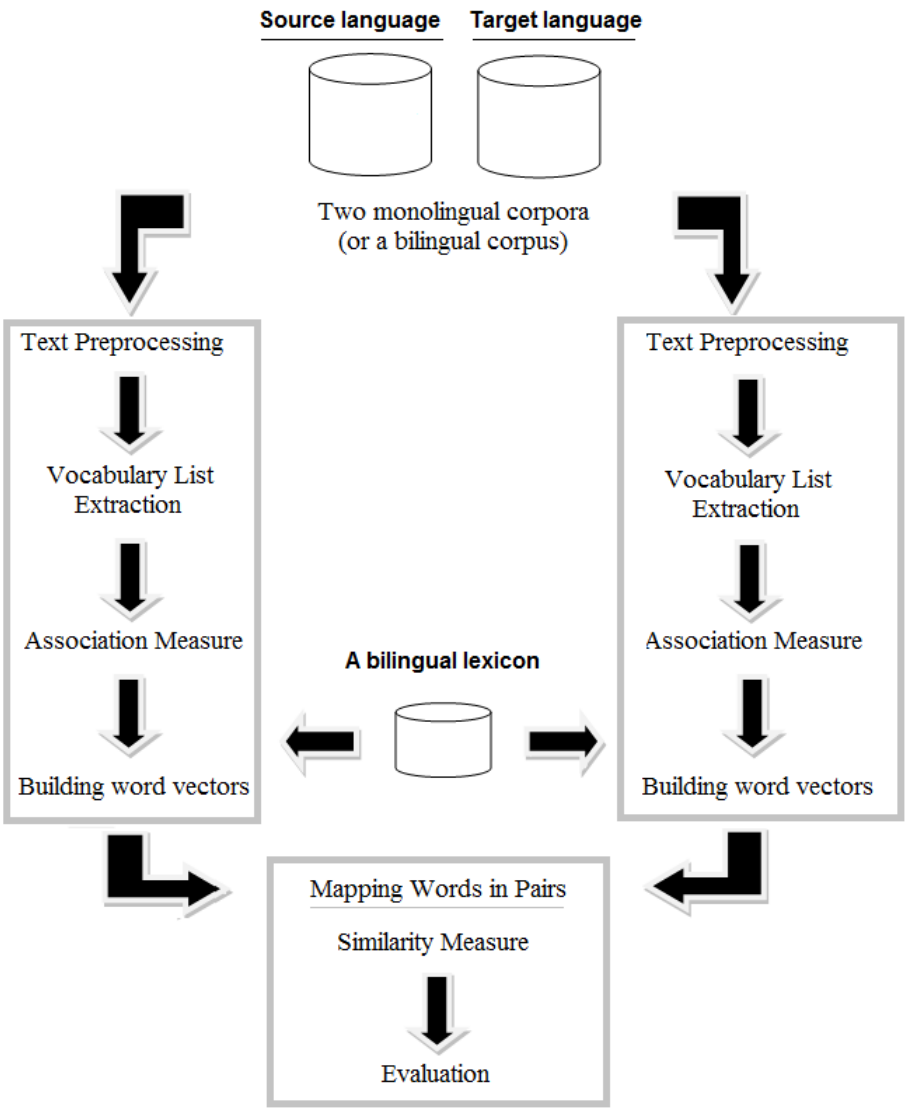

Figure 3.1: A general framework for learning a bilingual lexicon from bilingual corpora

Next requirements are the source and target word vocabulary lists; which would be obtained from the corpora. The lists consist of two separate word lists, namely one list contains the source words and another list contains the target words. Potential translation pairs would be selected among words from these two lists during the matching process.

Once all the requirements mentioned earlier are ready, the next requirements are context term lists, which are learnt from corpora for each words in the source and target word vocabulary lists. Context terms provide occurrence information that is useful to map translation pairs during the extraction process; commonly, the context terms are among highly occurrence words in a context of a (source or target) word. Prior to the extraction process, a context window would be required to define the location where context terms can be collected and their occurrences can be counted; whereby an association measure would be required to define the degree of association (based on co-occurrence) between a word and each of its context terms.

Words occurring in corpora of different languages but sharing equivalent contexts are assumed to be potential translation pairs, however, this assumption would only become useful when their similarity can be measured and ensured. Hence, an initial bilingual lexicon and a similarity measure are required to help obtain translation pairs. Last but not least, an evaluation measure is required to evaluate the outcome.

3.3 Related approaches

Each component mentioned in the previous section is essential to a bilingual lexicon extraction task; therefore, obtaining the right components can be viewed as acquisition problems in the task. Some issues involved with each of the components maybe small but fair attentions are still required. This section presents several approaches pertaining to these components and issues.

3.3.1 Corpora acquisition

Corpora are the main resources required to learn translation pairs. Different approaches for acquiring corpora are extensively described in this section as addition to general methods presented in Chapter 2.

Somers (2001) has noted "fully annotated aligned multilingual parallel corpora are becoming increasingly available through various coordinated international efforts". However, Somers was also concerned about the number of different languages featured, which according to him, "...is still rather small". Insufficient text collections in terms of their amounts or domains coverage would probably threaten any extraction attempt. Hence, acquiring corpora is an issue that requires serious attention.

This section includes extensive descriptions about several approaches that have been considered by previous studies for acquiring parallel or non-parallel corpora. The descriptions presented in this section are not primarily related to bilingual lexicon extraction, because corpora acquisition is a topic related to many studies of different fields and might also stand as a research topic of its own.

3.3.1.1 Existing text corpora

International organizations, such as The Linguistic Data Consortium (LDC) and The European Language Resources Association (ELRA), have built text corpora to provide the corpora internationally through their major efforts. However, most of these organizations do not provide the resources freely. Existing corpora are often encumbered by fees or licensing restrictions (Resnik and Smith, 2003), including the following corpora:

- **The Brown Corpus**
  The Brown Corpus of American English is the first modern, computer

readable corpus, compiled by Francis and Kucera at the Brown University. It contains 1 million words of sample texts from fifteen different categories that were printed in 1961. The corpus is considered undersized but still makes a useful resource for language processing. It is a balanced corpus with different genres, ranging from press reportage, fiction, scientific text, legal text, and to many others.

The Lancaster-Oslo-Bergen (LOB) Corpus and the Kolhapur Corpus are the examples of corpora that were compiled to match the Brown Corpus. The LOB Corpus contains texts in British English sampled from fifteen different categories that were also printed in the year 1961. The corpora comprise different number of texts for each category and the length of each text was about 2000 words. The Kolhapur Corpus also took sample texts from fifteen different categories. The corpus contains 500 texts for each categories, with each text contains about 2000 words. The materials were taken from Indian text printed in 1978, mainly of government texts, press reports, catalogues and fictions.

- **The Canadian Hansards** The Canadian Hansards corpus is a bilingual corpus containing large compilations of parallel Canadian parliamentary proceeding texts in two languages, which were the English and the French. The texts are translations of one another. According to Resnik and Smith (2003), previous studies in machine translation field were focused heavily on the French-English translation because of the Hansards; which were the only large parallel corpora available at that time.

As addition, the Linguistic Data Consortium provides corpora comprise the United Nation proceedings (refer to their website `http://www.ldc.upenn.edu`). Although these corpora are not freely available but the main advantage for acquiring these kind of resources is that the corpora are readily linguistically-marked-up.

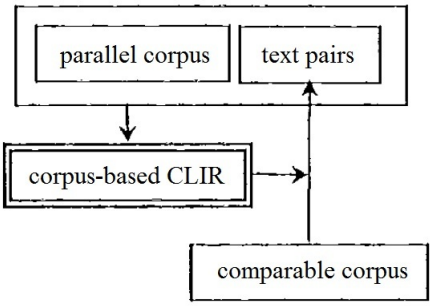

Figure 3.2: A bootstrapping method

Source: Masuichi et al. (2000)

Free sources are slowly becoming more accessible, for example, the Europarl (i.e., the European Parliament Proceedings) corpora comprise parliament proceedings in many European languages. Corpora could also be found in form of specialized documents, such as government documents or software manuals.

3.3.1.2 Deriving 'a subset' of corpora

Learning parallel or comparable texts automatically from collections of texts is an alternative approach to acquiring an existing corpora. Several techniques that based on this approach are described in the followings:

- **Using small parallel corpora to bootstrap bilingual text pairs from comparable corpora**
  Masuichi et al. (2000) proposed a model that extracts comparable text pairs from existing comparable corpora, but this model requires a very small parallel corpus to initiate the extraction process. A method based on this model is called the bootstrapping (see Figure 3.2 that illustrates the method). This method aimed to improve an existing corpus-based cross-lingual information retrieval (CLIR) system by collecting more bilingual text pairs from comparable corpora through bootstrapping.

The bootstrapping method consists of two iterative sub processes: 1. retrieval of bilingual text pairs from a parallel corpus, and 2. concatenation of the text pairs to the initial parallel corpora. The first sub process is used to retrieve more bilingual text pairs based on information gained from initial parallel texts, followed by the concatenation process where small numbers of the most reliable text pairs are to be concatenated to the initial parallel corpus.

For the bootstrapping method to work, Masuichi et al. used a CLIR method based on information mapping approach, which is a variant of the vector space model. In this method, a large word-by-word matrix is used for each languages, where all word vocabularies m of one language appear in the parallel corpus would correspond to the rows, and the most frequent words n of the same language would correspond to the columns of the matrix. Likewise, a similar matrix is built for another language. In other words, parallel corpora are used to determine all word vocabularies for the matrices, which appeared to be equivalences of different languages. Each matrix is used to represent a word space in a single language.

Based on the word space, each text in the comparable corpora is represented by a vector. The text vectors of different languages are then matched to one another using the cosine measure, and the k top most confident text pairs are concatenated to the initial parallel corpus. This process iterates for certain times.

Masuichi et al. conducted two experiments on artificial corpora consisted of English and Japanese bilingual texts; one experiment using the bootstrapping method and another experiment using an existing CLIR method. Better results were observed with the former when compared to the latter.

- **Extracting parallel sentences from two large monolingual corpora**

  Most studies assume parallel sentences to have similarities in terms of sentence length, sentence order and bilingual word context. Rapp (1995); Fung and McKeown (1994); Fung and Yee (1998) suggest that these characteristics should be used to help find correspondences between texts, sentences and words from comparable corpus. They have also suggested the use of documents of similar topics. However, Fung and Cheung (2004b) have argued that: as corpora become less comparable, sentence length and sentence order also become less reliable for finding correspondences.

Fung and Cheung's 2004 study focuses on the similarities of bilingual word context, parallel sentences and document pairs. They proposed a multi-level bootstrapping model that assume as follows: document pairs are useful to find sentence pairs, and sentence pairs are useful to find bilingual word pairs, and the bilingual word pairs improve document and sentence matching process. In this model, Fung and Cheung proposed a principle as follows:

"Find-one-get-more"

The principle emphasizes that all documents, which are found having at least a pair of parallel sentences, are likely to contain more parallel sentences.

Fung and Cheung used TDT3 data in their experiments, which is a quasi-comparable corpus containing various news stories transcription of radio broadcasting and TV news report from 1998 to 2000 in English and Chinese channels. Documents in the comparable corpus are segmented by words. Chinese words were glossed using Language Data Consortium LDC Chinese-English Dictionary 2.0. Each document is represented by

a word vector, where every element in the vector consists of weighted words according to their IDF values.

Munteanu and Marcu (2006) proposed another model to extract parallel sentences from two large monolingual news corpora. However, this method requires parallel texts. The method is divided into three stages as follows: 1. article selection, 2. sentence pair candidate selection, and 3. parallel sentence selection.

In the first stage, the two monolingual corpora are divided into documents on similar topics. Similar documents are matched using a tool (for example, Munteanu and Marcu used a tool kit provided by Ogilvie and Callan (2001)). From each pair, all possible sentence pairs are identified and sent through a simple word-overlap-based filter to find candidate of sentence pairs. Finally, A maximum entropy (ME) classifier could be used to make decisions whether these candidates are mutual translations of one another. A system as described above would require parallel sentences to initiate the extraction task, and small amounts of parallel data to train the ME classifier.

Munteanu and Marcu used this model to utilize parallel sentences extracted from comparable corpora to help improve the quality of machine translation. In addition, they also conducted similar experiments to find parallel sub-sentential fragments from non-parallel corpora. Figure 3.3 shows the architecture of a system based on this model.

Munteanu and Marcu suggest that document pairs to be extracted first, followed by sentence pairs extraction before a parallel corpus could be formed. Their work follows a principle of:

"Find-topic-extract-sentence"

According to this principle, parallel sentences could be found in documents pairs that have high similarity scores.

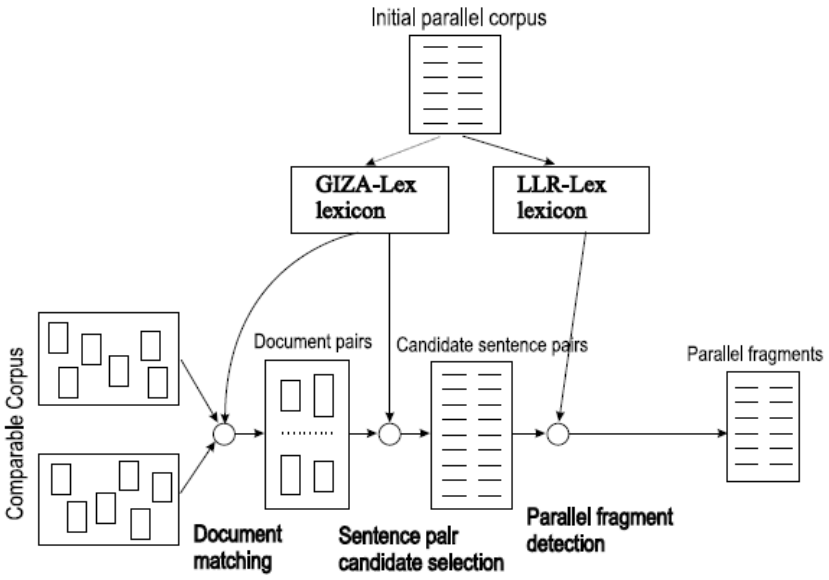

Figure 3.3: An architecture of a parallel fragment extraction system

Source: Munteanu and Marcu (2006)

- **Compiling a specialized comparable corpora using linguistic analysis**

  Goeuriot et al. (2009) built a specialized comparable corpora compilation tool that they used to compile French-English specialized comparable corpora in searching for quality comparable corpora close to one manually-compiled.

  This tool could be used to determine the comparability at three levels: i.e., domain, topic and type of discourse. First, it filters the domain and topic of documents with keywords that are used through web search, before detecting the type of discourse using linguistic analysis. This analysis includes the followings: the structural dimension, the modal dimension and the lexical dimension. Structural dimension is about the structure and the context of creation of documents (see Figure 3.4 for a list of structural dimension features that includes meta tags, number

Feature	French	Japanese
URL pattern	×	
Document's format	×	×
Meta tags	×	×
Title tag	×	×
Pages layout	×	×
Pages background	×	×
Images	×	×
Links	×	×
Paragraphs	×	×
Item lists	×	×
Number of sentences	×	×
Typography	×	×
Document's length	×	×

Figure 3.4: Structural dimension features

Source: Goeuriot et al. (2009)

of sentences, pictures and videos).

The second type of analysis (i.e., the modal dimension) focuses on tone and linguistic elements that are used in texts to define the relationship between an author and a reader. For example, the author 'speaks directly' to the reader in some text, whereas, the tone sounds neutral in a scientific document. Forms of verbs, adverbs and politeness characterize such modalisation.

Meanwhile, the lexical dimension considers variations between texts. For example, a scientific document contains more complex lexical unit and specialized vocabularies (see Figure 3.5 for a list of lexical dimension features that also includes word lengths and punctuations.

Existing work on collection of parallel corpora from the web has not been used in this study. All of the models assume huge knowledge resources are available. The last model also requires a vast linguistic analysis. Moreover, the tool is not language independent.

Feature	French	Japanese
Specialized vocabulary	×	×
Numerals	×	×
Units of measurement	×	×
Words length	×	
Bibliography	×	×
Bibliographic quotes	×	×
Punctuation	×	×
Sentences end		×
Brackets	×	×
Other alphabets (latin, hiragana, katakana)		×
Symbols		×

Figure 3.5: Lexical dimension features

Source: Goeuriot et al. (2009)

3.3.1.3 Harvesting the web

Comparable corpora are more accessible resources compared to parallel corpora. A survey in 2003 predicted that more than 50% of web contents will be in languages other than English. In 2007 the online populations had seen 70% of Internet users were among non-English native speakers. With the increment of non-English users, the number of web pages in languages other than English had also increased.

Multilingual news feed produced by multilingual news agencies such as Reuters, CNN and BBC, multilingual documents from a government web site such as the Hong Kong government web site $Http://www.info.gov.hk$, and contents of online multilingual encyclopaedia such as Wikipedia are among web sites that can be good sources for compiling comparable corpora. Nevertheless, multilingual web pages can also be collected by crawling the web, for example by using a web spider. A web spider is a tool that traverses the web by following hyperlinks to collect web pages. Each collected multilingual set taken from the web is assumed comparable. To improve the set, a task is required to identify parallel pairs among the web pages. Parallel web pages (i.e., bitexts),

consist of translated versions of identical pages.

Many methods, which were based on an approach that acquire corpora by compiling texts from the web, automatically, have been discussed in previous studies. Some methods were quite straightforward, for example, the one that extracts bilingual articles by using input from the RSS system. There were other methods that use certain classification, such as a word-overlap-based filtering, word alignment computation and feature extraction, to extract parallel corpora from news article taken from selected news agencies.

- **Building parallel corpora using structural elements**
  Somers (2001) built parallel corpora through simple 'tricks' like using the names of filenames (.fr, .en) or anchors in the text. Somers proposed a model that finds candidate pairs based on the content of the text, i.e., the amount of text available between each anchor. Among techniques discussed were as follows: identification of anchor points, matching texts between anchors, and the use of machine readable dictionaries and other language specific resources.

- **Structural Translation Recognition, Acquiring Natural Data**
  Resnik and Smith (2003) searched for parallel texts by utilizing the Internet Archive. They used a web mining architecture, which was known as STRAND, to identify pages that were of mutual translations. An algorithm used to realize the goal is as follows: firstly, all pages that might have parallel translations are located, translation pair candidates are generated through URL-matching algorithm, and structural filtering is applied to throw out non-translation candidate pairs. These algorithms were used to build English-Arabic corpora of about 2,000 parallel pages.

  To locate the pages, Resnik and Smith used a search engine's advance search to look for *parent* and *sibling* pages. A parent or anchor page is a page containing hypertext links to different-language versions of a child

document. The hypertext links should be occurring reasonably close to each other. (See Figure 3.6 for an example of parent page). A sibling page is a page in one language containing a link to a version of the same page in another language.

In this method, two child pages of different languages that are linked by the parent page would be paired, and the sibling page would be paired to the page that are linked by the sibling page itself. A URL-matching is used to find pages with parallel organization. A substitution rules are hand crafted for the matching. For example, two pages with the following URLs are likely to be candidate pairs:
$Http://mysite.com/english/home\_en.html$, and
$Http://mysite.com/english/home\_es.html$.
Document lengths are other possible matching criteria.

Structural filtering could be used to analyse the underlying HTML of each pages to determine a set of *pair-specific structural values*. As an example, `<FONT COLOR = ‘‘BLUE’’>` produced a `[START:FONT]` token followed by a `[Chunk:12]` token and ended with an `[END:FONT]` token. These values could be used to determine whether the pages are translations of one another.

- **Assembling parallel corpora from RSS news feed**
  According to Fry (2005), previous web mining approach for discovering parallel texts on the web are useful but also have a few drawbacks as follow:
  - The quantity and, especially, the quality were unpredictable because the researchers have had no control of what have been picked by search engines or web spiders.
  - Most of the approaches were slow. The researchers had to generate sets of hand-crafted substitution rules, and applied the rules to

Figure 3.6: An example of a parent page

Source: Resnik and Smith (2003)

all URL candidates to decide which would become the new URLs before the researchers could check the new URLs for content.

- Some of the methods appeared to be quite complex and required adequate knowledge expertise in many aspects to be implemented.
- Some of the approaches simply considered the full cross product of web pages on each site as possible translation pairs.
- There were cases where some web page pairs were misidentified as translations.

Fry proposed a model that aims to assemble quality parallel corpora in a faster, simpler manner compared to the web crawling. This model takes advantages of trends of news delivery over the web, which includes the following: 1. the trend for many multi-national news organizations to publish articles in multiple languages on the web, for example, the CNET Networks website that offered news in information technology in English, Japanese, German, Korean and French (see the CNET websites at

Table 3.1: List of RSS feeds used to construct parallel English-Japanese corpus

URL of main news site	RSS feed
http://hotwired.goo.ne.jp	http://hotwired.goo.ne.jp/news/index.rdf
http://japan.cnet.com	http://japan.cnet.com/rss
http://japan.internet.com	http://bulknews.net/rss/rdf.cgi?InternetCom
http://www.itmedia.co.jp	http://bulknews.net/rss/rdf.cgi?ITmedia

Source: Fry (2005)

`http://www.cnetnetworks.com`), and 2. the trend of using the RSS (i.e., the XML-based syndication format) that provides news-like content in many sites.

The RSS allows readers to subscribe to RSS feeds of a site rather than checking the sites manually for new content. There were cases in which a translated story in one language was linked to the original story in another language. When the original story was published over the RSS, both stories were made available. Hence, the use of the RSS may allow parallel texts to be collected without requiring a web-crawler.

In order to employ this model, Fry subscribed to a few websites that not only provide the RSS but also publish news stories in one language together with links to the original articles in another language. News feed updates would be received through emails and each incoming RSS feed could be processed as the emails arrived. This process including as follows: 1. Extracting the URL of a new story, 2. downloading the articles using the URL, and 3. finding the link to the original story in the articles. For any links found, both URLs to the new story and the original story are saved to a file and article pairs are extracted.

To test this model, Fry subscribed to RSS feeds of four Japanese websites (see Table 3.1 for the list of the websites). He processed the RSS feeds from the sources over a period of five weeks. Fry found that this basic

RSS-based method collected very small number of parallel news articles over a short test period. Thus, recursive crawler was then used on the past archives of the websites on top of the basic method, which allowed a quick access to a huge list of article pairs up to 20,000 items. The only drawback with this method is that it is only feasible for language pairs with a substantial online news media representation.

- **Parallel Text Identification**
  Chen et al. (2004) built an automated tool to facilitate the construction of parallel corpora by aligning pairs of parallel document from a collection of multilingual documents. The system is called the Parallel Text Identification (PTI). One of the strategies is to take the advantage from a common practice by the web master of a multilingual web site to keep track of the files by languages.

  Similar to other web-crawling based method, the system fetches parallel multilingual documents by crawling the web using a web spider. Two different modules are used in this method to determine the "parallelism" between potential document pairs: 1. filename resemblance checking, and, 2. contents analysis. Parallel documents that are aligned through any modules are then archived to form a parallel corpus. Chen used the system on the Hong Kong government websites, and successfully built an English-Chinese parallel corpus consisting parallel document pairs. However, the websites were actually composed of parallel documents that are direct translations of one another.

- **BootCat**
  Baroni and Bernadini (2004) were the first to propose a tool (i.e., known as the BootCat) for the automated extraction of specialized corpora and technical term using web mining approach. This method is based on a simple idea: a small set of technical terms are used as queries to find similar documents through a search engine and a corpus is built from the returned documents. From the corpus, new (single) word terms are

extracted. Then these processes are iterated by sending queries that contain the newly extracted terms into the search engine. The returned documents create another corpus, and so forth. However, the BootCat simply ignored documents in non-textual formats such as PS, PDF and word documents, though documents in these formats tend to be content-rich such as of scientific papers.

- **Building comparable corpora using local relevance feedback**
  Collier et al. (2003) applied CLIR techniques, including the Local Relevance Feedback, to obtain comparable corpora from database collections of English and Japanese news stories. The researchers interpret the task as *multi-lingual threading of news articles* that finds all related news articles on a particular topic in a different language to an initial query.

  For this task, the goal of the CLIR is reformulated to be a bilingual text matching task. Given a query in Japanese contains a list of Japanese terms, a CLIR-based model finds English documents in the collection of news articles most closely to the query by using the following steps: First, the Japanese query is translated into English by using some linguistic analysis. Secondly, the list of English terms is converted to a query vector. Each English news article is represented by a document vector, and each document vector is matched to the query vector within search date range. A relevance score, which is determined by the number of the matching terms and their distributional characteristics in the document as well as in the whole collection, is calculated. If the system returns a number of highly-matching English documents, the documents are automatically used to weight new sets of terms to refine and expand the query in a local feedback loop, before the search is repeated.

  To test the model, Collier et al. compiled English and Japanese daily news articles, which were produced and posted by Reuters on the web for duration of about five months, i.e., from December 1996 to May 1997.

In total they obtained 6782 English news articles and 1488 Japanese news articles. They found that the local relevance feedback performed well only when the term expansion involved just a few highly scored documents.

- **Querying for comparable documents via search engine**
  Prochasson et al. (2009) compiled English-French-Japanese comparable corpora by exploring the web via search engine. Such method uses some specific themes, such as *diabetes* and *nutrition* of scientific discourse.

  Prochasson et al. used the PubMed to obtain the English documents; the PubMed is accessible from the website `http://www.ncbi.nlm.nih.gov/PubMed/`. As an example, the query they used was:

  "Diabetes Mellitus/diet therapy" [MESH]
  OR
  "Diabetes Mellitus/etiology" [MESH]
  OR
  "Diabetes Mellitus/prevention and control" [MESH])
  AND
  ("nutrition" or "feeding") with limit to "English language"

  Then, all documents returned from the search engine results is manually extracted. To convert into simple, usable text, the documents are sent to a pre-processing stage, where non-informative parts such as *References* are manually removed. Prochasson et al. obtained 257,000 token words for the French corpus, 235,000 token words for the Japanese corpus and 250,000 token words for the English corpus. They considered the corpora as small-sized specialized comparable corpora.

- **Compiling corpora from Wikipedia**
  Laroche and Langlais (2010)'s study aims to retrieve sets of the French and English document pairs from Wikipedia. A method they proposed

involved the use of reference translations to obtain the most relevant target language documents. Sets of bilingual Wikipedia documents could be retrieved using the NLGbase Information Retrieval tool, which is available at NLGbase website (`http://nlgbase.org`).

Using this method, Laroche and Langlais obtained different sets of data ranging from 50,000 to 90,000 token words for the English corpus and 10,000 to 50,000 token words for the French corpus. Laroche and Langlais have noted the advantages of using Wikipedia, which including as follows: most of the Wikipedia document pairs are relevant to one another, some of the Wikipedia document pairs contain a handful of parallel sentences, and the Wikipedia is suitable for mining medical terms.

As a summary, certain characteristics could be used to measure the 'parallelness' of document pairs, including the filenames, the contents and the structures of the HTML texts. For news reports, the comparability of texts collected could be measured by putting some constraints on the texts themselves, for example, the domain, the date and the title of the texts. These characteristics could be used to generate comparable corpora from the web or from existing corpora. Parallel corpora could also be viewed as a subset of comparable corpora, thus, a few researchers also tend to extract parallel texts from comparable corpora.

In addition, text corpora compiled from the web might come in different sizes. The total number of words that occur in both corpora might be about the same, though, more different word types occurred in the English corpus. Many researchers agree that corpora of the same size are not mandatory requirements for bilingual lexicon extraction task to work (Diab and Finch, 2000; Chiao and Zweigenbaum, 2002). All corpora contain a million token words, or less, were considered small in the previous studies. Hence, very extremely small comparable corpora containing less than 1,000 documents might be used to represent an extreme setting, especially in a minimally-supervised approach. Last but

not least, we agree with Diab and Finch (2000) who noted "there are languages that are less represented in electronic forms let alone in translations into another language".

More work related to corpora acquisition can also be found in the proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (BUCC). The workshop was held during the ACLHLT 2011 at Portland, Oregon, USA.

3.3.2 Text pre-processing

Type of text available would define the approaches to be taken for initial text processing (Manning and Schütze, 2002). Texts usually are marked up according to certain structures or formattings. These formattings, which include document headers, footers and diagrams, should be removed from the texts. For example, Laroche and Langlais (2010) used 40 different regular expressions to remove all Wikipedia symbols pertaining to its particular syntax.

Other pre-processing approaches including dividing the input text into sentences with a sentence detector, and dividing the sentences into tokens using a tokenizer (Manning and Schütze, 2002). Tokens are small units with each token consists of a word or as small as a symbol, which would probably provide useful information to an extraction task.

A more specific approach or special case treatment could also be done during the pre-processing stage. For example, if each sentence in the collections is required to be ended by a period, other punctuation marks that end a sentence should be replaced with a full stop.

Extra knowledge resources and tools are necessary to solve some particular pre-processing issues, such as the word segmentation (Shezaf and Rappoport,

2010), the POS tagging (Otero and Campos, 2008), or the lemmatization (Rapp, 1999; Gaussier et al., 2004; Otero and Campos, 2008). POS tagging is used to extract certain type of expressions using basic patterns of POS tags provided in the tool, for instance, V for a verb and NN for a noun phrase. Lemmatization is the process of extracting lemmas of words, and might involves the name entity recognition, in which allows the identification of proper nouns (i.e., mono or multi-word units). However, most systems could still performed well without undergoing the lemmatization process; nonetheless, correct lemmatization would improve results, but noisy lemmatization may reduce the quality of the results (Laroche and Langlais, 2010).

3.3.3 Stop words

Rapp (1999) and Koehn and Knight (2000) suggest that commonly occurring strings that do not help in processing natural language data should be removed, and Fung (1995) suggests removing the function words from the texts to increase the values of many nouns. Similar approach is also found in the IR field.

In an IR setting, dividing text vocabularies into two classes (i.e., the stop words and the content bearing words) is a custom. A stop list, or negative dictionary is a device usually used in an automatic indexing system to filter out words that would make poor index terms. For example in the searching process, English words such as "a," "and," "is," and "the" are left out of the full-text index since they are deemed unlikely to be useful for searching. The advantage of using the stop list is that this technique helps reduce the size of an inverted index up to by half, hence effective indexing can be achieved. However, once the stop list is applied in a system, phrases containing stop words are totally removed and can never be searched again in the system. Hans Peter Luhn, one of the pioneers in IR, should be credited for coining the phrase and for starting the concept in his design. Stop lists have been

constructed for the English languages and most of the major European languages. Developed for English based on frequency statistics of a large corpus (Zipf, 1932) such English stop lists can be easily retrieved online.

Stop words in text collections can be generally divided into two types: 1. generic stop words, and 2. domain stop words. A generic stop list includes words that can be eliminated at any circumstances whereas a domain stop list includes stop words which can only be effective in certain domain.

An English generic stop list typically consists about 200-400 words includes articles, prepositions, conjunctions and some high frequency words. The domain stop list contains repetitive words in domain specific documents. For example, words such as `states`, `system` and `government` appear too frequently as candidates for translation when a bilingual lexicon is learnt from the Europarl, in which consisting parliament proceedings. To make up for high frequency there is a suggestion to reduce the dispersion weight of distributional criteria as follows:

$$w`(t) = \frac{w(t)}{d(t)}$$

where

$w(t)$ is the weight that t had as a candidate for some term,

$d(t)$ is the number of times t has been proposed as a candidate.

A stop word is often associated with low variance and comparatively high frequency in the whole corpus. Conventionally, stop lists are supposed to include the most frequently occurring words. However, in practice, it may also include infrequent words, and not all most frequent words. A classic method by Christopher Fox in 1990, which were manually aided by frequency statistics of the Brown corpus (this corpus contains 1,014,000 words that had been drawn from a broad range of English literature), had generated a stop list containing 421stop words that might differ from other lists available today. This method kicked off with a list of tokens occurring more than 300 times in the Brown

corpus. From this list of 278 words, 32 were culled on the grounds because they were too important and had potential to be index terms. Then, 26 words were added to the list as these words occurred very frequently in certain kinds of literature, and 149 words were added to the list because the finite state machine based filter, in which this list is intended to be used, was able to filter them at almost no cost. The final product was the list of 421 stop words that was used to filter most frequent words occurring in English literatures in the past.

Previous studies for constructing stop lists, automatically or semi-automatically, are available. One of the studies is based on the term frequency, a careful manual elimination process and an assumption that: not every most frequent words in the stop lists should be considered. This study focuses on eliminating terms that carry significant information, although the terms are found to be occurring quite frequently in the corpus. In this study, an experiment on document collections that were restricted to a specific politics domain was conducted to create a stop list. In this experiment, certain words carrying significant information (such as "President" and "France") were found to be highly ranked, thus these word were then eliminated manually from the stop list.

Other methods uses an automated statistical testing based on the IDF to identify stop words in a collection. For these methods, a stop word is seen as a word that has the same likelihood of occurring in documents that are not relevant to a query as in documents that are relevant to the same query. The strength of a term and how strongly the term's occurrences correlate with the subjects of documents in the database are measured. If term occurrences are random then there will be no correlation and the strength will be zero, but, if for any subject the term is either always presents or never present the strength will be one.

Although the IDF provides a useful global weight for terms, the frequency of a term in the database is not the only factor bearing on its usefulness as a key term for document retrieval. Infrequently used terms might also not relate to the specific content of documents. A statistical method of judging the function of a term might be needed.

There is also other automatic model, which is based on a complex statistical model that assigns weights on each term using the Kullback-Leibler divergence measure (TszWai et al., 2005). A stop list constructed based on this term-based random sampling approach requires less computational effort, however, the quality of the stop list is slightly worse than the classical stop lists constructed on term frequency. A merging between this stop list and Fox‘s classical stop list is suggested in this study.

Stop word identifications for other languages than English are also discussed in many previous studies, including by Hao (2008) and Alajmi et al. (2012). For the Chinese language, text tokenizing would be more difficult than in other natural languages because the word boundaries are not well defined. Therefore, a segmentation algorithm has to be employed first before a statistical model can be built for engineering the stop list. Generally, this statistical model is also based on the term frequencies, but the term frequencies are then normalized using document lengths before the probability of each potential stop word become a stop word is calculated.

As a summary, removing stop words is a common practice to reduce index size without affecting the accuracy of an IR system. Likewise, a similar practice is found in bilingual lexicon extraction field. A stop list could be generated automatically. A stop list for generic use is best learnt from a very large corpus. Standard stop lists for English and some other major languages are already available, however, for most languages, the construction either manually or automatically is still required.

3.3.4 Source word and target word vocabulary lists

For some researchers, the task of extracting bilingual word pairs from non-parallel corpora is far too difficult and too ambitious; hence, less ambitious task is adopted. Otero and Campos (2008) noted "... some preferred to work on a less ambitious task, for instance, to choose between several translation alternatives previously selected from a bilingual dictionary". Furthermore, a source word and a target word of translation pairs corresponding one-to-one across the languages might be assumed.

Most frequent words in comparable corpora

The most used approach is to select translation pairs between two high frequency lists containing corpus words of different languages. To obtain a high frequency list, corpus words are sorted and ranked according to the occurrence frequencies, and words with occurrence frequencies of a certain range would be selected to be in the list Gaussier et al. (2004); Fung and Cheung (2004a).

Most previous studies seem to agree that the source word and the target word vocabularies should be among the most frequent corpus words of the source and target languages, respectively (Haghighi et al., 2008; Rapp, 1999). According to Laroche and Langlais (2010), more frequent words have a greater chance of being correctly translated.

Word pairs that co-occurring most frequently in parallel corpora are assumed to have parallel frequencies. Hence, word frequencies could be useful to help extract bilingual word pairs from parallel texts. For comparable corpora, frequent words in one corpus might also have their equivalents occurring frequently in other corpus. The correlations between the occurrence frequencies of word pairs might be rather small, hence, the correlations might not be strong enough to help identify bilingual word pairs from comparable corpora. However, choosing translation equivalents among high frequency words for a

high frequency word might be quite practical.

High frequency words might have more to offer compared to other words in comparable corpora, such as follows:

- `An advantage for statistical methods`
  Higher occurrences might lead to higher level of possibilities.

- `Minimize the computational cost`
  Limiting the number of candidates to only among high frequency words for the matching requires lower costs. Considering all words in a corpus is not a good idea.

- `Minimize error score`
  Limiting the candidates to only among high frequency words should yield less error score.

However, the approach might still have a few disadvantages:

1. *Spurious translation candidates*
   False candidates usually produced because of certain words in a corpus, such as function words, noise words, words with similar context occurrences, and words with similar spelling but not equivalent to the source word. One disadvantage of this approach is that these words highly occurring as high frequency words. A general approach mentioned previously is to eliminate stop words from each corpus during the initial stage. Other approach is to consider only certain word types, for instance, Haghighi et al. (2008) built their bilingual lexicons from the most frequent noun word types, which were the first 2000 nouns that were found in each corpus of the source and the target languages. Similar approach is found in Shezaf and Rappoport (2010)' study.

   In addition, the domain of corpora might also contribute to spurious translations. For example, choosing high frequency words from domain specific corpora (such as the Europarl) might cause certain content words

to appear as significant noises. For example, one of the most frequent words `council` appears to be a 'perfect' but not a correct translation candidate for many Spanish words. Fung and McKeown (1994) avoid such errors by choosing mid-frequency content words to be seed words.

Furthermore, high frequency (test) words of a single language might also have similar context features among them if many words occurring in their contexts also occur highly in the corpus. Commonly, this problem is related to noise words but could become worse when a small-sized, limited coverage seed lexicon is used during the extraction task. This type of seed lexicon might not help much in distinguishing the contexts of words. Nonetheless, using a large-sized seed lexicon does not guarantee that the lexicon would cover across all corpus words.

2. *Translation equivalents were not found among the high frequency words of the other language*
   For comparable corpora, the translation equivalents for the most frequent words in one language may not be the most frequent but may still occur notably in the corpus of the other language (Fung and Cheung, 2004). There is also a possibility of some target words occur very frequently but have been missed because of polysemy.

Interestingly, there is also a study with a far more ambitious goal, which is to find bilingual word pairs among low frequency words. The aim is to fully utilize the corpora words as an addition to the high frequency word approach (Pekar et al., 2006). An experiment was conducted in this study, and as expected, the results were not good.

Last but not least, a study by (Diab and Finch, 2000) has conducted experiments using the source and target word vocabulary lists of a single language, hence, seed and reference lexicons would not be required in the experiments.

3.3.5 Initial bilingual lexicons or seed lexicons

According to Fung (1998), the use of a huge amount of initial bilingual lexicon would probably avoid the prohibitively expensive computational effort. Previous studies assume the availability of a large bilingual dictionary to be used in transforming vectors into a similar word space. General domain bilingual dictionaries are the most used lexicons in many experiments, for example, Rapp (1999) and Fung (1998)used large general bilingual lexicons containing 16,000 to 20,000 entries. Chiao and Zweigenbaum (2002) used a specific domain dictionary, which contains more than 18,000 'simple' medical word pairs (but later, they suggest taking a large general lexicon into account when they experienced less success with the specific bilingual lexicon). On the other hand, Koehn and Knight (2002) were be able to transform vectors of many words just by using small seed lexicons containing less than 1000 entries, which were constructed automatically based on identical word pairs found in comparable corpora. Likewise, a similar approach was introduced by Prochasson et al. (2009), who used transliterated elements and scientific compounds to build the initial English-Japanese and French-Japanese bilingual lexicons.

Laroche and Langlais (2010)'study investigated the impacts of different lexicons' sizes and contents on their systems' performance. They considered a general lexicon of 5,000 bilingual entries, and mixed lexicons of 5,000, 7,000, 9,000 and 11,000 entries (with 2,000 of the entries were related to medical domain). They observed that more accurate translations were produced when mixed lexicons were used compared to the general lexicon, but only by a small margin. Hence, they conclude that lexicons of smaller size (of about 7,000 to 9,000entries) and containing mostly content words would be appropriate to be used if the unavailability of general bilingual lexicons is assumed. According to Laroche and Langlais, seed lexicon per se might not have great impact on the systems's performance, however, they have not investigate bilingual lexicon with limited size to compare the effects.

Koehn and Knight (2002) discusses a technique using cognate pairs, which were extracted from comparable corpora, to be the seed words in order to alleviate the needs of large initial bilingual lexicons. Haghighi et al (2008) followed this idea by constructing a very small seed lexicon that contained identical word pairs from comparable corpora. Nonetheless, Koehn and Knight (2002) pointed out that majority of word pairs would not show much resemblance if unrelated language pairs (such as, the German-English languages) are used.

3.3.6 Context windows

In the context-based approach, context windows are used to collect co-occurrence frequencies. The sizes of the context windows may vary, ranging from as small as the 2-word window (Rapp, 1999; Diab and Finch, 2000; Koehn and Knight, 2002; Gaussier et al., 2004), over the 3 word-window (Chiao and Zweigenbaum, 2002), to as large as the 25-word window (Prochasson et al., 2009) and the window size of a sentence. Fung (1995) suggests larger context window size is used to improve the lexicon, while Gaussier et al. (2004) insist that the small 2-word window is used because it would help alleviate noise in a multi-dimensional word space.

Laroche and Langlais (2010)'s study investigated the effects of different context window size by taking the 5-word window, the 25-word window, the window size of a sentence and the window size of a paragraph. They observed that their context-based systems performed better when the window size was based on sentences.

3.3.7 Association measures

In the standard approach, each word is associated with a set of context terms according to how likely the terms co-occurring with the word in similar context. Fung and Yee (1998) have noted that not only the number of common

words in context would be the clue to word similarity but also the actual ranking of the context words. Hence, weighted context words are more preferable compared to the actual co-occurrence counts. Rapp (1995) suggested that the correlation between word co-occurrences in different language texts could be strengthened by using association measures.

No specific best measure to compute the weight or the degree of association has been reported. Fung (1998) suggests the IDF instead of using the term frequency. The term frequency considers all words that highly occurred in the context of a test word, including general usage words that might cause the performance of a system to weaken. In contrast, the IDF de-emphasizes these general usage words by taking the context globally. However, this IDF measure is often used to derive stop words rather than finding highly associated words.

Quite often the LLR is used to find highly associated context words (Melamed, 1997; Rapp, 1999; Prochasson et al., 2009; Laroche and Langlais, 2010). This LLR measure determines how likely two words will co-occurring together. Nonetheless, Chiao and Zweigenbaum (2002) mentioned about their attempt has failed when the LLR was used in their experiments, but unfortunately, they did not publish their experimental findings in detail.

Other popular association measure is the PMI (Shezaf and Rappoport, 2010; Andrade et al., 2010). This PMI uses relative frequencies that considers how much occurrence of a context word makes the occurrence of a test word more likely (Andrade et al., 2010). This measure assigns high weights to frequent events (Manning and Schütze, 2002). Andrade et al. (2010) proposed a smoothed PMI to avoid the bias. However, they observed that their version of PMI performed only slightly better than the original LLR, especially, when only positively associated pivot words were considered during their experiments.

3.3.8 Similarity measures

A similarity measure is a critical component of a bilingual lexicon extraction. This measure is used to find matching pairs.

Different similarity measures are used in the previous studies. The most used similarity measure is the cosine (Fung, 1998; Chiao and Zweigenbaum, 2002; Gaussier et al., 2004; Prochasson et al., 2009). Chiao and Zweigenbaum (2002) conducted experiments to compare the similarity measures, including the Cosine, Dice and Jaccard measures. Their systems performed slightly well when Jaccard measure is used compared to the Cosine measure. In addition, Gaussier et al. (2004) employed Fisher Kernels to compute the similarity between complex multi-dimension word vectors and compared the results to another system with similar settings using the cosine, and they found that the cosine works better compared to the Fisher Kernels.

Other similarity measures that have been discussed in the previous studies include the simple city-block measure (Rapp, 1999), non-aligned signatures similarity scoring (NAS) (Shezaf and Rappoport, 2010), and other new versions of existing measures (Fung, 1998; Andrade et al., 2010). Nonetheless, Chiao and Zweigenbaum (2002)mentioned another failed attempt when they used the city-block measure, but decided not to publish the details of the findings because the results were too poor to be presented. However, Chiao and Zweigenbaum believe that having larger size of corpora and more general type of lexicon, and considering word order were among factors that help improved the results in Rapp's experiments.

3.3.9 Evaluation methods

Similar to other NLP systems, an evaluation measure is used to evaluate bilingual lexicon extraction models. Current evaluation methods may vary but mostly are influenced by the notions of evaluation introduced for IR, i.e., the

precision and the recall (Gaussier et al., 2004; Haghighi et al., 2008; Shezaf and Rappoport, 2010). Based on the precision and the recall, the average precision-recall score could be computed using a measure called the F_1 score. Other simple measure that is available is the accuracy score (Fung, 1995; Melamed, 1997; Rapp, 1999; Koehn and Knight, 2002).

3.3.9.1 Accuracy

Accuracy score is a very simple, straight forward measure. The score indicates the percentage of items right for the system to select, or not to select. The measure is defined as follows:

$$accuracy = tp + tn$$

where

tp is the number of cases the system succeed by selecting the target (correct) item, which is also called the *true positive*, and

tn is the number of cases the system succeed by not taking the wrong item, which is also called the *true negative*.

However, the accuracy score might not be a good choice because the measure emphasizes on less important cases indicated by the huge tn value. It is more interesting to consider other smaller cases, for example, the number of cases the system succeed to ignore the wrong items.

3.3.9.2 Precision and Recall

In the IR field, the precision is defined as "a measure of the proportion of selected items that the system got right', and the recall is defined as "the proportion of the target items that the system selected" (Manning and Schütze, 2002)'. The precision p can be measured as follows:

$$p = \frac{tp}{tp + fp}$$

whereas the recall r can be measured as follows:

$$r = \frac{tp}{tp + fn}$$

	Actual	
System	target	⌐ target
selected	*tp*	*fp*
⌐ selected	*fn*	*tn*

Figure 3.7: The contingency matrix represents the concepts used in the precision and the recall

Source: Manning and Schütze (2002)

where

fn is the number of cases the system failed to take the target item into account, which is also called the *false negative*.

Likewise, a similar definition of the precision in the bilingual lexicon extraction is provided, which is defined by the fraction of selected items, where the system has chosen the right ones. However, the number of *fn* cases, where the system has failed to consider the right target items, is typically ignored in the previous equation of the recall). Therefore, this study would follow Haghighi et al. (2008) who define the recall as the fraction of possible translation pairs proposed (i.e., comprise the items that the system succeed to select, regardless whether the items are right or wrong targets). Figure 3.7 shows the contingency matrix to help illustrate the concepts in a simpler way.

3.3.9.3 F_1 score

The F_1 score, or the F measure, is a single measure that combines the precision and the recall to obtain the overall performance score. This measure addresses a well-known *trade off* issue between the precision and the recall (i.e., if every items are selected in order to obtain 100% recall, low score is expected from the precision). This F_1 score is defined as follows:

$$F_1 = \frac{1}{\alpha\frac{1}{p} + (1-\alpha)\frac{1}{r}}$$

where

α is a factor that determines the weighting of precision and recall.

For equal weighting of p and r, α is set to 0.5.

Equal weighting is commonly used. With the α value = 0.5, this F_1 measure is simplified to become as follows:

$$F_1 = \frac{2pr}{(p+r)}$$

The F_1 score is quite similar to accuracy as they both bias to certain cases encountered by the system. In the IR, the F_1 measure is sensitive to the numbers of correct cases, whereas the accuracy is sensitive only to the numbers of errors. In bilingual lexicon extraction, the F_1 score is more sensitive with every items selected by the system regardless whether the item is correct or wrong.

This sub section has described methods that mostly used in many previous studies during the evaluation stage. As addition, (Gaussier et al., 2004)'s have also insisted that the rank of the candidates should be treated as an important feature at this stage.

At this point of this study, comprehensive information has been obtained and presented in this thesis. The next section describes initial experiments conducted in this study.

3.4 Base experiments

The objectives of conducting initial experiments in this study is a two-fold: (a) to test the standard context-based approach models, and (b) to determine a system that could be the baseline for this study.

These tests are divided into three different sets based on the following: the corpora, the similarity measure and the initial bilingual lexicons. In addition, another set of tests is held to compare the use of high-dimensional data to the lower dimensional data. The latter involves the canonical correlation analysis (CCA) to be incorporated into a context-based system in order to generate data in the lower dimensions.

3.4.1 Basic methodology

A basic context-based method, which is used in one of the experiments in this study, is briefly described as follows:

1. Context vector models $\vec{x}$ are built for all selected source words s in the source language.
2. Context vector models $\vec{y}$ are built for all translation candidates t in the target language.
3. For all pair of a source word and a target word (s, t), the similarity score $sim(\vec{x}, \vec{y})$ is computed between the source word vector, $\vec{x}$, and the target word vector, $\vec{y}$.
4. The output is ranked according to the similarity score.
5. The N highest ranking t is chosen as the translation candidate for s.

Pre-processing:
Each corpus is sent through similar pre-processing jobs. All formatting and tags are removed completely from the corpus using a set of regular expressions. Numbers, special characters such as #, or currencies symbol such as $ are also removed. Each corpus is decomposed into sentences using sentence boundary detection. From each sentence, words are extracted using a tokenizer and collected to form corpus words.

Prior to the tokenizing process, a series of special case processing is used to ensure a standardized form of texts could be obtained, including as follows:

all sentences should end with a full stop; symbols like a question mark (?) and an exclamation mark (!) are 'normalized', also into a full stop (.); and language specific symbols (such as − that typically used in Malay language for plural nouns, for example, the Malay translations for English words `player` and `players` are *pemain* and *pemain-pemain*, respectively) are removed. The latter means that plural and singular nouns would be treated differently in this study. No other major pre-processing is required.

Stop lists for a few major languages can be downloaded from the web (for example, website `http://www.lextek.com/manuals/onix/stopwords1.html` and `http://www.thebananatree.org/stoplist.html` provide the English general stop lists). The English stop words are also used to help construct stop list for other languages manually, if its translation equivalent is available in the other language). More stop words are obtained using the IDF scores of each corpus words, which indicate words occurring in most documents as those that should not be favoured. This approach was introduced in the bilingual lexicon extraction field by Fung (1998).

3.4.2 Experimental setups

This sub section describes the experimental settings that were used during the evaluation stage in this study.

- *Data*

  For the first set of evaluations (i.e., based on different corpora), several corpora were obtained from different sources to form different sets of corpora as follows: (a) the British National Corpus that contains a large collection of English texts and a subset of the ECI/MCI multilingual corpus that contains Malay texts form the Malay-English comparable corpora; (b) Wikipedia documents were downloaded, from which comparable document pairs were compiled to form a small comparable corpora; and (c) a subset of the Europarl corpora is derived to form comparable data.

For other evaluation sets, only the third corpora in the list were used. All corpora used in this study are further described in the followings:

The ECI/MCI multilingual corpus

A multilingual corpus could be obtained from The European Corpus Initiative Multilingual Corpus I (ECI/MCI) website, which is initiated by The European Network of Excellence in Human Language Technologies (ELSNET). This organisation is based in the Netherlands. The ECI/MCI corpus is a not balanced corpus because the corpus contains many genres, ranging from government documents and lecture notes, to fiction novels and instruction manuals. From the ECI/MCI corpus, Malay monolingual texts containing 546,653 token words are retrieved; where 26,670 of the token words are unique words. In this study, we refer to the collection as the MalayMCI corpus.

The British National corpus

The British National Corpus (BNC) is described as a balanced text corpus, which containing more than 100 million words, with morpho-syntactic annotations. Figure 3.8 shows excerpts from BNC texts, in which, the top represents the original annotated texts, and the below shows the texts after the formattings were removed.

The Wikipedia

The Wikipedia is one of the free sources that provides large collections of texts in many languages. We have mentioned in the precious section about Laroche and Langlais (2010) who retrieved Wikipedia articles to compile comparable text. Related work that identifies word translations using Wikipedia can also be found in Rapp et al. (2012)'study.

For this study, 1000 Malay-English article pairs were collected from the Wikipedia, semi-automatically. A list of titles of the Wikipedia articles in both languages is obtained from the web. We assume articles in different languages having an identical title are article pairs, for example,

all the Wikipedia articles across languages with the title of "Margaret Thatcher" are all about the same person that represents as the former prime minister of the United Kingdom. Only Malay and English articles with identical titles and contains more than 20 sentences each were considered. We refer to this small Wikipedia collections, which containing about 200,000 Malay word tokens and more than 350,000 English word tokens, as the MyWiki.

The Europarl parallel corpora

The Europarl parallel corpora are initiatives from Phillippe Koehn (Koehn, 2005) who provides these corpora freely. From these Europarl parallel corpora, a few subsets of English and Spanish texts were derived using the following procedure: (a) Both corpora were divided into three different parts comprised texts published according to certain range of year (i.e., year 1996 to 1999, year 2000 to 2003 and year 2004 to 2006); (b) for one of languages, only the first part of the corpora in that language was considered (for example, 40,000 sentences were taken from the first part of the Spanish corpus (i.e., published in (year 1996 to 1999)), and (c) for the other language, only the second part of the corpora in that language was considered (for example, another 40,000 sentences were derived from the English corpus published in (year 2000 to 2003). We refers to these text collections as the MyEuroparl. This procedure is originated from Koehn and Knight (2002)'study. Similar procedures can be found in Fung and Cheung (2004a) and Haghighi et al. (2008).

- *Source and target word lists*

  To obtain the source and target word vocabulary lists, all corpus words of the source and target languages were sorted and ranked according to their occurrence counts. Noise words were removed from the two vocabulary lists based on the stop lists of the two languages, respectively. Every word with single occurrence was removed from the lists. From the remaining words left in the lists, a number of words that occurred

```
</txtClass>  </profDesc>  <revDesc>    <change n=1>      <date
value=1994-11-24>        1994-11-24      </date>      <respStmt>
<resp>          Initial accession to corpus        </resp>
<name>          dominic        </name>      </respStmt>
</change>  </revDesc></header><text complete=Y org=SEQ
decls='CN000 HN001 QN000 SN000'><div1 complete=Y n=1 org=SEQ
type=item><head type=MAIN><s n=001> <w NP0>HIV <w PNP>IT<w
VBZ>'S <w DPS>YOUR <w NN1>CHOICE</head><pb n=1> <gap
desc="Newspaper cutting omitted" ed=OUP> <pb n=2><p><s n=002> <w
AT0>Every <w NN1>day <w AT0>the <w NN1>virus <w AJ0-VVG>causing
<w NN1>AIDS <w VBZ>is <w VVG>infecting <w AV0>more <w AJ0>young
<w NN0>people<c PUN>.<s n=003> <w AT0>A <w NN1>friend <w VM0>can
```

```
every day the virus causing aids is infecting more young people.
a friend can infect you without your knowing.  complete y no
cure for aids once you re infected the virus may destroy your
natural defences for over years without you realising.  and you
an pass on the infection again without knowing.  there is no
cure.  once infected you will sooner or later develop fullblown
aids.  friends or partners may soon be ill too.  complete y the
truth about aids the doctor says hiv is the virus that causes
```

Figure 3.8: Excerpts of BNC texts before and after the pre-processing stage

frequently were selected to form the vocabulary lists of the source and target words.

For the first evaluation, only the first 50 words in the vocabulary lists were selected, whilst, for the second and the third evaluation sets, the first 2000 words were selected because larger corpora were involved in the latter.

- *Association and similarity measure*
  The PMI was used to calculate the degree of association between a word and its context words in each evaluation set. In addition, the cosine is employed to measure the similarity between matching word pairs.

- *Initial bilingual lexicons*
  Seed lexicons are used to build word vectors before the vectors can be mapped in order to find matching pairs. For these experiments, different sets of of seed lexicons were used, including as follows:

- Lex_{700} - a bilingual lexicon containing 700 cognate pairs, which the entries were manually compiled from a few Learning Spanish Cognate websites such as follows:
  - `http://www.colorincolorado.org`, and
  - `http://www.language-learning-advisor.com`.
- Lex_{100} - a bilingual lexicon contains as small as 100 bilingual entries, which was constructed semi-automatically from the most frequent words in the source corpus that share similar spelling with one of the target words, i.e., the first 100 source words that have their translation equivalents found among the first 2000 target words.
- Lex_{160} - a particular bilingual lexicon containing word pairs of different languages that share similar spelling. Only 160 word pairs, which have the edit distance value of less than 2 and their lengths were longer than 4 characters, were considered. This approach is not appropriate for the unrelated Malay-English comparable corpora that we obtained previously.

A reference lexicon is another important component of evaluations, which provides known translation pairs. In this study, the English-Spanish reference lexicon was extracted from `http://www.wordreference.com`, which is a free online dictionary. This extracted bilingual lexicon has a low coverage. For Malay-English language pair, about 5000 bilingual entries were compiled from an online dictionary provided by Dewan Bahasa dan Pustaka (a large government organization focusing on Malay language development in Malaysia), which can be accessed at `http://dbp.gov.my`. In addition, only candidate pairs that were found in the reference lexicon were considered during the evaluation, whilst, the rests of them were treated as unknown words and removed from the candidate pair list, regardless whether the translation pairs were correct or not.

- *Stop List*

  We obtained 598 English stop words, 350 Spanish stop words and 297 Malay stop words.

- *Evaluation*

  The precision, the recall and the F_1 scores were used to evaluate outputs of the systems based on the a reference lexicon. The precision score p were given at certain recall values r, which is denoted by p_r. For example, $p_{0.1}$ is the precision score p at the recall value 0.1, which also means the percentage of being correct at the 10% recall.

  The outputs comprise a list of candidate pairs of translations. Only the first 2000 of the (s,t) candidate pairs, which s only have the highest ranked t (or top 1) were considered. This study has ignored the word types, however, after all stop words were removed from the vocabulary lists words left in the lists were mostly of content words.

3.4.3 Evaluation results

This sub section presents the results of different systems in the experiments.

Performance of systems using the cosine

These experiments involved two different types of data values, i.e., $CB + CosReal$ that denotes the system using cosine on real values, and $CB+CosBit$ that denotes the system using cosine on binary value. $Lex_7$00 and MyEuroparl were also used in these experiments.

Table 3.2 shows the results of the experiments. The systems based on the $CB + CosBit$ and the $CB + CosBit$ models recorded 52.6%, and 43% of the F_1 scores, respectively. In details, the $CB + CosReal$ scored higher precision values at lower recall with almost 90% of precision score at 10% of recall value, and followed by 73% of precision score at 25% of recall value. On the other hand, the $CB + CosBit$ yielded higher precision score at higher recall, which

are 64.8% and 55.2% as shown at 33% and 50% of the recall values, respectively. The $CB + CosBit$ has slightly outperformed the $CB + CosReal$ by less than 10% of the best F_1 score, which means the former proposed more correct translation pairs, although more correct candidate pairs were observed for the first 20% of candidate pairs generated by the latter.

Setting	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$BestF_1score$
CB+CosBit	58.3	61.2	64.8	55.2	52.6
CB+CosReal	89.7	73.5	63.5	47.3.2	43.0

Table 3.2: Performance of systems with the cosine of different values

Performance of systems using different corpora

This evaluation considered the Lex_{100} as the bilingual lexicon to help build the basis term vector in this evaluation set. Each of the models involved in this evaluation are denoted according to the particular corpora each system learned from. The $CB + MalayMCI$, the $CB + MyWiki$ and the $CB + MyEuroparl$learned from The MalayMCI, the MyWiki, and the MyEuroparl, respectively.

Setting	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$BestF_1score$
CB+MalayMCI	0.0	10.0	12.0	8.2	7.6
CB+Wikipedia	5.0	5.1	6.0	4.7	5.6
CB+MyEuroparl	52.0	53.0	47.2	44.8	66.4

Table 3.3: Performance of systems with different corpora

Table 3.3 shows the results of the second set of the evaluations. The $CB + MyEuroparl$ that used larger comparable corpora and related language pairs scored an outstanding F_1 score of more than 65% compared to the other two systems, which were both underperformed.

Performance of systems using different bilingual lexicon sizes

Each experiment in this evaluation set used initial bilingual lexicons of different sizes and the cosine to learn bilingual word pairs from the MyEuroparl. Each model is denoted by $CB + 700$, $CB + 160$ and $CB + 100$.

Setting	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$BestF_1score$
CB+700	58.3	61.2	64.8	55.2	52.6
CB+160	68.5	56.8	48.8	48.8	62.4
CB+100	52.0	53.0	47.2	44.8	66.4

Table 3.4: Performance of systems with different sizes of bilingual lexicon

Surprisingly, when the system used the smallest bilingual lexicon (i.e., $CB + 100$), the system outperformed the other systems by over 65% of performance score. The $CB + 160$ scored the second best with the F_1 score 62.4%, followed by $CB + 700$ of 52.6%. See Table 3.4 for the results of systems using different initial bilingual lexicon sizes. Overall, the system works quite well in these settings.

3.4.4 Discussion

This section discusses some of the experimental findings to justify the best components to be used within the baseline system's settings for this study.

3.4.4.1 Using the cosine for measuring the similarity

The experimental findings show that either real value and binary value datasets could be used with the cosine. The metric is directly proportional to the actual source and target word values; thus, the similarity value is easy to be affected by the values. Although the source word and the target word might not be closely correlated, the matching score could still be high if these words have many high context term values. It would be ideal if all of the high context terms are the important context terms, but this situation is not likely with

sparse data.

Changing the real values simply into binary values could avoid the bias, hence, the slight improvement was observed in the experiments. Nonetheless, the cosine measure favours word pairs that share the most number of non-zero context terms values. A flat representation of context data with similar weights would only causes some important context terms to be missed. However, the results might have also shown that the use of the cosine on the binary dataset is appropriate because of the real characteristics of the corpora. Unfortunately, the 'relatedness' and the 'comparableness' of corpora are not easy to be measured. Experimental findings might be different when other types of comparable corpora are to be used. (See Chapter 7 for further discussion).

3.4.4.2 Minimum requirements of corpora

Through the second evaluation set, the results achieved by the system were not good when the unrelated and less comparable corpora were used. On the other hand, the results were better when the MyEuroparl was used, which is probably due to having highly comparable texts.
In addition, the performance of the system that used the MyWiki (i.e., the corpora that also contains related texts although the texts are of two unrelated language pairs) is not as expected. The problem might probably caused by the seed lexicon, which is very small in size and has very low coverage. This issue has been introduced in the previous sections (see Sub Section 3.4.4.2).

With the current technologies, acquiring comparable corpora becoming an easier task. Compiling article pairs from the Wikipedia is an interesting approach, however, the compilations might have limited size, and most words in the compilations might only found to be occurred in a single documents, or two, which causing many words to have insignificant occurrence counts and easily been missed. As suggestions, larger data should be acquired and some careful measures are required to improve the extraction process. Other idea is

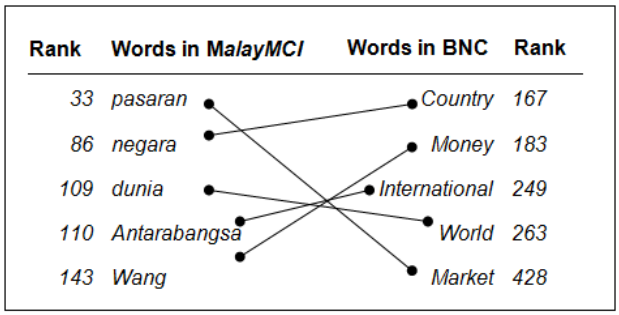

Figure 3.9: Sample of translation equivalents found in MalayMCI-BNC, and sorted according to their ranks. Each line shows the matching translation pairs.

to use the MyWiki as a starting point to search for more data available from the internet. (See Chapter 6 for details).

Very low score near zero achieved with the MalayMCI and the BNC was to be expected because these two corpora contain extremely unrelated texts. Each corpus has different genre composition; i.e., the MalayMCI contains mostly collections of original Malay texts and translations from English of technical books and a few novels; whereas the BNC contains general texts and also text examples of both spoken and written languages. We analysed the vocabulary lists extracted from both corpora and observed that there was less than 5% of Malay unique words would find their equivalents in the English word list, i.e., too many potential target words were missing. In addition, if that target words are available, the possibility for the target words to be similarly ranked with the source words is low because these source and target words have extremely different occurrence counts. Figure 3.9 shows some examples of translation equivalents sorted according to their ranks. These translation equivalents were extracted manually from the English-Malay vocabulary lists used in this study.

Larger size and wider coverage corpora would likely help a system to produce better results. In this study, the *MyEuroparl* of 50,000 sentences is considered small and now become the minimum requirements to ensure the feasibility of

the extraction process.

3.4.4.3 Choosing an initial bilingual lexicon

The initial seed lexicon plays a major role in extracting bilingual lexicon from comparable corpora. A few different approaches that could be used to derive seed lexicons have been described in the previous sections. The Lex_{700}, the Lex_{160} and the Lex_{100} were derived using different methods.

The F_1 scores of the system using Lex_{100} was much higher compared to the system using Lex_{700} especially when the cosine was used to measure the similarity between matching pairs. Thus, adding more word pairs of high frequency words to the Lex_{100} might improve the results because the words frequently occurred in the corpora. Although the size of Lex_{700} is the largest in the experiments but the lexicon contains many general words; hence, it is not surprising to find most of the words were never occurred in the corpora, such as the English word *volleyball* and word *romantic*.

The results with the Lex_{160} is more interesting because the lexicon was derived automatically from the corpora. However, the relationships between the language pairs used in the experiments might have largely affected the results; i.e., the reason for the approach to be unsuccessful when we were working on the unrelated Malay-English translations. Nonetheless, the texts in the MalayMCI corpus are considered as an old collection. Nowadays, many loanwords have been incorporated into the Malay language and have been widely used in modern texts; for example, a Malay loanword *bajet* (budget) has been used more often than the original translation equivalents (i.e., *belanjawan*) since the loanword was introduced in the past three years.

Hence, larger initial bilingual lexicon does not guarantee better results; smaller bilingual lexicon might still helps generate better results as long as the con-

tents of the lexicon is wide coverage, especially of the corpus words.

(We might have concluded with the following: having a small, good quality initial lexicon is better than having large, less quality bilingual lexicon. However, the experiments conducted were too small for us to make such a conclusion. The Lex_{700} may represent a small, general bilingual lexicon with the entries randomly occurred in the corpora of unrelated language pairs, and on the other hand, a small hand-compiled lexicon can be considered as an initial outcome of a minimally-supervised approach. The latter approach is more preferable than the former because the approach is automatic and dynamic in terms of the flexibility of language pairs and corpora used. If any initial bilingual lexicon lexicon depending too much on the availability of the lexicon, or the corpora per se, a new bilingual lexicon would be required whenever different corpora are involved. Hence, the Lex_{700} can be considered as a language-independent and corpora-independent lexicon.

3.4.4.4 The advantage of stop word removals

Previously, Fung (1995) suggests a filtration of all commonly occurring words that do not help in processing natural language data (i.e., the stop words). A few approaches have been described in the previous section. This idea of removing stop words sometimes seems as a negative approach to the natural articles of language. We analysed the corpus words to observe the occurrence of the stop words and their effects to the system's performance.

Figure 3.10 shows a rank-frequency curve of word frequency in the MalayMCI. The plot is in log coordinates where x-axis indicates the rank of a word in the frequency table and y-axis indicates the total number of the word occurrences. It is clearly shown that the curve corresponds to the Zipf's law, with a few words occurred very often and too many words occurred only one time in the corpus.

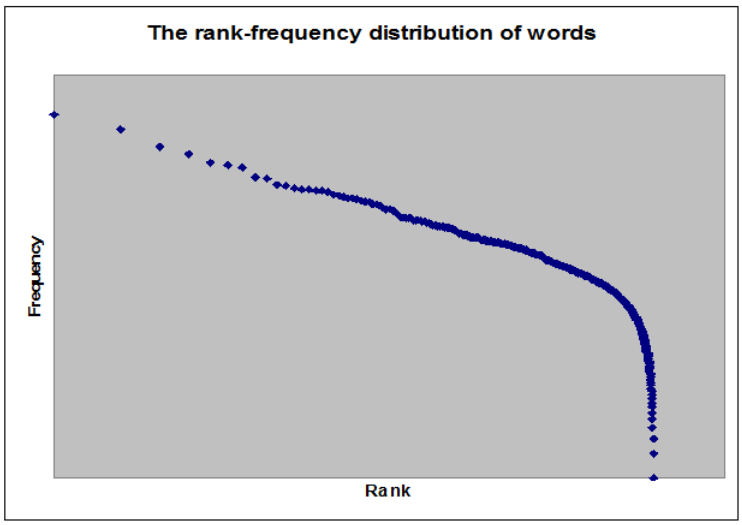

Figure 3.10: The rank-frequency distribution of words in the MalayMCI corpus

Then, we analysed the groupings of words in the corpus according to their occurrence frequencies. We sorted and ranked the words in a descending order, from the highest to the lowest frequencies. We observed the followings:

- The first 25 words, mostly, consisted of prepositions and pronouns such as the Malay words *ini* (`this`) and *dia* (`he` or `she`). Surprisingly, almost 7% of the corpus words consisted only of these two Malay words: *yang* (`that` or `which`) and *dan* (`and`).
- The first 100 words comprised prepositions, pronouns and some nouns. For example, the Malay word *bank* (`bank`) was ranked at 26th, while the Malay word *wang* (`money`) is ranked at 52nd. Other words include *bandar* (`city`), *sistem* (`system`) and *nilai* (`value`).
- The rest of words, which occurred more than 10 times in the corpus comprised many verbs and common nouns. We grouped these words as the medium frequency words.
- More than 40% of the unique words are considered as *hapax legomena*, in which, the words occurring only once through the whole corpus. This

The first 2 words	**36939**
yang	22236
dan	14703
%	6.76
The first 6 words	**65231**
yang	22236
dan	14703
di	9372
dalam	7462
itu	5854
dengan	5604
%	11.93
The first 50 words	**154549**
%	28.27
Hapax legomena	**11168**
%	2.04

The total of token words = 546653.

Table 3.5: A brief analysis of word groupings of the MalayMCI corpus

group also represents many typos, for example, `Bahnk` (supposedly, `bank`) and ***anaknyaitu***, which we suppose it to be ***anaknya itu*** (`her child` or `his child`).

Table 3.5 shows the details of the analysis.

In this study, we have applied the stop word removal approach to all experiments. A system based on the standard approach (i.e., the $CB + 700$) is one of the experiments involved with the stop word removal process, which recorded 52.6% of the best F_1 score. To observe the effects of the stop words, another model (i.e., X) was implemented without using the stop word removal approach during its pre-processing stage. As expected, the performance of the system X is very poor. From our observation, the problem was mainly caused by the internal approach that took the first n-word from the ranked word list;

without stop word removal, the list contains too many noise words, hence, these noise words were chosen to be among candidates for further process. As a result, no correct translations was proposed by the system.

We conclude that stop word removal is very useful in order to obtain an improved set of test words, i.e. the source words and the target words vocabulary lists. These stop words do not have any effect to the representation of words as long as these stop words are not among the initial bilingual entries. Nonetheless, most stop words are high occurrence in corpora, hence, a more effective system would be achieved if these stop words are removed, completely.

Source word (in English)	**Target Word (in Spanish)**
beauty	natural
diplomacy	crisis
airspace	trafico
digital	television
tourism	economico
banks	central
system	deficiencias
opinions	expertos

Table 3.6: Interesting incorrect pairs

3.4.5 Common errors from a basic context-based model

A post-experiment error analysis was also conducted in this study. Too many common errors arose from semantically related words, which had strong context feature correlations. Table 3.6 shows some examples of interesting incorrect pairs we found in the English-Spanish translation pairs. Figure 3.11 shows five context words found for the first two source words, together with their incorrect and correct target words. There are also types of errors that were difficult to categorize but occurred quite often, such as the English word (e.g.

adequacy was proposed as the translation to the Spanish word *sociedades* (societies)).

Source word	Target word	
	Incorrect	Correct
beauty {areas, necessity, protection, Europe, spots, ...}	*natural* {*Europa*, (Europe) *necesidad*, (necessity) *estado*, (state) *bella*, (beauty) belleza, (beauty) .....}	*bella* (adj) {*contribuir*, (to contribute) *Europa*, (Europe) *naturaleza*, (nature) *medio*, (average) necessidad, (necessity) .....} / *belleza* (noun) {*directiva*, (directive) *natural* (natural) *reacciones*, (responds) *contribuir*, (to contribute) *piratas*, (pirates), ...}
diplomacy {development strategies activities nations joints ....}	*crisis* {desarrollo (development) pablo (nation) bienestar (welfare) resultantes (resulting) sector (sector) .....}	*diplomacia* {crisis (crisis) desarrollo (development) estados (states) preventiva (preventive) relacionar (to connect) .....}

Figure 3.11: Sample contexts of incorrect and correct translation pairs

3.5 High dimensional data vs. low dimensional data

As addition to the models based on the standard approach, we also employed the CCA-based models in order to compare the effects of different dimensions of data; i.e., high dimensional data and lower dimensional data. The CCA are usually used to resolve the problems of high-order co-occurrences.

Terms do not need to be in the context of a word to be associated. Associations could also occur in a latent space. Generally, a CCA-based model involves two distinct views or datasets; where each of them represents data of different languages. In this study, the CCA-based models employed in the experiments comprised two different sets as follows: (a) a basic CCA-based model, and (b) an extended model that combines both standard and the CCA approaches.

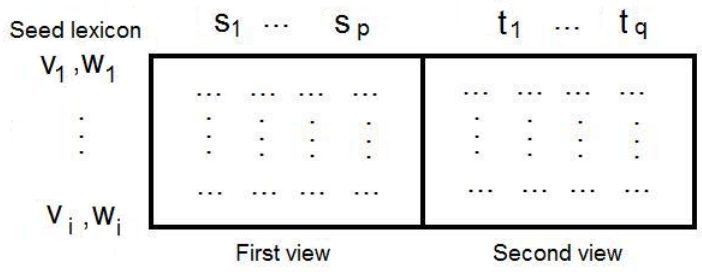

Figure 3.12: An illustration of two distinct views

Let S denotes the set of source words and T denotes the set of target words. The CCA-based algorithm requires two separate data; one consists of all context vectors of S in the source language and the other consists of all context vectors of T in the target language. The source and target word vectors are represented by $\vec{x}$ and $\vec{y}$, respectively. Data can be seen as two distinct views provided by the context vectors $\vec{x}$ and $\vec{y}$, and supported by each bilingual entry of (v, w) in the seed lexicon. Figure 3.12 shows an illustration of the data representation.

Data for each language can be represented by a matrix. Given the two distinct matrices, the basic algorithm using the CCA is elaborated as follows:

1. *Product matrix M computation*
   The partitions of a correlation matrix are combined to produce a single product matrix or canonical matrix, M. If the relationship of a set of source words S to a set of target words T is analysed, the correlation super matrix corresponding to the data matrix can be illustrated as shown in Figure 3.13. The super matrix consists of four sub matrices:

   - the sub matrix of inter-correlations between source variables, R_{SS},
   - the sub matrices of the cross-correlations between source variables and target variables, R_{ST}, and its transpose, R_{TS},

- the sub matrix of inter-correlations between target variables, R_{TT}.

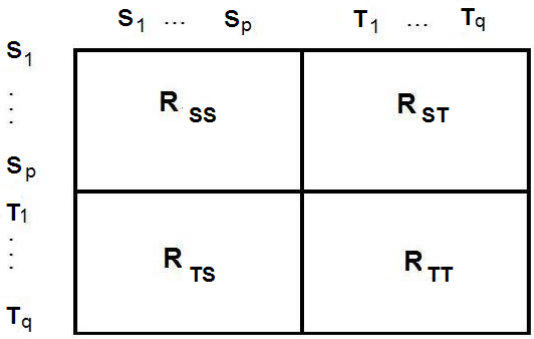

Figure 3.13: An illustration of correlation super matrix

The canonical equation is shown as follows:

$$M = R_{TT} - 1R_{TS}R_{SS} - 1R_{ST}$$

The values in the matrix M can be regarded as expressing the ways in which the data interrelate or overlap. The product matrix M is an asymmetric matrix. It is, typically, defined as follows:

$$M = A - 1B$$

where

A, a symmetric matrix, is equal to $A = R_{TT}$, and

B, also a symmetric matrix, is equal to $B = R_{TS}R_{SS} - 1R_{ST}$.

2. *Latent root extraction*

The latent root or the eigenvalue of a square matrix M is a number, λ, that satisfies the following canonical equation:

$$MX = \lambda X$$

where X is a column vector (or known as a latent vector of M).

Given the characteristic equation of M:

$$|M - \lambda I| = 0$$

where I is an identity matrix.

The λ values when X is not 0 can simply be determined by solving the characteristic equation of M since the canonical equation may also be written as:

$$(M - \lambda I)X = 0$$

Each significant λ means there is a significant common pattern identified across the two datasets. The most significant λ is when $\lambda = 1$, during which, the most likely canonical correlation will be statistically significance. Figure 3.14 shows some examples of eigenvalues in the significance test that were computed for a set of Spanish-English translation pairs in this study. In these examples, if P >0.05, we accept the null hypothesis that the two sets are unrelated.

Chi-square Tests with Successive Roots Removed.

Removed	Eigenvalue	CanCorr	LW	Chi-sqr.	df	P
	0.9968	0.9984	0.0000	9670.8744	72	0.0000
1	0.9799	0.9899	0.0006	5478.8015	56	0.0000
2	0.8409	0.9170	0.0274	2627.0129	42	0.0000
3	0.7995	0.8941	0.1719	1285.2523	30	0.0000
4	0.0740	0.2720	0.8574	112.3387	20	0.0000
5	0.0512	0.2263	0.9259	56.2217	12	0.0000
6	0.0241	0.1554	0.9759	17.8427	6	0.0066
7	0.0000	0.0034	1.0000	0.0082	2	0.9959

Alpha = 0.05

Figure 3.14: Our example showing the results of significance test of the latent roots

3. *Obtaining the canonical weights*

These canonical weights indicate the involvement of each of the source

words S and target words T in each of the common patterns, recognized by a significance latent root. Let $S = s_1, \ldots, s_p$ and $T = t_1, \ldots, t_q$. We compute a vector of canonical weights in each language. Each vector of canonical weights B_k for the t_i, where $1 < k < q$, is computed as follows:

$$B_k = \frac{C_k}{\sqrt{\theta}}$$

where

C_i is one of the co-factors C of any row of $M\lambda I = 0$,
θ is the value of sub matrix R_{TT} pre and post multiplied by both of the co-factor C values.

Similarly, a vector of canonical weights for s_i are computed using sub matrix R_{SS}. This particular step is repeated for all $s \in S$ and all $t \in T$. The set of vectors for S is called U-variates, or the vectors of left hand weights, whereas the set of vectors for T is called V-variates, or the vectors of right hand weights.

4. *Finding the correlation*
   Canonical weights provide the degree and the direction of involvements of each of the source words and the target words in the common pattern. The similarity between the degree and the direction indicates a correlation between the source word and the target word in the underlying dimension.

   Figure 3.15 shows examples of variates containing canonical weights in the U-variate and V-variate tables; each rows of variate represents a vector for a test word. In these examples, the most significant common pattern is observed between the sets of canonical weights, which is shown at the first column of U-variate and V-variate tables. A weak example of a pair of variates having similar degrees and directions are shown by the second vector in the U-variate and the second vector in the V-variate, i.e. 0.0410 and 0.0250. These pairs of variates represent the vectors for

```
------------------------------------------------------------
  Uvariates (left hand) =

   1st Common Pattern   ........   7th Common Pattern
         -0.0819                .          .
          0.0410
          1.0111                .          .
         -0.0537
         -0.0032
          0.1175
         -0.0057                .          .
          0.0514                .          .
          0.1177                .          .

------------------------------------------------------------
  Vvariates (right hand) =

   1st Common Pattern   ........   7th Common Pattern
         -0.0473                .          .
          0.0250
         -0.1052                .          .
          0.0013
          0.0026
          1.0483                .          .
          0.0048                .          .
          0.0220                .          .

------------------------------------------------------------
------------------------------------------------------------
Notes: Variate in rows and weight functions in columns.
```

Figure 3.15: An example showing the U-canonical functions and V-canonical functions

a Spanish-English translation pair of (*transparencia*, transparency). Figure 3.16 illustrates the CCA-based method in general.

3.5.1 Experimental setups

For the experiments, the $MyEuroparl$ and the Lex_{700} are employed. In addition, the Lex_{100} is also used in order to observe the performance of the CCA in a smaller-sized dimension.

3.5.2 Evaluation results

The extended CCA-based algorithm is a combination of the standard and the CCA approaches. The standard approach is used to provide confident candidate translation pairs in a ranked and sorted order, and the CCA-based approach is used to 'verify the candidate pairs proposed by the standard approach. We have chosen two systems of the best settings from the evaluation

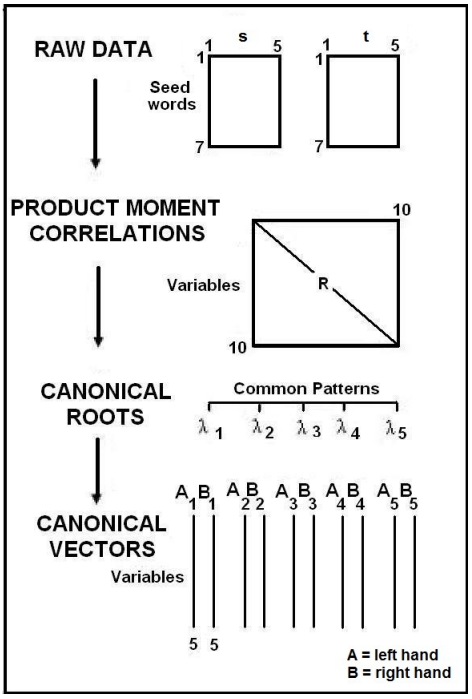

Figure 3.16: An illustration showing some examples of the steps required to acquire data in latent space

sets described in the previous sections to be compared to the systems that used the CCA-based models.

When the CCA was first used in the system (i.e., the $CB + 700 + CCA$), the results was too poor that no score is recorded. In this system, large datasets were used, and as expected the performance was seriously affected by noise that commonly occurs with a multi-dimensional data. Another CCA-based system, i.e. $CB + 700 + Cos + CCA$, only took datasets from confident candidate pairs proposed by a standard context-based system. This system yielded a low F_1 score of 29.7% and unable to outperform the standard context-based approach.

We took another approach by grouping the ranked candidate pair list into

Setting	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$BestF_1score$
CB+700+Cos	58.3	61.2	64.8	55.2	52.6
CB+700+CCA	0.0	0.0	0.0	0.0	∞
CB+700+Cos+CCA	62.3	38.2	31.4	24.4	29.7
CB+700+Cos+CCA*	65.6	65.5	59.2	44.6	42.0
CB+700+Cos+CCA**	80.0	71.4	62.5	50.0	57.5
CB+100+Cos+CCA	0.0	0.0	0.0	0.0	∞
Haghighi's	91.1	81.3	80.2	65.3	58.0

Table 3.7: Performance of different CCA models compared to the context-based model $CB + 700 + Cos$

small groups of 10. Each group was processed separately using the CCA-based system. We only considered candidate pairs with high similarity in terms of degrees and directions. The model is denoted by $CB+700+Cos+CCA^*$. The performance of the system based on this model was improved by 10%. The results were not impressive especially because multiple translations occurred in different lists were selected. When words with multiple translations were put into the same group, the performance of the system further degraded. The reason might be caused by the insignificant context terms provided by the seed lexicon, hence, resulting weak features in the latent space.

Another approach similar to Haghighi et al. (2008) was then employed for $CB + 700 + Cos + CCA^{**}$. In Haghighi et al. (2008), the EM was employed to find source words and target words that were expected to match and to maximize the matching pairs to find the best parameters in order to refine the next matching pairs. In the $CB+700+Cos+CCA^{**}$ system, a list of the first 10 most confident candidate pairs was used as a starting point, before more candidate pairs were added to the list gradually. When any of the candidate pairs affected the common pattern badly, this candidate pair was discontinued from participating in the matching and replaced with another candidate pairs. Only the first common pattern was considered. However, the larger the

datasets the less significant the common pattern would become.

Nonetheless, the CCA-based system managed to outperform the other systems with 57.5% of best F_1 score, although it was only a slight improvement over the system using standard context-based method. The best extended model (i.e., $CB + 100 + PMI + CCA$) performed very poorly, especially when only 100 entries were available in the initial bilingual lexicon.

The performance of the basic context-based system within similar settings was not as expected. The poor performance might be related to the initial bilingual entries that contains mostly high frequent corpus words. Table 3.7 shows the results yielded by this system. The last row of the table shows the results recorded in Haghighi et al. (2008)' study specifically involving the context features for comparison purposes. Their best F_1score is slightly higher than the score achieved with our best system. Their approach is certainly more refined with the EM, thus, better performance is expected with their systems.

3.5.3 Discussion

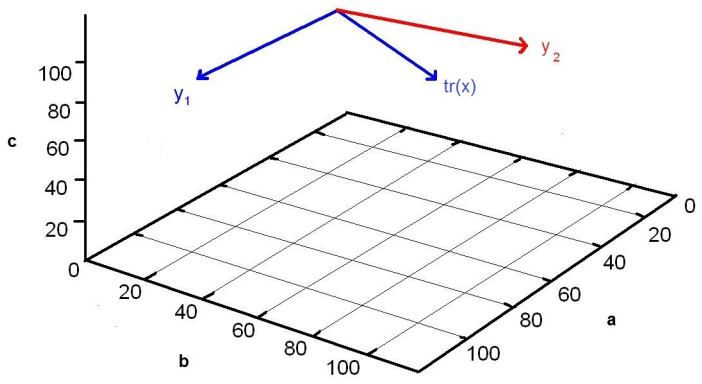

Figure 3.17: Word pairs can be mismatched in a high dimensional space

In the vector space model, each unique, basis term of one language in the initial bilingual lexicon is considered as a dimension of a word space. If this word space represents the source language, the source word is then represented as a vector in the space. Similarly, a target word is represented as a vector in other word space for the target language. To allow matching, one of the vectors has to be translated into the other language. Hence, a vector-based approach for bilingual lexicon extraction is strongly dependent on the initial bilingual lexicon. An m-sized bilingual lexicon represents an m-dimensional word space. Hence, a large bilingual dictionary provides a high dimensional space.

A distance between two vectors in high dimensional space can be misleading. For example, let a, b and c be the dimensions in the word space. A translation of a source word vector into the target language is denoted by $tr(x)$ and a set of two target word vectors is denoted by $y = \{y_1, y_2\}$. According to the distance in Figure 3.17, $tr(x)$ is closer to y_2, though, y_1 is the correct translation for x. This problem may happen due to missing or insufficient relevant features, or highly-occurring irrelevant features. Moreover, in these word spaces, terms are considered independent from one another. However, this is contrary with the common knowledge because the term dependency can arise in written or spoken languages in many different ways. The most common are through synonymy and polysemy.

If one assumes the availability of a large, general bilingual lexicon of 10,000 entries, wider coverage would be expected. It means that the bilingual lexicon may be able to cover widely, if not completely, the context features for each test word. Hence, efficient matching would be performed. On the other hand, a large bilingual lexicon can also contribute to irrelevant features that can mislead the similarity measure. Missing important features in the lexicon may also cause a similar problem. However, more serious problems would be expected with smaller bilingual lexicons. Nonetheless, sparse corpus data also

contributes to these problems.

Some believe that a latent space could resolve most problems with the high dimensional space. The dimensions are reduced using some dimension reduction techniques. The selection of dimension reduction techniques depends on the data we have in hand and data that we are looking for. Some techniques involve different starting data matrix, and others involve two distinct matrices. These techniques also compute different outcomes for the eigenvalues. The CCA is preferable for bilingual lexicon extraction than other dimension reduction techniques, most probably, because the CCA could represents two datasets.; which makes the CCA feasible to process data of two different languages. We have mentioned previously that CCA helps find linear combination of the source and target word vectors, which have maximally correlated with each other. It could also shows the dimensions that are shared by both words (i.e., the variables are intercorrelated between the two sets) using the canonical scores.

However, a reduced dimension might have incorrectly conflated the critical dimensions, especially when different irrelevant features occur too highly in the source and the target language. In other words, large dimensions might introduce more noise to the CCA, thus, the system would become less efficient. For example, a careful measure helps the $CB + 700 + CCA$ to perform better compared to the time when all data were considered in this system. Hence, the matching process is not only affected by the correct matching pairs and the sufficient number of context terms, but also by the dimension sizes.

3.6 Summary and conclusion

This chapter has discussed many important components required in order to perform the bilingual lexicon extraction task. General approaches introduced or used in the previous studies have also been described. To gain experiences,

some base experiments have been conducted and reported in this chapter, from which we selected the baseline systems for comparison purposes in this study.

Through experiences, we observed that a little 'trick' such as eliminating noise words from a corpus helps improve the performance of the systems. We also found that the context-based approach is quite simple, straightforward and works fine. We first thought that using latent data will produce outstanding results compared to using high dimensional data, but the results we achieved with the CCA-based models were not very outstanding despite of the hassle we had gone through with the work. Moreover, the approach is difficult to apply and the results can easily be misinterpreted. We conclude that the reliability of the context-based approach is quite difficult for other approaches to beat. Last but not least, we would require more than just a trick to obtain better results, but perhaps not as difficult as the CCA.

The next chapter will introduce new, novel methods for bilingual lexicon extraction proposed in this study. The baseline system that has been chosen for the experiments is the system using the cosine, the $MyEuroparl$ and the Lex_{700}. Each model we propose in this thesis aims to obtain higher precision bilingual lexicons than the baseline system. We demonstrate the experiments using minimal resources, which mostly have already been described in this chapter. We called this approach as the minimally-supervised techniques due to the resources that have been used in the experiments, although they are likely to perform better with larger, quality resources.

Chapter 4

Utilizing Contextually Relevant Words

This chapter discusses a minimally-supervised technique that aimed to improve the source word and the target word vocabulary lists. As discussed previously, removing noise words from the vocabulary lists is a good, simple technique, which allows more correct translation pairs to be collected. This time, another technique based on contextually relevant words is introduced. This technique is demonstrated using the small-sized MyEuroparl comparable texts. Through experiments, a system based on this technique has shown a slightly better performance compared to the baseline system. As an addition, the technique was also applied to the spelling-based model, and as a result, the system performance of a system based on this approach achieved more than 20% of best F_1 score when this system was compared to the baseline system.

4.1 Introduction

The most ideal source word and target word vocabulary lists are the ones containing target words that are among acceptable translations for each source word of the vocabulary lists, and vice versa. The most inspiring example was found in Rapp (1995)'s study; which includes a simulation of the co-occurrence patterns between a set of 100 English words and a set of 100 Spanish words,

where all the Spanish words were actually the translations of the English words. Identical common patterns were clearly shown in that example. Likewise, a similar pattern was simulated and discussed by Fung (1995).

Any specific approach, with regards to the source words and target words, seems to be taken lightly in previous studies. In general, most of the studies seemed to agree that the source word and the target word vocabularies should be among the most frequent words of the source and target language corpora (Haghighi et al., 2008; Fung, 1995). Rapp (1999) used frequent words, with a frequency of 100 or higher. Some other researchers, such as Fung (1995), took less ambitious approach by considering known translation pairs, which were hand-compiled from a bilingual lexicon but still among frequent words occurring in the corpus. Fung (1995) also suggests removing noise words, especially, function words such as the English words `by` and `from` from texts to improve the source and target word vocabulary lists. Nonetheless, there were also some researchers, such as Haghighi et al. (2008) and Koehn and Knight (2002), who only considered noun word type to be useful vocabularies. In general, each approach in previous studies was not much different to one another.

Nonetheless, using most frequent words is a good approach, however, this approach alone seems to be quite loose due to the limitations of comparable corpora. Many translations of the source words could still be missing from the target word vocabulary list and the occurrences of other non-equivalent target words in the context might have affected the matching results. Having good source word and target word vocabulary lists might avoid many mismatching to happen; one good practice is to eliminate stop words, especially when high frequency words are to be involved.

In this study, a new technique is discussed, which aimed to generate good source and target word vocabulary lists from within similar contexts. The technique relies on a list of cognate pairs to find each a set of context words that we called as the *contextually relevant words*. These contextually relevant

words are then utilized to form the source and target word vocabulary lists. That means, each of the cognate pair would be provided with a set of source words and a set of target words of its own, which were taken from within a context (of the cognate pair). The source words and the target words are then matched and evaluated. Based on this technique, a system in this study has achieved higher precision scores compared to a baseline system, where a precision score of 79% at 50% of recall value was recorded (which means that most candidate pairs in the first half of the output list, produced by the system, were correct translations).

Likewise, a similar technique was applied to a spelling-based model. Based on this technique, a spelling-based system has obtained 85.4% of precision score at 50% of recall value. The system has performed better when only the highest ranked target candidate (top 1) was chosen for each source word. In addition, by using a string edit-distance versus precision curve, we also reveal that this contextually relevant words technique has allowed the systems to correct word pairs efficiently.

4.2 Methodology

Previous work in bilingual lexicon extraction, such as Haghighi et al. (2008); Rapp (1999)'s studies, used lists of high frequency words that were obtained from a bilingual corpus of a source and a target languages to be the source and target word vocabulary lists, respectively. In this study, the aim is to extract higher precision bilingual lexicon using an improved approach. Instead of just using ordinary high frequency words to form the source word and the target word vocabulary lists, the lists could be further improved by considering the contextually relevant words (i.e., context words that highly co-occur with cognate pairs). This technique would be used to restrict the contexts of the source and the target words, thus, this technique could help increase the possibility to find correct matching pairs. Figure 4.1 shows an illustration of

the model.

Cognate pairs could be derived automatically by mapping or finding identical words, or very similar spelling words, that occur in the high frequency word lists. Figure 4.2 shows an illustration of the cognate pair extraction process.

In order to utilize the contextually relevant words of a cognate pair, a cognate of one language would be used to find its own set of contextually relevant words. The pair of another language would follow a similar process. An association measure such as the LLR could be used to identify the contextually relevant words.

In this technique, once all contextually relevant word lists are obtained, the lists are then sorted and ranked. Only the top lists are to be considered as the source and target word vocabulary lists, and later, would be used in an extraction task. Figure 4.3 shows some examples of matching source words and target words that we found within the context of identical cognate pairs (`civil`,*civil*). This approach is meant to be used at the initial stage of a bilingual lexicon extraction process, hence, this technique could be applied to a context-based or a spelling-based model. Sub Section 4.2.1 describes a method based on this technique.

4.2.1 Using cognate pairs to restrict the contexts

The contextually relevant words technique only considers the lists of the source words and target words that co-occur within the context of a cognate pair to be in the extraction process. A method based on the technique is elaborated as follows:

1. *Acquiring cognate pairs automatically*
   Two high frequency word lists $HFWs$ and $HFWt$ were extracted from

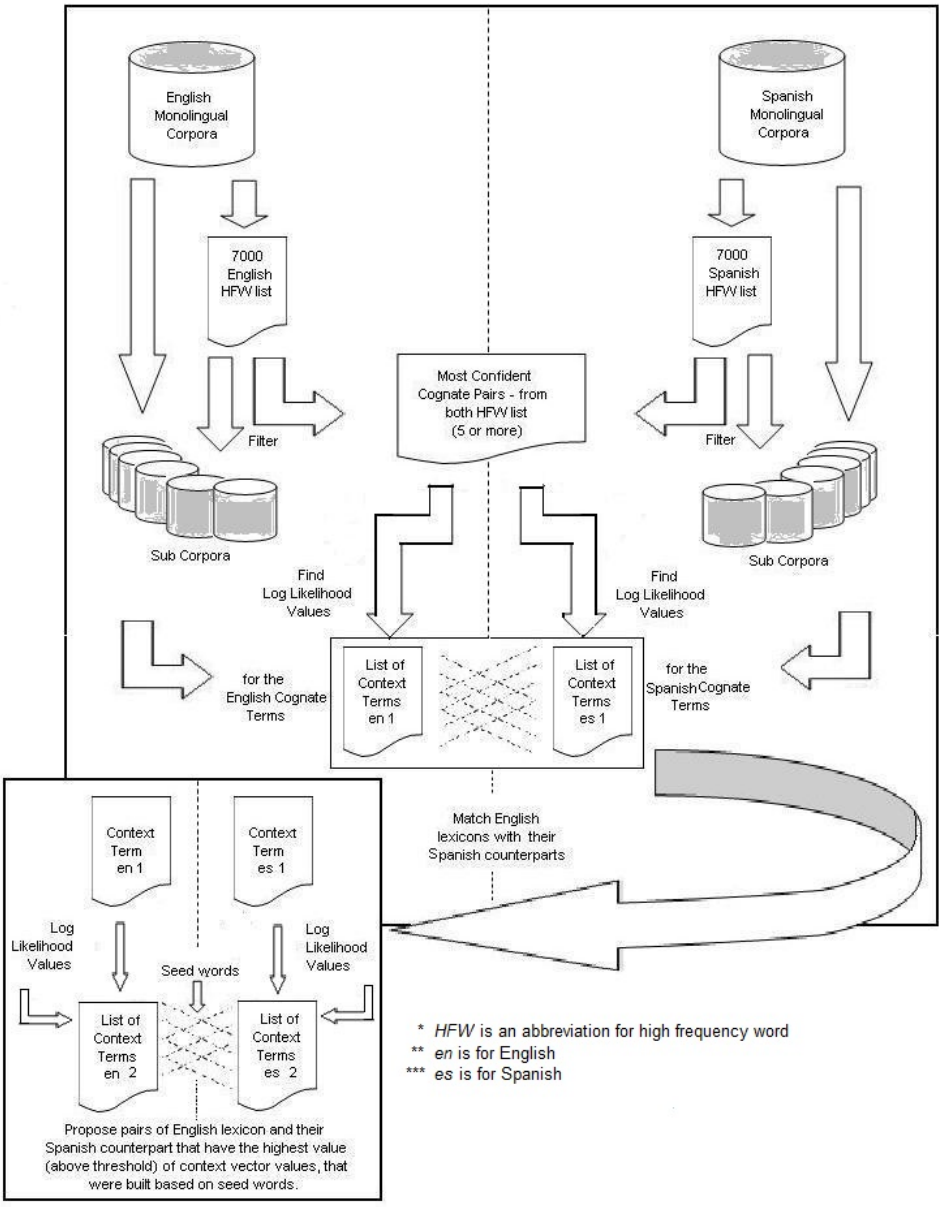

Figure 4.1: An illustration of a model using cognate pairs to derive contextually relevant words in order to form the source word and the target word vocabulary lists

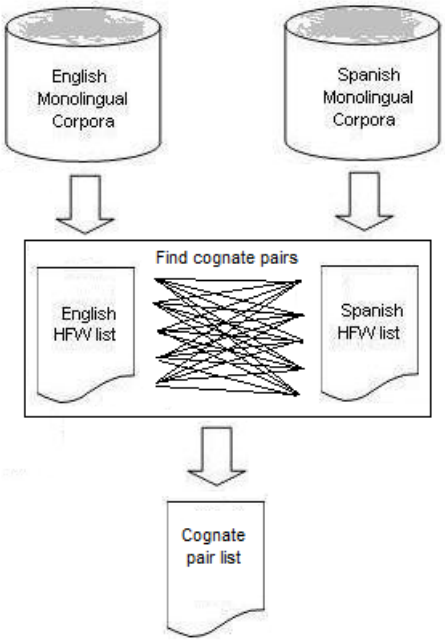

Figure 4.2: Cognate pair extraction

comparable corpora of different languages; each corpus is denoted by $Corpus_S$ and $Corpus_T$ according to their respective languages. Each word in $HFWs$ is initially paired and then compared with each word in $HFWt$. For all word pairs of $HFWs$ and $HFWt$, the word pairs having non-identical and less similar spelling words were removed immediately. The remaining word pairs in the lists were considered as cognate pairs (CS, CT).

2. *Finding contextually relevant words*
   Given cognate pairs $(CS, CT) = \{(cs_1, ct_1), \ldots, (cs_n, ct_n)\}$.
   For every cs_i,

   (a) All sentences containing cs_i were collected to form $SubCorpus_{cs_i}$.

   (b) Using window size of a sentence, the LLR was computed for all words that co-occur with cs_i.

Figure 4.3: Examples of bilingual word pairs that were found within the context of the cognate word `civil`

(c) All the context words (or contextually relevant words $CRcs_i$) for cs_i were then sorted and ranked according to the LLR values.

The steps, from (a) to (c), were repeated for every ct_i to obtain its contextually relevant words $CRct_i$.

3. *Obtaining the source word and the target word list*
The top 100 of highly ranked contextually relevant words for each $CRcs_i$ and $CRct_i$ formed the source word s vocabulary list and the target word T vocabulary list, respectively.

4.2.2 Building context vectors

A similar method for building context vectors as described in the previous chapter was conducted in this stage. Context terms that were not found among entries of an initial dictionary or a seed lexicon were completely removed. Only the remaining context terms with their known translations were then used to build the context vector for s, or t, respectively. As an example, the words `community` and `democracy` occurring in the seed lexicon were among the high occurrence context terms of the source word `powers`. Thus, the words `community` and `democracy` has become the term vectors (i.e., the elements of a word vector) representing the word `powers`. Their translations were used to represent word vectors in a word space of the target language.

C[i,j]	***community***	¬ ***community***		
powers	124	1831	**1955**	*C(powers)*
¬ ***powers***	11779	460218	**471997**	*C(¬ powers)*
	11903	**462049**		
	C(community)	*C(¬ community)*		

Here $C[i, j]$ denotes the count of the number of sentences in which i co-occurs with j.
Total corpus size, N = 473952 in the above

Table 4.1: Contingency table for observed values of target word $t = powers$ and context word $b =$*community*

LLR was used in this study as the association measure to help judge whether the term was likely (or unlikely)to occur by chance in the context. In order to compute the log-likelihood of target word t occurring with context word b, a contingency table was created. The contingency table contained the observed values that were taken from a given corpus (see Table 4.1 for an example of the contingency table for the target word `powers` and the context word `community`). For each set of a target word t and a context word b, the LLR value was given by:

$$LL(t,b) = 2 \sum_{i \in \{t,\neg t\}, j \in \{b,\neg b\}} p(i,j) N \log \frac{p(i,j)}{p(i)p(j)}$$

where

$p(i,j)$ = observed joint probability of i with j

$p(i)p(j)$ = expected probability

$p(i.j)$ can be estimated by $\frac{C[i,j]}{N}$. Similarly for $p(i)$ and $p(j)$.

4.2.3 Measuring the similarity

Bilingual word pairs were extracted by matching each source word s with every target words t using their context features. Let x denotes the source word vector and y denotes the target word vector. To determine whether the $\{s, t\}$ is a translation pair, the similarity between their context vectors was computed automatically using a vector similarity measure such as the cosine measure. Prior to this step, the term vector values were first transformed from real values into binary values (see previous chapter regarding the issues with the cosine measures).

In addition, spelling similarity could also be used to match potential bilingual word pairs. In this study, a list of candidate pairs were extracted by matching the source words and the target words of the vocabulary lists using the string edit distance.

4.3 Experimental setups

This section describes the experimental setups used in this study. Some of the settings have been described in Chapter 3, thus requiring no detailed description.

1. *Corpora acquisition*

   We used the English-Spanish MyEuroparl comparable texts, derived from the Europarl parallel corpora (Koehn, 2005) (see Chapter 3 for details). The reason was that: although the use of such corpora would

not be able to illustrate a real extreme scarce resources problem, but the minimal use of the resources deemed to be suited for demonstrations of a minimally supervised technique.

2. *Pre-processing*

   The MyEuroparl corpora were passed through similar pre-processing jobs as described in Chapter 3. For corpus pre-processing, only sentence boundary detection and tokenizing were conducted on raw texts. All tags were cleaned up and stop words were removed completely from the corpus. Some special case processing jobs were also conducted.

3. *Seed lexicon*

   We used the Lex_{700} that was already mentioned in the previous chapter.

   The approach used to obtain the lexicon is very much simpler than acquiring general dictionaries of 10,000-20,000 bilingual entries (Rapp, 1999; Fung and McKeown, 2004), or acquiring the seed words automatically (Koehn and Knight, 2002; Haghighi et al., 2008). However, this approach could only work if the source language and the target language are fairly related and both share lexically similar words that have same meaning. Otherwise, general bilingual dictionaries might be the only option.

   In addition, the size of a small seed lexicon is defined as the size ranging between 100 to 1,000 word pairs. Hence, seed lexicon containing 700 cognate pairs were still considered as a small-sized lexicon.

4. *List of cognate pairs*

   79 identical cognate pairs were successfully obtained from the top 2000 high frequency lists, which were extracted from the source and target language corpora. Figure 4.4 shows an excerpt of high frequency word lists that we obtained from the comparable corpora and Figure 4.5 for an excerpt of cognate pairs that were found among the high frequency word

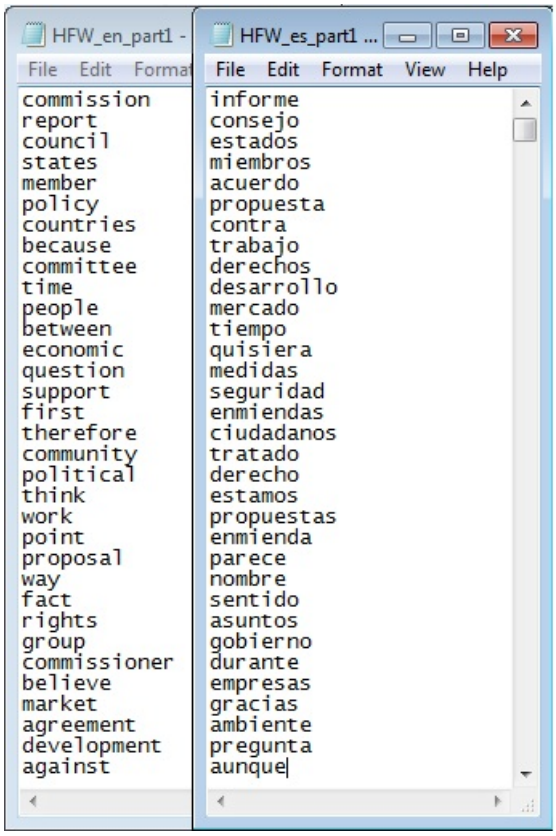

Figure 4.4: An excerpt of high frequency word lists that were kept in separate text files according to their languages

lists of different languages. However, only 55 of them were considered for having at least 100 contextually relevant terms highly associated with each of the cognate pair. These cognate pairs might have also been included in the source word and the target word vocabulary lists because some of them might also co-occurred with other cognates that were used to restrict the context at that time; hence, if these happen, the cognate pairs should be removed immediately from the lexicon.

5. *Baseline system*

   A baseline system based on the context-based model is described in the previous chapter. Another baseline system based on the spelling-based

```
crisis,crisis
civil,civil
central,central
amsterdam,amsterdam
language, language
control,control
debate, debate
particular,particular
fundamental,fundamental
decision,decision
global,global
normal,normal
regular,regular
helsinki,helsinki
individual,individual
europe, europeo
european, europea
respect,respecto
important, importante
resolution,resolucion
situation,situacion
embargo,embargo
possible,posible
programme,programa
problem,problema
national,nacional
cooperation,cooperacion
general,general
social,social
moment,momento
sector, sector
```

Figure 4.5: An excerpt of English-Spanish cognate pairs derived from high frequency word lists

model was also built. The spelling-based model is originated from Koehn and Knight (2002). In the model, high frequency word lists of both languages are matched to one another based on identical and similar spelling features. In addition, a threshold of a 4-letter word length, was introduced; the length for all test words may only be equal to or more than four to be considered.

6. *Reference lexicon*

   A reference lexicon that was extracted from the *Word Reference* was used in this study (see Chapter 3 for details).

7. *Evaluation*

   There were two sets of evaluation in this experiments; one having multiple translations for each source word, and another that restricting the translations to be one (the most confidence) candidate only for source

word.

For the first evaluation,the threshold, $\theta = 2000$, was set for the size of the proposed candidate pair list, which means that only the first 2000 of (s,t) candidate pairs from the proposed list were considered, including word pairs containing source words with multiple translations (i.e., any source word s having paired with multiple target words t_i, where each t_i is the i-th target word in the n-sized target word list where $1 < i < n$). This evaluation was conducted for both context-based and spelling-based models, but in separate sets of experiments.

For the second evaluation, the method have been described in the previous chapter. Similar threshold value for the proposed list to the first evaluation was used, but this time, the evaluation only involved the first 2000 of (s,t) candidate pairs, where s having paired only with the highest ranked t, or top 1.

4.4 Evaluation results

This section presents the experiment results for systems of different models that were conducted in this study.

4.4.1 General candidate pair lists

For evaluating general candidate pair lists, several systems were involved. The systems for context-based and spelling-based are labelled with *ECS* and *ESS*, respectively. Whilst, the baselines for context-based and spelling-based are labelled with *CS* and *SS*, respectively. From the results, systems based on both *ECS* and *ESS* models achieved over 50% of the F_1 score. However, the results were only 1% to 2% of error reduction over the baseline systems, hence, the performances of both systems were not quite impressive. Nonetheless, in terms of precision scores at higher recall values, significant improvements were observed especially with the system based on the context-based model. The contextually relevant words approach in the ECS have allowed over 30% of

precision score improvement at 10% recall value. Table 4.2 shows the full experiment results.

Table 4.2: Performance of the ECS and ESS systems compared to baseline systems for 2000 candidates below certain threshold and ranked

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
CS	42.9	69.6	60.7	58.7	49.6
SS	90.5	74.2	69.9	64.6	50.9

(a) from baseline models

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
ECS	78.3	73.5	71.8	64.0	51.2
ESS	95.8	75.6	71.8	63.4	51.5

(b) from our proposed models

4.4.2 Top 1 candidate pair lists

Table 4.3 shows the full experiment results for the second evaluation. Systems based on both *ECST* and *ESST* models yielded almost 60% of best F_1 score.

By using the new, extended method in the spelling-based system (i.e., the ESST model), a significant improvement of 20% of best F_1 score was recorded compared to the baseline system (i.e., the *SST* model). The former obtained 85.4% of precision score at 50% of recall value.

Meanwhile, the context-based systems achieved precision score of 79% at 50% of recall value when the system was based on the new, extended context-based model *ECST*. However, the *ECST* system has not recorded a significant difference over the baseline *CST* system as expected; when only 57.1% of best F_1 score was recorded, compared to 52.6 % of best F_1 score recorded by the baseline system.

Table 4.3: Performance of the ECST and ESST models compared to baseline systems for 2000 candidates of top 1

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
CST	58.3	61.2	64.8	55.2	52.6
SST	84.9	66.4	52.7	34.5	37.0

(a) from baseline models

Setting	**P**$_0$.1	**P**$_0$.25	**P**$_0$.33	**P**$_0$.5	**Best-F1**
ECST	85.0	81.1	79.7	79.0	57.1
ESST	100.0	93.6	91.6	85.4	59.0

(b) from the proposed models

The overall performances for these four models, represented by precision scores for different range of recalls, are illustrated in Figure 4.6.

4.4.3 String edit distance value vs. precision score

It is important to see the inner, underlying performance of the *ECST* model with a further analysis. A string edit distance value (*EDv*) versus precision score curve for the *ECST* and *CST* models are introduced and shown in Figure 4.7. The curve could be used to measure the performance of the *ECST* model in terms of capturing bilingual pairs with less similar orthographic features, those that might not be able to be captured using spelling similarity.

The graph shows that although the *CST* has higher precision score than the *ECST* at *EDv* of 2, it was not significant because the difference was less than 5% and the spelling between the word pairs was still very similar. On the other hand, the precision for the proposed lexicon with *EDv* above 3 for the ECST (where the spelling of the source word s and the proposed translation equivalent t becoming more dissimilar) was higher than the *CST*. The most significant difference between the precision scores was almost 35%, in which the *ECST* achieved almost 75% of precision score compared to the *CST* with

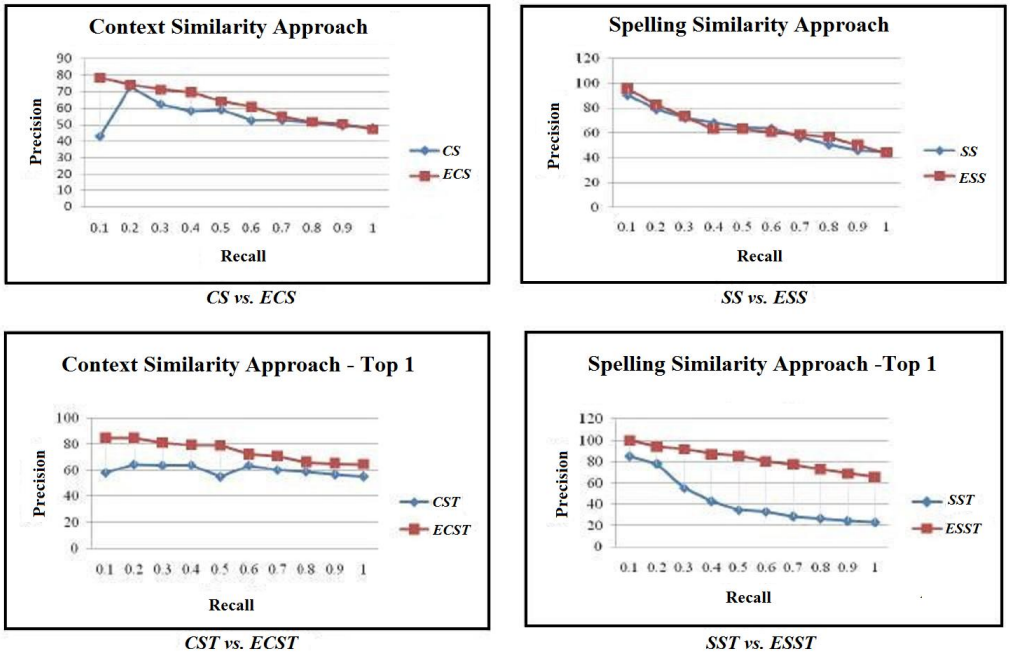

Figure 4.6: Performance of different models

40% of precision score at *EDv* of 4. It is followed by the *ECST* with almost 50% of precision score compared to *CST* with precision score less than 35%, offering about 15% improvement of precision score at *EDv* of 5.

4.5 Discussion

Discussions are provided in the following sub section.

4.5.1 Contextually-relevant word based model vs. baseline model

The *ECST* system recorded some improvement of the best F_1 score over the *CST* (baseline) system but it was not significant. However, by utilizing the contextually relevant terms, the *ECST* system was likely to gather more correct candidate pairs in the proposed candidate pair list, especially when it comes to word pairs with dissimilar spelling. These findings were aided by the

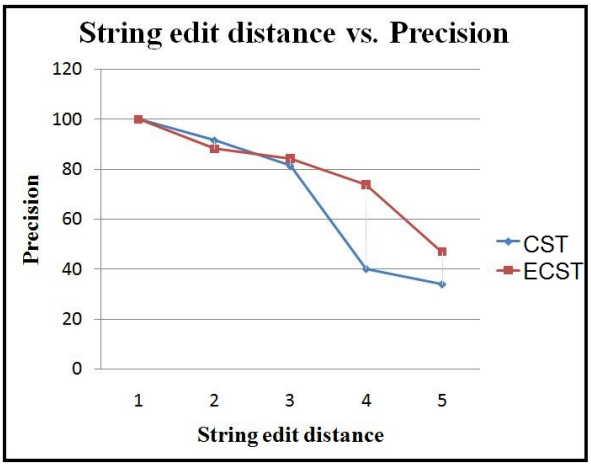

Figure 4.7: String Edit Distance vs. Precision curve

string edit distance value (*EDv*) vs. precision curve. That means the *ECST* was able to add more correct translations compared to the *CST*, though the number of cases happened were still considered small.

Some examples that show the *ECST* has some advantages over the *CST* are as follows: the Spanish words `autentica` and `fortalecimiento` were incorrect translations for the English word `clause`, but both Spanish words were ranked highly in the proposed list (output) of the *CST* system for the word `clause`. Surprisingly, the *ECST* was able to avoid this mismatching by effectively removing the false target words at the initial stage. (See Figure 4.8 for many other examples).

4.5.2 Context-based model vs. spelling-based model

The performance of an extraction system seems to be more efficient with the spelling-based approach when related language pairs were involved. The most outstanding result for such approach, in terms of precision, is shown in the CS vs. ECS graph, especially, at 10% of the recall value. However, this spelling feature is not quite useful to extract most word pairs in the comparable cor-

Source	Target	ECST		CST		Rank
		Candidate found	Sim. value	Candidate found	Sim. value	
clause	*clausula*	***clausula***	***0.402015126***	*autentica* *fortalecimiento* *economico* *respeto* *vigor* ***clausula***	*0.447213595* *0.430331483* *0.412478956* *0.40824829* *0.402015126* ***0.402015126***	1 2 < > < > < > < >
pillar	*pilar*	***pilar***	***0.547722558***	*daramente* ***pilar*** *basada* *comercial* *iniciado* *exterior* *agricola*	*0.632455532* ***0.547722558*** *0.53935989* *0.516397779* *0.516397779* *0.478091444* *0.447213595*	1 2 3 4 4 5 6
state	*estado*	***estado***	***0.433012702***	*derecho* ***estado*** *respeto*	*0.43519414* ***0.433012702*** *0.412478956*	1 2 < >
confidence	*confianza*	***confianza***	***0.424264069***	*errores* *desarrollo* *haberse* *demuestran* *deficiencias* ***confianza***	*0.447213595* *0.447213595* *0.447213595* *0.447213595* *0.447213595* ***0.424264069***	1 1 1 1 1 2
welfare	*bienestar*	***bienestar***	***0.40824829***	*hubiera* ***bienestar***	*0.500000000* ***0.40824829***	1 < >

Figure 4.8: Some underlying examples that show the effectiveness of the *ECST* compared to the *CST*

pora. We contend that combining both *ECST* and *ESST* models would help a system to produce more correct translation pairs.

The experimental findings in this study has shown that there are some advantages of using contextually relevant word technique, specifically; and context-based, generally; which are as follows: 1. the technique helps reduce some possible errors due to mismatching, and 2. a context-based model has more potential to extract most word pairs from comparable corpora. Both of the corpora used in this study were of similar domains, hence, the potential of this technique in its entirety has not been observed. We would like to test the technique with different type of corpora in the future.

Nonetheless, the approach of using cognate pairs as seed words is more appropriate for language pairs that share large number of cognates, or similar spelling words with similar meanings. Otherwise, one may have to rely on general, bilingual dictionaries.

There might be some other possible supporting strategies, which could be used to further improve the precision score. For example, other techniques based on the noise reduction such as the re-ranking method might be useful for this study.

4.5.3 Word hypernymy and hyponymy

Many spurious translations because of many false target words have higher correlations with the source word compared to the correct translations are still in concern. Some of the examples have been shown in Chapter 3 (see Table 3.6, page 149). The most common errors detected in the top 100 were of semantically related words, which had strong context feature correlations.

We contend that the problem was naturally caused by imbalance corpora that we used in this study; the corpus in the target language might have smaller coverage compared to the corpus in the source language, hence, more general target word is proposed to a less general source word. An error analysis, which was conducted on the translations of the source words in the example, showed that their correct translations were among the top 5 in the output list. Thus, we relate the problem to the hypernymy problem. Nonetheless, the problem could also be related to the hyponymy, which is usually caused by documents in the source language having smaller coverage compared to documents in the target language. Unfortunately, we could not find any related examples in this study to support the thought.

4.6 Conclusion

In this study, as we were working on the English-Spanish comparable corpora, we could have focused solely on spelling similarity feature to obtain a high precision bilingual lexicon; especially, because the performance of systems using this spelling feature usually recorded high performance. Related experiments that were conducted in this study based on this particular feature have recorded 100% of precision score at lower recall, especially when only the highest ranked target candidate was considered for each of the source word.

However, we were more interested with context features. One main reason is because the context feature would have more potential to extract most word pairs from the comparable corpora, especially if less related language pairs are involved. Another reason is that the approach of taking in word pairs using cognates alone to extend the initial bilingual lexicon might not be a reliable approach, because sometimes correct translations are not always a cognate even though a very much identical or similar spelling word for the word is available. The potential of the context-based approach compared to the spelling-based approach could be shown using the string edit distance value versus precision

score curve.

This chapter has mainly presented a new technique that utilizes contextually relevant words. These set of contextually relevant words could be used to form 'an improved version' of vocabulary lists of the source word and the target word. To extract the contextually-relevant words, cognate pairs are to be used. Through experiments, we have shown that this technique could help improve the learning process on small-sized, non-parallel but comparable corpora, especially when only high recall is considered.

Based on this technique, the spelling-based system has obtained 85.4% precision score at 50% of recall value. Precision of 79% at the same recall value was recorded when the technique was applied to a context-based model. More significant achievements were recorded, especially when only the highest-ranked translation candidate was considered for each source word. In addition, the new spelling-based system was found to be as efficient as expected, but the number of correct pairs proposed by the system seems to be limited compared to the output of a context-based system that based on similar technique. However, both were able to capture words efficiently compared to the baseline systems, thus, showing the potential of the contextually relevant words technique. Moreover, the experimental findings might have offered strong evidence that any vocabularies to be involved in an extraction process should be carefully selected beforehand; although this approach might not be able to guarantee great performance.

We have applied a technique that restricted the contextual boundaries of the source word and target word vocabulary lists, which helps avoid some of the mismatching from happening, but thus far, we were still getting other incorrect translation pairs. This issue would be further addressed in the following chapter.

Chapter 5

Using In-domain Terms in Context Vectors

In the previous chapter, the author described a technique that transforms the initial stage of the bilingual lexicon extraction task. This chapter discusses another new technique, but this time the stage of the task that relates to the context terms would be involved. The novelty of this technique relies on the notion of the in-domain terms *which can be thought of as the most important context term sets for each source or target word. A system, which was built based on this technique, has yielded over 80% of the best F_1 score. This score was 15% more than the score achieved by a baseline system in similar settings. Furthermore, the results were more significant (with nearly 30% difference of best F_1 score) when compared to another baseline system, which has beed described in the previous chapter. Hence, a model based on this technique would be useful for an extraction process to help produce high precision bilingual lexicon. In addition, a novel method for measuring similarities between words of different languages without the use of existing bilingual lexicon is also presented.*

5.1 Introduction

In the standard context-based model, the source word and the target word need to be associated with their context vectors before these two words can be compared to one another using an initial bilingual lexicon and a similarity measure. However, there are possibilities that not every important context terms would be occurring in the training corpora; whilst, the ones occurring might have low occurrences and could be easily missed. Whilst, some may occur but with low frequency and can be missed. Hence, learning from comparable corpora might be particularly problematic due to scarce data. In addition, the limitations with small-sized initial bilingual lexicons could further hurt the learning. Such kind of initial bilingual lexicons could also have contributed irrelevant, or less relevant, features that could mislead the similarity measure, especially when the numbers of dimensions are large. (See Chapter 3 that provides discussions on the vector-based approach's problems).

An approach that might help overcome the problems is based on an assumption that: *if two highly associated terms share certain features, their corresponding translations should also be highly associated and share similar features.* The features used here are the sets of context terms that mutually occur in similar domain though the idea may be extended to other kind of features. Figure 5.1 shows an example of the context term set.

In the example, the source word `powers` is highly associated with the word `delegation`; and both share common context terms such as the English words `parliament` and `affairs`. Now the translation equivalent of the English word `delegation` is a Spanish word *delegacion*. The word *delegacion* is highly associated with the word *poderes*, which is a potential translations for the source word `powers`. The common context terms shared by the words `powers` and *poderes* are the terms *parlamento* (`parliament`) and *asuntos* (*affairs*). Hence, the translation equivalents of the words `powers` and `delegation` in the target language are not only highly associated but they also share common

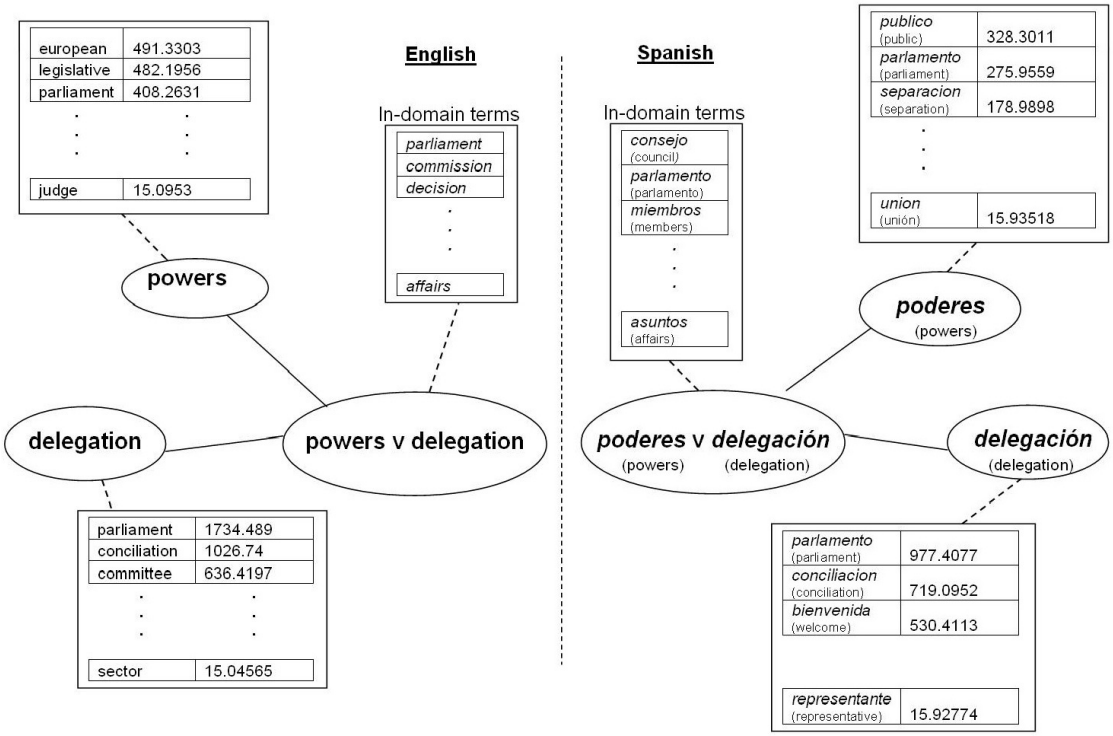

Figure 5.1: An example of in-domain terms that co-occur in English and Spanish. The source word is `powers` and the target word is *`poderes`*. The words `delegation` and *`delegacion`* are the highly associated words with the source word and the target word respectively. Their in-domain terms, as shown in the middle, can be used to map the source word in context of word `delegation` to its corresponding target word in context of *`delegacion`*.

context terms that are the translation equivalents of the words `parliament` and `affairs`.

It is observed that the common context terms are simultaneously the *first-order* and *second-order* context terms of the target word. They are the shared context terms between the target word and its highly associated context term. These terms are defined as the *in-domain terms*. These in-domain terms can be used to map words to their corresponding translations. The highly associated context terms can be thought of as sense discriminators that differentiate the different uses of the target word. Figure 5.2 shows how the word `delegation` helps in selecting between the "control or influence" sense of the word `powers` while rejecting the "ability or skill" sense.

It is quite clear that the method can be adapted for building sense disambiguated bilingual lexicons. However, the focus of this study is not on sense disambiguation.

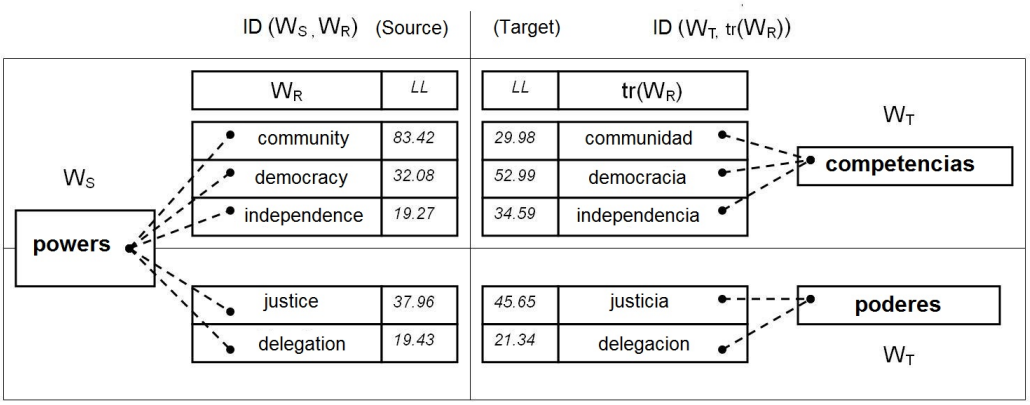

Figure 5.2: An example of English-Spanish lexicon learnt for the source word `powers`. On the top, the system suggested target word ***`competencias`*** and rejected target word ***`poderes`*** when the word `powers` is associated with the word `community`, `democracy` or `independence`. The word ***`poderes`*** is suggested when the word `powers` is associated with the word `justice` or ***`delegation`***.

5.2 Methodology

To use the in-domain terms in extracting bilingual lexicon, each set of the terms in form of a word vector need to be well-identified for each source or target word. Similar word vectors of different languages represent translation pairs.

A method used in this study based on this in-domain terms technique is further elaborated as follows:

1. *Identifying highly associated words*
   Once the source and target word vocabulary lists were established, the context terms would be identified for each source and target words in the lists based on their occurrences, which were derived using the LLR. A threshold was then set for the LLR value to help pick out another list consisting only highly-associated words from the source word's context term list. The step was repeated for each target word to acquire its own list of highly-associated words.

2. *Identifying in-domain terms*
   Given the list of highly-associated word from the previous stage, the context terms for each highly-associated words were then identified. Using both context term lists for the source word and its highly-associated words, in-domain terms (i.e., the common context terms occurring in both list) were extracted to form an in-domain term list for the source word. This step was repeated for each target word and its highly associated words.

3. *Measuring the similarity*
   The aim of this stage is to find potential translations of the source word. Each source word, represented by its set of in-domain terms, was then matched to each target word, also represented by its set of in-domain terms.

5.2.1 Identifying in-domain terms

Only two steps (namely, Step 1 and Step 2) play the important roles for extracting the in-domain terms. However, Step 1 is a part of a common bilingual lexicon extraction so no detailed explanation is required. This sub section describes Step 2 in more details.

Let assume S is a set of unique word $s_1, \ldots, s_n$ that occur in a corpus of the source language, $Corpus_S$, and T is a set of unique word $t_1, \ldots, t_n$ that occur in a corpus of the target language, $Corpus_T$. Each t is a potential translation in the target language for each s. Let R_S denotes a term highly associated with s with log-likelihood $LL(R_S, s) > threshold\ t_1$. Let $tr(R_S) \in Corpus_T$ denotes the translation equivalent of R_S. We assume that an initial lexicon, from which $tr(R_S)$ could be found, is provided.

Indirectly, $tr(R_S)$ also denotes that a highly associated word for a target word t is found in the initial bilingual lexicon. Let that specific highly associated word is denoted by R_T, and the initial bilingual lexicon is denoted by Lex. Lex contains p number of (A, B) pairs, where A is the set of entries of the source language, and B is the set of entries of the target language. Assume that $A = a_1, \ldots, a_p$ and $B = b_1, \ldots, b_p$. Formally,

$$tr(R_S) = R_T \cap b_i$$

where b_i is the i-th entry in the Lex that identical to R_T.

From the previous example, s was referring to the word `powers`, whilst the words *`Competencias`* and *`poderes`* were its potential translations (see Figure 5.2). Given $LL(\texttt{powers}, \texttt{community}) = 83.42$ and $t_1 = 15.0$, the word `community` was one of the highly associated words for the word `powers`, where $LL(\texttt{powers}, \texttt{community}) > t_1$.

Furthermore, if the word pair (`community`, *`comunidad`*) could be found in the initial seed lexicon, then the word `community` would be used as R_S and the

word *comunidad* would be used as $tr(R_S)$.

Let $CT(W)$ denotes the set of context terms of a word W. To get the in-domain terms, all identical context terms co-occurring with both s and R_S need to be identified. The in-domain terms of s given the context terms R_S are given by:

$$ID(s, R_S) = CT(s) \cap CT(R_S)$$

The English words `Programme` and `public` were the examples of in-domain terms of the word `powers` given the word `community` as its highly associated term. Likewise, similar procedures would be performed to obtain all context terms co-occurring with both t and $tr(R_S)$. Among the in-domain terms of $ID($*poderes*, *comunidad*$)$ includes the words *programa* (`programme`) and *publico* (`public`).

Note that the $ID(s, R_S)$ is a context vector in the source language and the $ID(t, tr(R_S))$ is a context vector in the target language. The $ID(s, R_S)$ refers to the in-domain context vector of s with respect to R_S. $tr(s|R_S)$ is used to denote the translation proposed for s given the highly associated term R_S. We compute $tr(s|R_S)$ using:

$$tr(s|R_S) = \underset{t}{\operatorname{argmax}}\, sim(ID(s, R_S), ID(t, tr(R_S)))$$

Based on the in-domain terms technique, learning translation pairs are conditioned on the highly associated words (R_S). Table 5.1 provides a sample of the English-Spanish lexicon learnt for the word `power` with different R_S.

In the next section, we introduce a similarity measure that operates on the context vectors of the source and target languages without requiring a seed lexicon.

English		Spanish		Sim
s	R_S	$tr(R_S)$	t	
powers	community	comunidad	**competencias**	**0.9876**
			poderes	0.9744
			independiente	0.9501
	democracy	democracia	**competencias**	**0.9948**
			poderes	0.9915
			independiente	0.9483
	independence	independencia	**competencias**	**0.9939**
			poderes	0.9745
			independiente	0.9633
	justice	justicia	**poderes**	**0.9922**
			competencias	0.3450
			independiente	0.9296
	delegation	delegacion	**poderes**	**0.9568**
			competencias	0.9266
			independiente	0.8408

Table 5.1: A sample of translation equivalents learnt for *powers*

5.3 Rank-binning similarity measure

Most existing methods for computing similarity cannot be directly employed to measure the similarity between the in-domain term context vectors of different languages. A bilingual dictionary can be assumed, but that greatly diminishes the practicality of the in-domain terms technique.

We propose a new measure based on an assumption as follows: the relative distributions of in-domain context terms of translation equivalent pairs are roughly comparable in the source language and in the target language. Figure 5.3 depicts an example of the LLR values of the in-domain terms for the translation pair `agreement-`*`acuerdo`* (conditioned on the highly associated term `association-`*`assosciacion`*). The example shows that the distributions of the

in-domain terms are comparable, although not necessary identical. Thus, the comparable distributions could be used as a base for a new similarity measure to compute the similarity among the in-domain term vectors.

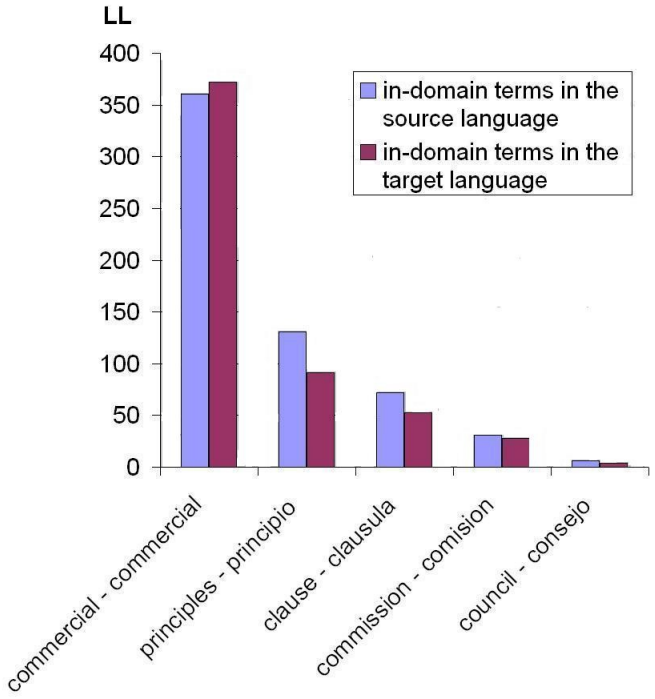

Figure 5.3: Similar distribution of in-domain terms for `agreement` with `association` and *acuerdo* with *asociacion*

Rank-binnings or rank histograms are usually used as a diagnostic tool to evaluate the spread of an ensemble rather than as a verification method. Wong (2009) have used the method of rank-binning to roughly examine the performance of a system on learning lightweight ontologies. In this study, a similar method is proposed to measure the similarity of word pairs. This method is based on transformed rank values of context terms that would be used to set the parameters of bins, however, only in-domain terms with transformed rank values of certain range could be used. This method builds two comparable

sets of bins using similar procedures, and later, both sets can be compared to one another based on the number of in-domain terms resides in each bin of specific location of each set.

$CT(\mathtt{s = powers})$			
Context term	LL	$rank$	z_k
european	491.33	1	0.00000
legislative	482.19	2	0.00406
parliament	408.26	3	0.00813
:	:	:	:
:	:	:	:
:	:	:	:
public	16.96	245	0.99186
programme	15.40	246	0.99593
representatives	15.32	247	1.00000
$n = 247$			

Table 5.2: Some examples of transformed values of each term in $CT(\mathtt{powers})$

Pre-processing step:

1. Let s be a word in the source language and $ct_1, ct_2, ..., ct_n$ be the set of n context terms ranked in descending LLR values of s (see Table 5.2).

2. We transform the rank values of context terms ct_k into the range [0,1] using:

$$z_k = \frac{rank(ct_k) - 1}{n - 1}$$

Binning procedure

The interval $[0, 1]$ is divided into g bins of equal length. Let $b_1, \ldots, b_g$ denote the g bins. Then, the in-domain terms vector $ID(s, R_S)$ is mapped into the binned vector $b_1, \ldots, b_g$. For each $ct_k \in ID(s, R_S)$, this mapping is done by

using the corresponding z_k from the pre-processing step. For each bin, the number of different in-domain terms that are mapped into this bin is counted. For example, the range of the first bin b_1 was [0, 0.009] and the $ID(s, R_S)$ consisted of the words *parliament, councils* and *affairs*, but the word *parliament* was the only term mapped into b_1 (i.e., $b_1 = 1$).

The bins are then normalised by dividing their counts with $\mid ID(s, R_S) \mid$. Likewise, a similar pre-processing step and binning procedure are repeated on the target word side. In the previous example, the $ID(t, tr(R_S))$ contained the words *parlamento, consejo* and *asuntos*.

Rank binning similarity

We used the Euclidean distance to compute similarity between bins. Given bins $P = p_1, \ldots, p_g$ and $Q = q_1, \ldots, q_g$, the Euclidean distance is given by:

$$dist(P, Q) = \sqrt{\sum_{i=1}^{g} (p_i, q_i)^2}$$

5.4 Experimental setups

This section describes the experimental setups used in this study. Most of the settings have been described in the previous chapters, including the seed lexicon, data, pre-processing tasks and reference lexicons (see Chapter 3 and 4 for details).

Evaluation

In the experiments, the extraction task for bilingual English-Spanish lexicon considered about 2000 high frequency source words and 2000 high frequency target words. Only individual words with at least a hundred highly associated context terms that were chosen to be part of the initial seed lexicon. Different highly associated $tr(R_S)$ terms for a given t could produce similar (s, t) pairs.

In this case, only one of the (s, t) pairs was considered. We suggest that the remaining word pairs should be kept for word sense discrimination purposes in future work. In addition, only proposed translation pairs whose similarity values were above certain threshold were considered in this study.

Similar to other experiments we have conducted, F_1 score, the recall and the precision were again used to evaluate the proposed lexicon against the evaluation lexicon. When either s or t in the proposed translation pairs was not found in the evaluation lexicon, the translation pairs were considered as unknown word pairs. We would not include these unknown translation pairs, although the translation pairs were correct.

Baselines

The main baseline system that was used in this evaluation was $CB+700+Cos$ (see Chapter 3 for details). However, the other two baseline systems, namely $CB + 160 + Cos$ and $CB + 100 + Cos$, were also included as additions in this study, but in this chapter, we have used different notations to denote these models.

5.5 Evaluation results

In the experiments, the effects of using in-domain context terms to system performance were observed. The potential of rank-binning similarity measure was also examined.

5.5.1 From standard context vector to in-domain context vector

Most research in bilingual lexicon extraction so far has employed the standard context vector approach. In this study, in order to explore the potential of the in-domain context vectors, the systems based on the in-domain approach were compared against the baseline systems (in which, the latter was based on the standard context-based approach). Different sets of seed lexicon were

employed in each system for comparison purposes.

To simplify the explanation, the baseline systems are now denoted according to the sizes of the seed lexicon used in their context similarity measure, namely $CV+100$ for using Lex_{100}, $CV+160$ for using Lex_{160} and $CV+700$ for using Lex_{700}. These lexicon sizes were also used to distinguish the different variants of the in-domain term (IDT) models, namely $IDT+CV+100$ for using Lex_{100}, $IDT+CV160$ for using Lex_{160} and $IDT+CV+700$ for using Lex_{700}.

Based on the $CV+700$, the system achieved more than 50% of the best F_1 score. Using the same seed lexicon, the best F_1 score increased about 20 % when the system was based on the $IDT+CV+700$ model. However, another system based on the $IDT+CV+100$ recorded a score 15% higher than a system based on the $CV+100$ (i.e., 80.9% and 66.4%, respectively). Using an automatically derived seed lexicon and based on the $IDT+CV+160$ model, a system yielded 70 % of best F_1 score compared to 62.4% when the $CV+160$ model was applied. Table 5.3 shows the results of various precision scores p_x at recall values x.

Model	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$BestF_1score$
CV+700	58.3	61.2	64.8	55.2	52.6
CV+100	52.0	53.0	47.2	44.8	66.4
CV+160	68.5	56.8	48.8	48.8	62.4
IDT+CV+700	83.3	90.2	82.0	66.7	73.1
IDT+CV+100	80.0	75.8	66.7	69.4	80.9
IDT+CV+160	90.0	80.6	73.9	69.2	70.0

Table 5.3: Performance of basic context-based vs. IDT models in different settings

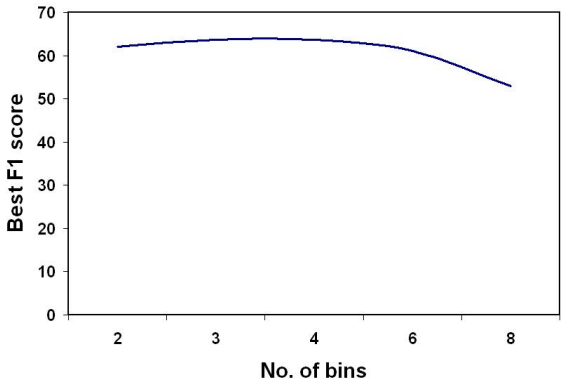

Figure 5.4: Performance of *IDT+RB+160* with different numbers of bins

5.5.2 Similarity measure using rank-binning

RB denotes a model based on the rank-binning approach. Running *RB* means that no seed lexicon would be involved in the similarity measure. Likewise, the similarity measure is also used in the *IDT*, namely $ID + RB + 160$ that employs Lex_{160}.

Several tests run using $IDT + RB + 160$ with different numbers of bins were performed. Figure 5.4 shows the overall results. A system based on the $IDT + RB + 160$ model yielded 63.7% of best F_1 score with four bins. However, the F_1 score starts to drop from 61.1% to 53% when six or eight bins were used. With three and two bins, the system based on the $IDT + RB + 160$ model yielded 63.7% and 62.0% of best F_1 scores, respectively. Nonetheless, using one bin was not be possible because all the values would fall into one bin. Thus, the rank-binning similarity measure for the rest of the experiments, where ever *RB* is mentioned, would refers to a four bins setting.

While all the systems that used the standard context similarity measure yielded scores higher than 50.0%, the *RB* alone only achieved 39.2 % of performance score. However, the advantages of the *RB* model were its ability to perform

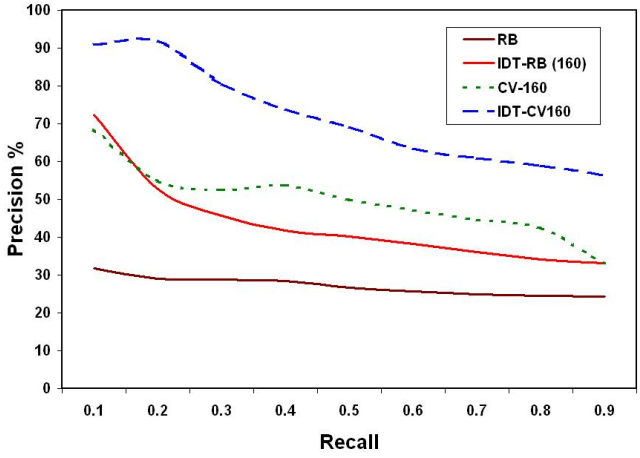

Figure 5.5: Performance of different unsupervised models

the matching without involving the initial lexicon and orthographic features. In addition, the system scored higher when the similarity measure was used in the IDT (i.e., $IDT + RB + 160$). Note that Lex_{160} was derived automatically, hence, the approach could also be considered as an unsupervised learning. This system's performance was slightly lower compared to the conventional system based on the $CV + 160$ model, however, another system based on the $IDT + CV + 160$ model has outperformed both systems. (See Figure 5.5 for details).

Overall, systems based on the in-domain terms yielded higher F_1 scores compared to the conventional context vector approach.

5.5.3 Comparison with a CCA-based model

In general, previous studies focusing on extracting bilingual lexicons from comparable corpora employed the conventional context vector approach. Haghighi et al. (2008) focused on applying CCA to improve the method. Their approach were quite similar to the approach taken in this study in terms of comparable corpora they used (which were the first 50,000 sentences taken from the

English Europarl, and the second 50,000 sentences taken from the Spanish Europarl), however, they used different initial seed lexicons. We replicated their model loosely, as described previously in Chapter 3. Within a similar setting, the system based on the CCA model yielded 57.5% of the best F_1 score.

5.6 Discussion

5.6.1 Potential of in-domain term approach

The experiments in this study have clearly demonstrated the use of the in-domain terms, which have allowed the systems to achieve higher performance scores compared to the systems based on the conventional methods. In addition, the former performed better than the systems based on the dimension reduction methods.

From observation, the number of incorrect translation pairs was further reduced when the context term lists were initially filtered. Nevertheless, this approach would be depending on the initial bilingual lexicon in order for the approach to work effectively in translation highly associated context terms into the source language. Table 5.4 shows some examples of the most confidence translation pairs proposed by the $IDT + CV + 100$ system.

We tested all the incorrect English-Spanish pairs that were previously presented in Chapter 3, which were possibly, of word hypernymy. Surprisingly, the system based on the in-domain terms was able to fix six of the incorrect word pairs by proposing the correct translations for the source words. For the remaining word pairs, for example, the English word `banks` was now incorrectly matched to the Spanish word `banco` (*bank*), and the English word `tourism` was still matched to the Spanish word *economico* (`economy`). The former might have been caused by the approach itself that treat each noun in its singular form differently to its plural form. This problem could be solved by using a lemmatizer or a stemmer, however, it might cause another problem

especially if inflected or agglutinative languages are involved.

English	Spanish	Sim score	Correct?
principle	principio	0.9999	Yes
government	estado	0.9999	No
government	gobierno	0.9999	Yes
resources	recursos	0.9999	Yes
difficult	dificil	0.9999	Yes
sector	competencia	0.9998	No
sector	sector	0.9998	Yes
programme	programa	0.9998	Yes
programme	comunidad	0.9998	No
agreement	acuerdo	0.9998	Yes

Table 5.4: Some examples of most confident translation pairs proposed by $IDT + CV + 100$ and ranked by their similarity scores

5.6.2 Similarity measure alternative for unrelated language pairs

One should have realized by now that the relationships between the language pairs of the respective monolingual corpora might largely affect the results. Thus, for systems involving unrelated language pairs, the rank-binning similarity measure might be a good alternative to be implemented.

5.6.3 Word sense discrimination ability

As mentioned in Section 5.3, each source word might have more than one highly associated context term, R_S. Different R_S might suggest different target words for the same source word. For example, given the source word `powers` and the highly associated word `community`, the word *competencias*was proposed as the best translations. On the other hand, for the same source

word `powers`, the target word *`poderes`* was suggested instead of the word *`competencias`* when the highly associated word was the word `delegation`.

5.6.4 Evaluation issue

This study focuses on a technique that could improve the sets of context terms used in the extraction process. It would be more interesting if the experimental findings of this study could be compared to more current, related work, such as found in Prochasson et al. (2009) and Andrade et al. (2010). Unfortunately, the goals and the settings they used were very different to this study. Both Prochasson et al. (2009) and Andrade et al. (2010)'s studies assume the availability of large amount resources. Moreover, the system discussed in Prochasson et al. (2009) required some specialised vocabularies in their setting, whilst, Andrade et al. (2010) took 100 noun pairs of technical terms into account; thus, we were not being able to replicate their models in this study. That was the reason why we only used the baseline systems for comparison during the evaluation. We highlight this issue in the final chapter of this thesis.

5.7 Conclusion

This chapter introduced the in-domain terms technique and discussed the experiments required in this study. The systems based on the in-domain technique performed quite well although without the availability of initial bilingual lexicons. Furthermore, this study might have revealed the potential of building word sense disambiguated lexicons.

The experimental findings of this study have also suggested the imperative of context terms that could help determine correct translation pairs, thus, the context terms should be selected carefully. In the next chapter, a different technique is discussed, in which the word vectors would include term elements in form of a single word and multi-words, and also the use of the web to search for additional data.

Chapter 6

Employing Data from the Web

In Chapter 3, the author described earlier work with regards to corpus acquisition task. This chapter discusses a method that acquires very small comparable corpora from the web, exploits the corpora to harvest more data (also from the web) and, eventually, extracts bilingual word pairs. More interestingly, the use of the web is not only limited to obtaining data but also to verifying the context terms at multi-word level. Surprisingly, the technique is able to eliminate most irrelevant and weak-relevant context terms, generating higher precision word pairs compared to using standard context-based system within similar setting.

6.1 Introduction

The question of building high precision bilingual lexicons for unrelated language pairs, especially in extreme settings, remains quite elusive to many practitioners. Hence, the type of research in this area is not eagerly pursued thus far; nonetheless, there is a pressing need for such learning methods to further improve existing systems. As described previously, we suggest improving the context term lists in order to improve the results of a bilingual lexicon extraction system. Therefore, the need to address an effective way to derive a good set of context terms from very limited resources becomes more urgent.

In this regard, investigating on how far information can be exploited under extreme settings would reveal many important insights, which could be used to develop robust systems than the current ones.

In this chapter, instead of using a single word per context term, we propose a technique that extends the context features to the multi-word level. Furthermore, in this study, all the experiments were conducted under-resourced situation (i.e., having limited capacity of languages). Primarily, the robustness of the under-resourced systems could be determined to highlight a worst case scenario, which is inevitable in today's demanding requirements. In other words, such extreme settings would stretch the systems' capability to the maximum, providing important new finding–vis-a-vis their normal capability under normal settings. To test the method, we utilized available data from the web given the limited English-Malay documents.

In this study, online news reports of an international football event and the context window of a sentence to get the context co-occurrence data were deployed. For instance, the English word `coach` and a sentence about Diego Maradona, the coach of Argentina football team during the World Cup 2010, were tested. Naturally, there were also other individuals sharing the same first name (i.e., Diego) in the event, such as Diego Forlan and Diego Lugano, who were members of the Uruguay team.

For terms `Diego` and `Maradona`, each may closely relate with the word `coach`. For example, one of the sentences that relate the words is as follows:

Argentina **coach Diego Maradona**, *whose touchline performance in the 1-0 win over Nigeria was better than the football played by most teams, is not too worried by the slow start.*

Similarly, the word `Forlan` also being mentioned together with the word

coach in the same sentence; though, these word are not directly related because Forlan was not the coach of the Uruguay team. For example,

"For the Uruguayan team **Diego Forlan** *is a very important player," said* **coach** *Oscar Tabarez.*

Table 6.1: A flat co-occurrence matrix for the standard approach

Context Term	**s = *coach***
Argentina	1
Diego	2
Maradona	1
Uruguay	0
Forlan	1

Table 6.1 shows the co-occurrence matrix using co-occurrence counts for the example using these two sentences. This matrix is called the flat co-occurrence matrix where a high count indicates the relations among the words are strong. In the matrix, the term Diego co-occurs much higher with the word coach than other words. Unfortunately, the word Diego is not a good discriminative context term for word coach as it could also co-occur highly with the word player. Thus, the English word coach may have been matched with its non-equivalent Malay words that occur in the same context, such as *pemain* (player), *pasukan* (team), or *prestasi* (performance), more than the correct translation equivalent, i.e., *jurulatih* (coach). Many researchers have attempted to eliminate words that occur highly in all documents, which are known as domain stop words. However, we believe that such words could still be useful by using novel, innovative techniques. Hence, for the model used of this study, each context term would be associated with one or more context terms to help emphasise the context it represents.

In addition to the above challenge, extraction tasks would be difficult to proceed when the needed resources scarce. For example, suppose that (Argentina,

Table 6.2: A depth co-occurrence matrix for the m-word level context feature approach, $m = 3$

Context Term	**s = *coach***
Diego	2
Argentina	1
Maradona	1
Diego, Argentina	1
Diego, Maradona	1
Argentina, Maradona	1
Diego, Argentina, Maradona	1
Uruguay	0
Forlan	1
Diego, Uruguay	0
Diego, Forlan	1
Uruguay, Forlan	0
Diego, Uruguay, Forlan	0

(`Argentina`, `Argentina`), (`Diego`, `Diego`), (`Forlan`, `Forlan`), (`Maradona`, `Maradona`), and (`Uruguay`, `Uruguay`) are the only entries that occur in the English-Malay initial bilingual lexicon. For the English word `coach`, a vector that represents the word `coach` in the English word space is represented as a linear combination of the basis terms of `Argentina`, `Diego`, `Forlan`, `Maradona`, and `Uruguay`. In the standard approach the terms are treated as individuals, which are `<Argentina>`, `<Diego>`, `<Forlan>`, `<Maradona>`, and `<Uruguay>`, respectively; and these terms are independent of each other in the context vector.

We propose a technique based on an assumption as follows: *two or more terms that are featured together as one may bring more meaning than a single term.* The terms used here refer to the context words co-occurring closely to one another within the same window for a word. In the model, the terms do not have to be a phrase or to occur next to one another.

Two terms that are treated together as one "multi-word" context term, such as `<Diego, Maradona>`, are expected to bring in more meaning based on the same word `coach`. On the other hand, `<Forlan, Argentina>` are less meaningful because they are not related, unless there was a report about any matches between Argentina and Uruguay teams. Otherwise, `<Forlan, Argentina>` should not be considered in the context term list. Arguably, more meanings would be rendered when the words `<Argentina, Diego, Maradona>` are also defined as the multi-word set of context terms.

Table 6.2 shows the co-occurrence matrix constructed based on the above approach. The matrix was divided into three groups in rows, with the top row containing a general word in both context and the other two rows representing a different meaning for the word. Simply, summing up all the non-zero occurrence counts separately for each row would show the word `coach` to be more related to the set of context terms in the middle row of the matrix.

Defining the multi-word context terms

Typically, the dimensions of each word vector, which are also known as basis term vectors, are defined by individual-word or single-word context features, comprising uni-grams that co-occur within a certain window around the word they represent, either for the source word or the target word. Extracting multi-word context terms would entail another fixed, small-sized windows that are to be used within the original context windows.

Figure 6.1 shows a sample of non-parallel English-Malay texts for the English word `coach` and its translation equivalent `jurulatih` in the sport domain. In the example, the block of lines shows the fixed boundaries that could be used to draw the multi-word context terms for the source word and the target word. As an example, for the block containing the word string "*Argentina coach Diego Maradona*", one of the multi-word context term vectors that was extracted is as follows:

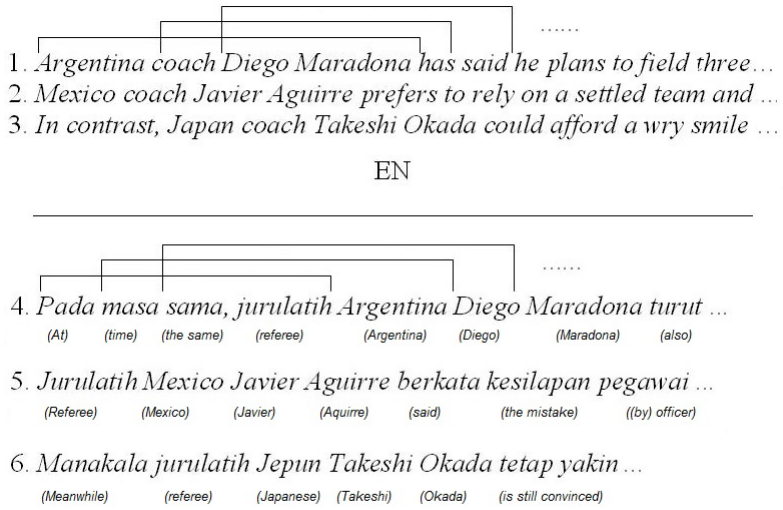

Figure 6.1: Sample of non-parallel English (EN) and Malay (MA) texts from comparable corpora. The EN contains the source word `coach` while the MA contains the target word that is equivalent to the source word, i.e., *jurulatih*. The block of lines in the first EN sentence showing some examples of 4-grams that could be drawn from the sentence.

<Argentina,Diego,Maradona>

Figure 6.2 shows part of the multi-word context terms for the source word coach and some of the multi-word context terms for its potential translation equivalent in Malay, i.e., *jurulatih*. The co-occurrence frequencies were obtained from very small comparable corpora.

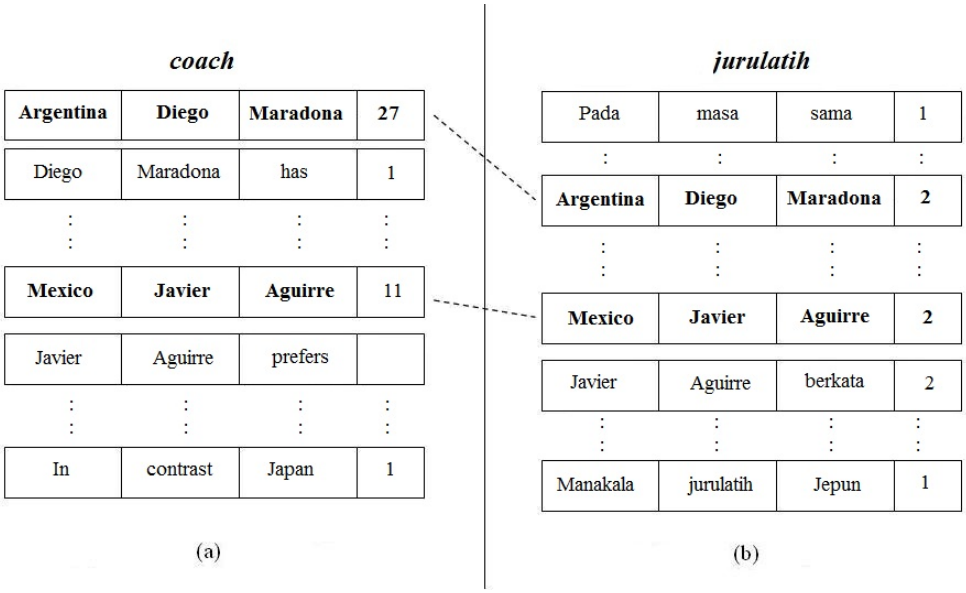

Figure 6.2: Some examples of the multi-word context terms for the source word coach and the target word *jurulatih*, deriving from (a) an English corpus, and (b) a Malay corpus

As has been observed, when a source word co-occurs highly with the multi-word context terms, its corresponding word in the target language also co-occurs highly with the multi-word feature correspondence. In addition, both words would share certain common multi-word features, as has been observed in the above example, where all the words between a multi-word feature pair were identical due to the names of the persons contained therein.

In the example, the co-occurrences of an identical multi-word context term <Argentina, Diego, Maradona> were observed on both sides of the texts.

A similar co-occurrences were also observed for the term `<Mexico, Javier, Aquirre>` in both languages. These multi-word context terms are the examples of features that could be used to map the source word to its corresponding target word. In order to construct the multi-word context terms in fixed, small word windows within a larger word window, we used the n-grams.

Using n-gram sized windows within a context window

Solving NLP problems using n-grams is not new as this method has important properties that allow multi-word units to remain together in a sequence. For example, with n-grams, specialized terminologies like *Lyme Disease* and *Carpal Tunnel Syndrome* can be identified easily if they occur in the corpora. Some examples of earlier work using n-grams are Haruno et al. (1996), who learnt bilingual collocations by finding similar word chunks of n-grams, and Yamamoto et al. (2001), who used various n-gram models to generate multi-word translation units for bilingual lexicon extraction. Essentially, they used the n-grams to extract multi-word correspondences from parallel corpora. In addition, both of these earlier work did not use the context vector. Hence, these previous work, apart from the use of n-grams, share no other similarity with the approach taken in this study.

In this study, the approach used the n-gram as a fixed, small-sized window to help capture a group of words co-occurring closely with one another within a larger context window of a test word. Nonetheless, word order can be different between non-related language pairs. Therefore, we did not treat the returned n-grams as n-grams, instead the content of n-gram was transformed into a single term vector. As such, the multi-word context feature is defined as a multi-word vector, each containing a set of context words that co-occur together in a small, fixed window within a context window of a test word. Through the use of these vectors, the bag-of-word concept was realized in this work.

In the next section, an approach to acquire very small comparable corpora automatically from the web is introduced. Though being very small in capacity, the corpora could be generated by fetching available data from the web.

6.2 Acquiring very small comparable corpora from the web

Typically, comparable corpora acquired from the web come from several sources, namely database collections, consisting of news reports, medical journals or government documents. In this study, the method used a slightly different corpora than the typical comparable corpora as the former was obtained mainly from daily news reports. In essence, the approach employed is similar to the approach based on RSS (Fry, 2005) and BootCat (Baroni and Bernadini, 2004); but, the former approach provides some control allows specific collection of data from news agencies relating to significant events.

6.2.1 Methodology

In this study, the method proposed comprises several steps to be followed in a sequence. These steps are described as follows:

1. *Identifying a specific, significant event*
   A specific, significant international event that was widely covered by the media was selected. Essentially, this type of important events are usually planned earlier by its organiser prior to its launching to ensure a successful outcome. The event may be held annually or once in every four years, such as national election campaigns and international sporting events (e.g., the Olympics, the Paralympics and the World Cup Series).

2. *Selecting relevant sources before the event*
   Several news agencies that provide the press coverage of the event were identified before the event takes place. The number of news agencies is not restricted and, more importantly, these news agencies must represent the languages required for the translation. Commonly, these news

agencies provide specific URLs for the proposed event, which could be assessed to check news update. Figure 6.3 shows two examples of the URLs. Once these paths have been identified, they will be kept in a list.

3. *Opening and accessing the relevant pages automatically*
   A web crawler and DnldURL tools were used to collect the relevant pages automatically. The data collection was conducted regularly on a daily basis until the closing end of the event (see 6.2.1.1 for details).

4. *Matching documents to find similar documents across languages*
   Many earlier studies, such as Resnik and Smith (2003) Chen et al. (2004), assumed some specific naming conventions of filenames or URLs to find parallel web documents. In addition, some of these studies also relied on matching similar articles according to the similar features that they shared (Patry and Langlais, 2011). More importantly, Patry and Langlais suggest using three types of entities in a document to get a set of sequential data, i.e., numerical entities, hapax words and punctuation marks. Moreover, document pairs are matched based on the proportion of the entities shared across the documents.

   In this study, a simple overlap measure using features such as the number of identical word, page length, and date range was used. The matching was done among document pairs within the same language and across the languages. The goal was to find a group of similar articles (which is most likely provided by different news agencies) in the same paired languages (i.e., in both English and Malay).

The following sub section describes some of the methods in further details.

6.2.1.1 Accessing web pages automatically

To access the relevant pages, an automated procedure was used on a list containing the parent web pages, URL_0. Thus, two important steps were carried out in a sequence: (a) first, each URL was queried regarding its contents and size using the $url_{0,i}$, where $url_{0,i} \in URL_0$, $1 < i < p$ and p is the size of URL_0,

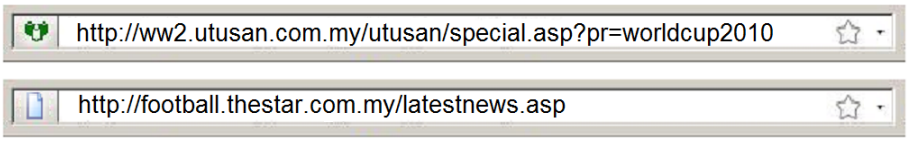

Figure 6.3: Two examples of the URLs provided by the Malaysian local news agencies, i.e., Utusan Malaysia (Malay) and The Star (English), reporting on the World Cup events in 2010

notation through a web crawler level 1. This process returned a list of child web pages, URL_1; (b) next, each URL site, which was identified by the URL_1, where $url_{1,j} \in URL_1$, $1 < j < q$ and q is the size of URL_1, was downloaded (using a download tool) and opened accordingly. In this study, both tools used in performing the two steps are freely available from a website, which can be retrieved from `http://www.devdaily.com/java/edu/pj/pj010011/pj010011.shtml`.

From the opened page, lines containing the html tag $< p >$ were extracted. Using this approach, cleaned texts were obtained by the removal of other HTML tags used in the web. Before saving the texts, a language filter was employed to identify the language of the written article. The filter utilizes a simple procedure that recognizes each language according to a set of most frequent words in the stop list. After the filtering process, the remaining unwanted HTML tags were removed from each line and the texts were saved separately in a specific folder of a directory according to the following order:

$$\ldots \backslash \text{DATE} \backslash \text{LANG} \backslash \text{AGENCY}$$

where

$DATE$ refers to the date the download takes place,
$LANG$ refers to the language used in the articles, and
$AGENCY$ refers to the specific news agency.

By using such an approach, the information pertaining to each text was indirectly preserved. For example, the directory for a folder containing the English

articles downloaded on the first of January, 2012 from The Star is represented by:

... \20120101 \En \ TheStar

This whole process was repeated daily until the event ended.

6.2.1.2 Matching similar documents

In this study, news reports in different languages that were published on the same date and shared many identical words were assumed to be article pairs. To find similar articles, we first obtained word corpus lists for both corpora and sought identical spelling word pairs based on several features, namely name of a person, name of a country, and others. On the other hand, finding cognate pairs were not easily carried out for unrelated language pairs.

In this matching process, a document vector was created for each article using the word pairs. These vectors were then used to match all document within the same date across languages. Finally, unmatched articles were removed and similar articles were saved in two folders according to the two languages involved. These two folders could be readily accessed to retrieve the required articles.

6.3 Acquiring more data from the web

The comparable corpora obtained from the previous step were too small; hence, they were likely to be unreliable for a bilingual lexicon extraction task (see the discussion on the effects in Section 6.6). In view of the pressing need for extra data, this study gathered more data from the web. The small comparable corpora were used as input to the process described in the following sub-sections.

6.3.1 Methodology

1. *Pre-processing*
   Sentence boundary detection and tokenizing were performed on the comparable corpora, which were obtained from the previous step, to divide the texts into smaller units. Stop words were removed from the texts to acquire the content-bearing words.

2. *Selecting the Source Word and the Target Word Vocabularies*
   To acquire the source word vocabulary list, all vocabularies in the text were first sorted according to their frequencies; and then, only words with medium frequency were selected to be in the source word vocabulary list. Likewise, the same procedure was repeated to obtain the target word vocabulary list. Altogether, both lists may include any word type.

3. *Generating the Multi-word Context Term*
   For each source word obtained from the corpora:

   (a) Sentences containing the word were extracted, thus deriving from each sentence the n-grams. Multi-word context terms were then collected from each block of strings provided by the n-gram.

   (b) A source word vector was generated from the multi-word context terms. Each basis term vector represented a weighted value for a multi-word context term in the source language of normalized co-occurrence counts, provided by a window of a sentence.

   (c) Sets of multi-word context terms were also generated, containing at least three different context terms. Each set was used as another query for the search engine. Again, using the similar step (i.e., Step 3) as described in Section 6.2, the output of the search showed a series of relevant pages. The sentences in the relevant pages were used to update the term vectors accordingly.

4. *Finding the Translation Pairs*
   Step 3 was repeated for each target language by directly matching each

source word vector to all target word vector using the cosine measure. The proposed translation pairs were then sorted in the descending order. For each English source word, the highly ranked Malay target words, for which the similarity values must not be below a threshold t_1, was treated as the most confident (i.e., reliable) translation.

In general, the only step we introduced in the basic method is the third step. The following sections provide the details of the ensuing process.

6.3.2 Learning multi-word context terms

In the standard approach, single-word context terms in the source language are collected within a certain word-window size surroundings a source word. In contrast, the model we propose takes multi-word context terms into account.

To simplify the multi-word context term collection procedure, we utilized n-grams to break each sentences into smaller units, with each containing at least three to six words. Before the terms can be used in the basis term vector together with the single-word context terms, two important steps should be performed as follows:

6.3.2.1 The n-gram extraction

A sequence of units, or block of strings containing n-words was drawn from a window size of a sentence using n-gram. Given a sentence containing the source word s, all the n-grams (where $3 \leq n \leq 6$) occurring in the sentence were located. In the previous example, the source word s was `coach`, and one of the sentences that contained the word `coach` is as follows:

Argentina coach Diego Maradona has said he plans to field three forwards against Nigeria.

If $n = 3$, examples of 3-gram that could be derived from the above sentence are "Argentina coach Diego", "coach Diego Maradona", "Diego Maradona has" and so forth.

6.3.2.2 Multi-word context term extraction

A set of n-grams was obtained from the previous step using a sentence that contained the source word s. Later, the n-grams were converted into the bag-of-word concepts. From here, the remaining words were no longer treated as a sequence of strings, but as a group of individual words. The source word s was removed from each n-gram (if any), and the remaining words in the group were sorted according to the alphabetical order to form a multi-word context term for s. Only unique context terms were considered such that any duplicate was removed.

Subsequently, each co-occurrence of the multi-word context term was counted, normalized and stored for the source word s in a simple co-occurrence metric as suggested by Rapp (1999). The process was repeated and the collections were updated for all sentences containing the source word s. Likewise, similar procedure was performed for each source word s and target word t.

6.3.3 Querying the search engine

A source word was accompanied by one set of multi-word context terms at a time and used as a query to the search engine. All n-grams, where 3 <n <6, containing the source word, were derived automatically from the returned documents. From these n-grams, the extracted multi-word context terms were used to update the existing context term list for that particular source word. A similar operation was done to all the source words and the target words. As a caution, each of the context words in the features must be part of the initial bilingual lexicon entries. Hence, a similar set of queries was used to obtain documents in the target language, except that the source word was replaced with each target word from the target word list, one at a time.

Later on, the query was submitted to the search engine, whereby producing a long list of hits (i.e., results). Using this list, the first 1,000 web documents were retrieved; this step was then followed by the automatic extraction of all

sentences containing the source word. Hence, for each source word, different numbers of sentences may be collected.

Using the extracted sentences, all context words were derived, and data were added to the existing context term list. Similarly, the same process was repeated for the target words. Thus, using this new added data, each source word vector and target word vector were matched accordingly.

6.4 Experimental setups

This section describes the experimental setups used in the study. Some of the settings have been described in the previous chapter, thus requiring no detailed description.

- *Data*
  We compiled the *World Cup 2010* online news articles from June 11^{th}, 2010 to July 11^{th}, 2010 to form the English-Malay small, comparable corpora. The compilation process involved several major online newspaper in Malaysia such as *Berita Harian*, *Utusan Malaysia*, *The New Straits Times* and *The Stars*. In addition, several news articles available from the *FIFA* official website were also collected.

  Using these data, very small English-Malay comparable texts were established, each containing 2,287 English and 1,304 Malay articles. We called the collection as $MyWC$.

- *The source word and the target word vocabulary list*
  The raw English corpus contained 27,642 unique English words. However, high frequency words tend to be noisy; thus we sorted the corpus words and removed the first 25 words from the frequency list. Additionally, words that occurred more than 15 times in the corpus (i.e., about 1500 words) were considered. From the list, the words were filtered to remove stop words that occurred in the initial bilingual lexicon entries and

also words of length less than four. The same procedure was repeated to obtain the target word list from the raw Malay corpus containing 9,810 unique Malay words.

- *Initial bilingual lexicons*
  The list contained many identical names and places for both languages, such as `Diego`, `Maradona` and `Argentina`; hence, these bilingual word pairs provided several advantages. To capitalize on the word pairs, the methods recommended by Koehn and Knight (2001) were used to obtain the initial bilingual lexicon. Furthermore, only word pairs with identical spelling were considered, which entailed the use of *string edit distance* with the zero distance. To perform this compilation, each word had to be longer than four characters.

- *Reference lexicon*
  For the evaluation of the lexicon, the English-Malay word pairs from the *Websters dictionary* were extracted. These word pairs could be retrieved from
  `http://www.websters-online-dictionary.org`.
  In addition, for English words having multiple Malay translations, only those correspondences that occur in the Malay corpus were considered in establishing the reference lexicon.

- *Evaluation*
  In the experiments carried out, the focus was to build the Malay-English lexicon of word-to-word correspondences. In particular, the proposed lexicon was evaluated against the evaluation lexicon. To compare the performances of the various methods used, the precision, the recall and the F_1 measures were applied in the experiments.

- *Baseline method*
  For comparison purposes, the standard approach was implemented in a similar setting. Generally, this approach is called the single-word level

context features approach, $SWCF$. Based on this approach, the single-word features were collected from the context words that co-occurred with the source word in the source corpus within a window of a sentence. The same collection procedure was repeated in dealing with the target word in the target corpus.

6.5 Evaluation results

This section reports the results of the experiments carried out in the study, namely the various precision values p_x at the recall values x and the F_1 scores of the systems based on the different models. Table 6.6 shows the different performances of the models used in the experiments.

6.5.1 Single word context feature vs. multi-word context feature

A given source word s and a set of context terms that co-occurred with the source word in a sentence were combined to derive a long query separated by the "+" symbol for the search engine. For example, this combination is written as follows:

"coach"+ "Diego"+"Maradona" + ... + "Argentina"

However, this approach might have introduced defective data to the context-based system, as evidenced from its poor performance denoted by $SWCF + Web$ in Table 6.3. More seriously, only 2% of the F_1 score were recorded, indicating the approach's underperformance. Hence, this approach lacked the required efficacy given the collection of data from the web that was less stringent. In inevitably, spurious data (i.e., noise) might have been gathered throughout the entire collection process. The $MWCF + Web$ model denotes the multiword approach. Based on this model, the system achieved almost 30% of performance score as shown in Table 6.3.

Setting	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$F_1 score$
$SWCF + Web$	4.74	3.79	2.87	1.89	1.87
$MWCF + Web$	33.00	28.57	20.00	20.00	28.57

Table 6.3: Performance of SWCF vs. MWCF using web data

6.5.2 *n-word* feature

Previously, two different approaches were used to acquire more data from the web. However, comparing their performance would be inadequate, at best, and misleading, at worst; the reason is that both approaches were implemented quite liberally. Hence, another experiment was conducted, where this time the number of context terms sent to the query was restricted by n-word, where 3 <n <6. The difference between this approach and the conventional n-gram approach is that the former may combine any context words as long as they co-occurred in a sentence, whereas, the latter approach only considered context words located near to each other within an n-gram window.

From the 1,000 returned documents,this model managed to locate sentences containing the source word and, at least, one of the context terms from the query. Through this approach, the results improved slightly by over 5% (see the performance score of the model denoted by $SWCF + Web + nword$ in Table 6.4).

Setting	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$F_1 score$
$SWCF + Web + nword$	9.48	7.56	6.55	4.86	7.79
$MWCF + Web + ngram$	33.00	28.57	20.00	20.00	28.57

Table 6.4: Performance of different models using n-words and n-grams

6.5.3 Top 1 evaluation

Table **??** presents some examples of the translation pairs learned by $MWCF+Web+ngram$ system. Clearly, the proposed lexicon suggested more than one target word for each source word. Moreover, the precision score could actually be improved further if only one candidate (i.e., the Top 1) was proposed for the source word. With the Top 1, the final precision for $MWCF+Web+ngram$ at 50% recall improved significantly from 20.00% to 43.56%. However, the approach did not work well with $SWCF+WC2$, where the precision at 50% recall dropped from 14.63% to 7.14%.

English	**Malay**	**Sim score**	**Correct?**
former	*presiden*	0.1294	No
president	*presiden*	0.1216	Yes
playmaker	*presiden*	0.0958	No
believes	*pengurus*	0.0540	No
coach	*jurulatih*	0.0318	Yes
league	*pemain*	0.0250	No
former	*pemain*	0.0250	No
coach	*pengurus*	0.0242	No
president	*liga*	0.0236	No
striker	*penyerang*	0.0214	Yes

Table 6.5: Some examples of the translation pairs learned by $MWCF+Web+ngram$ system and ranked by similarity scores

6.6 Discussion

6.6.1 The effect of using very small comparable corpora

The standard model (i.e; the $SWCF + MyWC$) was implemented using a very small comparable corpora. The system yielded 18% of best $F1$ score, which is considered poor. In contrast, a similar attempt on using the same corpora but based on the multi-word approach (i.e., the $MWCF$) yielded poorer performance. In other word, the latter attempt underperformed than expected. Hence, the use of such corpora are not reliable in the extraction task, thus entailing more data for any extraction system to perform at the desired level.

Setting	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$F_1 score$
$SWCF + MyWC$	0.00	9.75	14.78	14.63	18.00
$MWCF + MyWC$	-	-	-	-	-

Table 6.6: Performances of the different methods using extremely small corpora

6.6.2 Improvement in Performance from the Single-Word-Features to Multi-Word-Features

In this study, the findings from the experiments conducted clearly suggest that multi-word context feature method would achieve higher best F_1 score than the standard context-based when more data are queried from the web. The better performance of the former is attributed to taking more than one word for consideration that might alleviate the ambiguity of each feature. Moreover, the potential of the multi-word level context feature approach could be exploited through the use of massive data. However, this massive requirement is not easily met, thus entailing a new direction in focusing efforts to obtain the necessary data. One of the efforts that seems practical and feasible is by by incorporating data from the web.

Method	$P_{0.50}$
$MWCF + Web + 3 - gram$	33.33
$MWCF + Web + 4 - gram$	100.00
$MWCF + Web + 5 - gram$	100.00

Table 6.7: Effects on precision score at 50% recall for $MWCF + Web + ngram$ (with Top 1 evaluation) in different n-gram windows

6.6.3 The Effects of different window sizes of the n-grams

The findings of the experiments show that the precision could be improved with larger n-gram size. However, this improvement requires a certain number of such features, which is very limited. In this regard, larger windows of the n-grams are good discriminators, but they are quite sparse. In addition, major part of the few articles found in the web for n-grams with larger window were actually the same articles in the $MyWC$ corpora that were downloaded previously; which could have a positive impact of the reliability of the documents. Table 6.7 shows the effects of the window size observed in the experiments.

6.6.4 Data sparsity problem and the Use of web

Data sparsity is a potential major problem when using the methods that rely on massive data, but this problem could be relieved by incorporating the additional data from the web. However, as a caution, the single-word context feature based approach might have added more noise to the existing data with this approach.

Overall, the outcome of the approach adopted in this study provides several important insights. First, not all the entire multi-word level context features are highly relevant to the source word (or target word, respectively). Secondly, a feature verification model could be used to allow only highly recognized features to be considered as the multi-word level features. Thirdly, a potential

major drawback of the multi-word context feature approach is that the majority of the features does not occur together with the target words in the web, although the web offers huge relevant corpus.

6.7 Conclusion

This chapter discusses the experiences gained in conducting bilingual lexicon extraction tasks in an extreme setting. Through this setting, several systems based on the multi-word context terms have achieved higher precision performance compared to the single-word context terms approach. In the former technique, an issue related to the independent term vector is inevitable. However, the problem could be solved by incorporating data from the web to help make the technique feasible. More importantly, using n-grams keeps the technique simple and useful in the extreme setting. Although the performance of the whole system based on this technique has not been tested, nonetheless, the preliminary findings seem to suggest that this technique is capable of achieving better performance.

Furthermore, the experimental findings of this study also reinforce the imperative of reliable and quality resources as an essential requirement for the development of a bilingual lexicon extraction system. Potentially, these resources and information are readily available from the Web; however, they need to be carefully utilized to avoid introducing serious noise problem to the system.

Chapter 7

Summary, Conclusion and Future Work

7.1 Thesis Summary

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

7.1.1 Summary of Literature

In chapters 2 and 3, a review on the different methods employed by existing work in the field has highlighted several important findings, in particular those pertaining to the context-based approach. This review highlights that many methods that have used such an approach have done so by obtaining important information from a context window. In particular, the review in Chapter 2 indicates the general classification of the methods based on the way context information is acquired and exploited. More precisely, these methods are further divided into sub-categories according to some important details of a bilingual lexicon extraction task as discussed in Chapter 3.

The review has also discussed past research that had recorded considerable success based on work focusing on parallel corpora. This is not surprising

as parallel corpora provide complete (or almost complete data), which can be used to bind a correct matching pair efficiently. However, parallel data are quite scarce, hence, learning from comparable corpora becomes a feasible alternative. However, learning bilingual lexicons from non-parallel texts has its own inherent problem. Early research in this particular field has mainly focused on finding the best common patterns that could represent translation equivalents of different language pairs. As expected, the attempted efforts have faced some difficulties because of the characteristics of comparable corpora.

Later, it was learned that co-occurrence information deriving from a context window can be a good feature to match a word to its translation equivalents. In essence, co-occurrence information represents the relationship between a test word and its context word. A word in the target language and a word in the source language that share similar co-occurrence information are deemed to be a translation pairs. In view of this significant finding, the approach using co-occurrence information, known as the context-based approach, has since become a widely used method to tackle the problem of word-to-word correspondences. However, the methods that have been proposed to date are generally dependent on good quality comparable corpora and a bilingual lexicon, otherwise the precision can be compromised. In previous work, the translation pairs were typically learned from large comparable corpora, containing not less than a million words. A large, machine readable initial bilingual lexicon with 16,000 to 20,000 entries thus becomes a major requirement for such systems to function.

Co-occurrence information comprises many different data; however, co-occurrence counts and number of common words are not likely to correspond to each other significantly. Many researchers agree that the actual ranking of the context word frequencies provides important clue to the similarity. To determine the similarity, they suggest assigning a weight to each context word using association measures. In this regard, LLR has been widely used to achieve such

a goal. This method measures the likelihood of a context word to co-occur with a test word. Besides this method, there are other different association measures available that can be used to achieve the similar aim.

More progress has been made through current work, which has introduced new techniques to improve the measure, for example, by giving higher priority to the important context words or by retaining context words that are positively correlated. In addition, some researchers have carried out studies that focus on using latent data to extract translation pairs by finding common patterns in an underlying dimension. Likewise, there are other researchers who have performed experiments to examine the impacts of translation precision by using minimal resources. However, the results of these work are not quite robust to warrant their actual applications.

Chapter 3 mainly discusses studies that are related to the context-based approach. In particular, the discussion focuses on the critical components that are required in a bilingual lexicon task. The discussion that follows is centred on the general approach to be used when using the individual components.

To summarise, most previous work that have been carried out are based the context-based approach. Currently, research in the lexicon translation is progressing–albeit at a slow pace–with more new techniques being introduced to improve the conventional methods. More poignantly, research on minimal use of resources has not been carried out with greater depth. As a result, the field in translation lacks new, latest knowledge on methods that may benefit most language pairs. Based on the current literature, the debate on which technique represents the best method to obtain high precision bilingual lexicon is still being pursued. The surveys discussed in chapters 2 and 3 highlight the research gaps found in the literature, which pave the way for the pursuance of the study to address the first objective of this research (see Section 1.4).

7.1.2 Summary of Empirical Work

In this study, the researcher has developed a few novel techniques for learning bilingual lexicon from comparable corpora. Moreover, the conditions imposed on the techniques study was made very challenging with the absence of vast resources and major linguistic tools. Essentially, all the techniques have been tested in a series of experiments using minimal resources, the performances of which were compared the baseline measurements.

Initially, a conceptual framework was conceptualized, which is presented in Chapter 3, to highlight important components required in a basic bilingual lexicon extraction task. For each component of the framework, the general approaches that have been used in earlier studies were used as guidelines. Subsequently, a series of experiments were conducted to evaluate three different settings. The main objective of these evaluations was to identify and, subsequently, to justify the appropriate approach for each component, which was used as the baseline setting in this study. In addition, each of the experiments conducted was based on minimal use of resources. To ensure high reliability, only small comparable corpora containing about 40,000 to 50,000 words in each language were considered as the minimum requirement in this study. More precisely, in this research field, a bilingual corpus with less than a million words is considered small to render reliable results.

For the evaluation in the first setting, the systems that used the cosine measure on real and binary values have yielded good results. The result of this experiment strongly suggests that cosine measure provides a good measure of similarity. This important finding reinforces the imperative of this measure not purely due to its proven efficiency (and its popular utilization), but the ease in which this method interprets similarity value, ranging from 1 for identical through 0 for orthogonal, and -1 for very dissimilar context vectors, respectively.

For comparing comparable corpora in the second setting, the researcher set up three different sets of comparable corpora from different sources. As expected, the system that had used larger comparable corpora of related domain yielded encouraging results. However, this finding has be interpreted with caution as different sizes, types and language pairs may influence such results. Nonetheless, the corpora containing about 400,000 to 500,000 words used in this study is relatively small, thus the finding is quite significant. In the third evaluation, it was found that having larger size of initial bilingual lexicon did not necessarily ensure better performance of the system. Surprisingly, and more revealing, this study has shown that a system based on a small bilingual lexicon of 100 entries (which were hand-compiled from high frequency corpus words) could outperform other rich-resource systems. Furthermore, the researcher has demonstrated that a small high precision bilingual lexicon could be derived automatically from comparable corpora of related language pairs. However, in this study, the bilingual entries were quite restricted, thus prohibiting the translation of most translation pairs.

Finally, the analysis of CCA performed in this study has been proven to be very effective and robust. More importantly, it was observed that this type of analysis is quite delicate, entailing careful measures to eliminate serious noise that could produce spurious outcomes. Using such measures, a CCA-based model would have a better chance to derive a high precision bilingual lexicon. Nonetheless, this approach may have its own share of problems; thus, the results obtained needs careful interpretation.

In view of this potential problems, the researcher developed a system based on the basic approach, utilizing one of the best settings as the baseline whenever possible. The use of this approach helped the researcher to address the second objective of this thesis (see Section 1.4).

A bilingual lexicon task is mainly composed of several important components as discussed in the literature review, where each component has been studied

through small base experiments. More significantly, this study also dealt with two components deemed critical in the adopted approach, namely (a) the automatic construction of the source word and the target word vocabulary lists (see Chapter 4), and (b) the restricting context terms (see Chapter 5). Moreover, this study also paid greater emphasis on a particular technique related to context terms (see Chapter 6) in establishing an extreme setting, in which the source of the data acquisition was the World Wide Web. Hence, on closer examination, the objectives of chapters 4 and 5 are quite different from the objective of Chapter 6.

In Chapter 4, a novel technique that utilized contextually relevant words to form vocabulary lists is discussed, with a particular reference to cognate pairs to help derive these words. The technique was used as the initial step, allowing the constructed lists to be used right away by incorporating them in any basic system. Subsequently, the model was compared to the baseline system, which had been selected prior to the experiment. The findings suggest that a good set of vocabulary lists is essential for precise matching. Conversely, a lack of vocabulary lists would lead to poor performance (i.e., mismatching of word pairs).

Another important technique used to improve the set of context terms is presented in Chapter 5. The experiment carried out to test this technique produced an interesting finding, where the in-domain terms introduced in the systems managed to emphasize the significant context terms. Hence, this finding underscores the importance of these terms in the translation efforts.

Finally, this study was also performed to address another importance objective, namely to demonstrate a method that could be deployed to efficiently harvest data from the web (see Chapter 6). In many cases, an effective way to set queries in search engines and to retrieve results of comparable texts from the search is highly emphasized . To address this emphasis, a system entails some comparable data to begin with. Hence, the researcher proposed a novel

method to help acquire comparable texts from the web automatically. Added to this new approach, the researcher also introduced a new technique that could transform the context terms indirectly, rendering the term vectors for a word be dependent with one another; thus, the essence of a word would not be left undetected. In addition, a purely unsupervised task was performed using the automated bilingual lexicon system in this study. As expected, the results of experiment were not very impressive, which might be attributed to several reasons, notably the constraints of unrelated language pairs and the limited domain of the comparable corpora. In contrast, the systems performance has increased slightly when more data from the web were used, indicating the potential benefits of using web data.

Overall, the findings of this study, which involved a series of experiments, strongly suggest that minimally-supervised techniques are not only applicable but efficacious in translation work, as has been successfully demonstrated in the restricted settings. However, the findings of the experiments in this study need to be judiciously judged as the characteristics of the resources used in the computations were specific to the comparable corpora and initial bilingual lexicon used. Hence, the use of a particular resource may also play an important role as do the other components. More precisely, rather than focusing too much effort on finding the best technique for each component to function, a firm grasp on the unique impact of available resources is needed. This understanding would allow researchers and practitioners to deploy each component of a particular technique in an effective setting, thus utilizing the resources optimally. In other words, defining the right components for a particular set of resources must be carried out first to achieve better performance. As learned from the several findings, the researcher recommend the use of IDT as a better alternative that helps learn a bilingual lexicon accurately from corpora with similar characteristics (e.g., MyEuroparl). Furthermore, the techniques proposed would be able to perform over a spectrum of resources, ranging from low to high volume of materials. Based on the promising findings of a series of experiments using the novel techniques, the third objective (see Section 1.4)

of this study is thus addressed.

7.2 Research Contributions

The major contributions of this thesis are as follows:

- *Introducing new techniques in learning bilingual lexicon from comparable corpora within minimal supervised settings*
  In this thesis, each component in a bilingual lexicon extraction task has been treated as an important factor that is detrimental to the performance of the task. All the components in a bilingual lexicon extraction task have been dealt with as discussed in Chapter 3. In particular, two of the components have been given special focus as elaborated in Chapter 4, 5 and 6. The techniques have been deployed and tested, yielding results that have substantial impacts on translation efforts, notably in generating precise translations. Their promising performance owes much to these critical components that help minimize mismatching compared to the baseline systems.

  Moreover, the system has performed better through improved context terms based on IDT method than the ECST method, where both of them were tested under the same setting. Likewise, similar results using three different settings have also been observed. These findings have several implications in practice, in particular the greater impact of the context terms on the extraction task compared to the vocabulary list. Premised on this context, other useful techniques could also be developed by using each of the components of the framework in several different settings. More importantly, the research finding underscores the impact of IDT technique in building word sense from disambiguated lexicons, ensuring precise translations. Figure 7.1 summarizes the contributions made in this study.

Main issue		Approach	
	An automated bilingual lexicon	Context-based	By focusing on translation equivalents with similar contexts
	Lack of resources	Minimally supervised	By finding techniques that help learning a high precision bilingual lexicon from minimal resources
Less comparable word occurrence frequencies between corpora	Missing target words Domain stop words	ECST	By taking context terms of a cognate pair
Less balance corpora	Hypernymy Polysemy	IDT	By focusing on second order context terms of a test word shared with one of its context term (a cognate) at a time to restrict its context
	Hyponymy Synonymy	More novel minimally supervised, or less supervised techniques yet to be found	

Figure 7.1: Issues and approaches in the minimally supervised approach

- *The introduction of a novel way for mapping bilingual word pairs using ranking information*

  An initial bilingual lexicon is required to match context vectors in the same word space. This study has tested a novel method that could compute the similarity between a translation candidate pair without relying on any initial bilingual lexicon. More precisely, the similarity was computed using a method called the rank-binning by assuming relative distributions of in-domain terms of translation equivalents were roughly comparable. The performance of this method was observed to be exceptionally high, reinforcing the usefulness of this approach in effective

translation work.

- *The application of an automated initial bilingual lexicon technique*
  Automatic resource construction is not only relevant to but also urgently needed in this field. In this regard, this study has demonstrated the positive impact of an approach based on Koehn and Knights model in constructing the initial English-Spanish bilingual lexicon. In addition, several experiments have been conducted to examine the robustness and versatility of the new automatic approach in handling diverse lexicons, which resulted in encouraging results. Overall, the automated initial bilingual lexicon method can be used to run purely on unsupervised settings for under-resourced languages. Thus, many languages of many nations, which mainly contain less lexicon density, are now more accessible for translation efforts. In addition, this approach is deemed more suitable for related language pairs that share many similar spelling words, suggesting that a pair of languages that has many loanwords would facilitate better performance.

- *The application of CCA in a minimal supervised setting*
  Lower dimensional data may resolve problems with high dimensional data. The CCA-based approach seems a viable method, but its application in translation field thus far is scarce. Hence, the findings of this study provide useful, important insights and clues to better exploit the CCA approach in developing efficient systems. However, despite the supporting evidence to use this approach in translation work, it is worth to caution its full-fledged application because this method (in its current structure) is vulnerable to severe noise. Nonetheless, its potential benefits outweigh its drawbacks, making it a preferred approach in future undertakings.

In addition to the above contributions, several tools and resources (availability based on request to the author) that have been produced in this study could provide support for certain translation work. These materials include tools

for the corpus pre-processing, log-likelihood computation and similar spelling identification. Resources such as the small Spanish-English dictionary and the World Cup 2010 small corpora are also available. The remaining tools and resources will be made available once they are ready for download.

7.3 Recommendations for Future Work

In this section, the author provides several recommendations that are important in improving bilingual lexicon extraction tasks based on the lessons learned from this research. In doing so, the final objective of this study (see Section 1.4) is thus addressed.

- *Determination of the actual characteristics of resources*
  As demonstrated in this study, a minimally supervised bilingual lexicon extraction task entails a certain amount of resources for the translation process to be operational. Furthermore, the actual characteristics of each resource need to be accurately defined in the first place; only then, the relevant components for the task can be appropriately determined to ensure efficacious application. That being said, however, defining each characteristic of corpora and bilingual lexicon is not easily executed. Thus, the author strongly suggest the application of an automated tool, which will not only simplify this task, but also to perform the extraction with greater precision. Hence, some of these recommendation could be further explored in the future work.

- *Chain reaction between one component to another*
  Each component of the extraction systems has its own specific role in a bilingual lexicon extraction task. A firm understanding of each individual components unique role will highlight its inherent impact on the bilingual lexicon extraction task. More precisely, the use of several components in the systems will bring in the accompanying combined impacts, where each component could attenuate or amplify other components impacts, thus creating a chain of reactions among them. Depend-

ing on the level and magnitude of this combinatorial interplay among the system components, the performance of the systems could proceed in either directiongood or bad. Thus, the author recommends identifying the actual parameters of each component through some proper techniques that could screen out the best parameters deemed suitable for the components.

- *Definition of similarity in bilingual lexicon extraction*
  Interestingly, the similarity of two different features can be defined differently, which has been observed in the experiments conducted in this study. For example, using the cosine measure the similarity can be based on the component values (or vector weights) or the way feature vectors are generated, i.e., the size of dimension and variance of the norms of the vectors. Many other variant forms of cosine measure can therefore be generated. In addition, there are other approaches to defining similarity than just the cosine or Euclidean distance measures, namely the Dice, Jaccard, and Kullbeck-Leibler measures, which could as well be effectively used. Thus, predicting in advance the appropriate definition of a particular similarity measure to be used for a specific application is far more beneficial than attempting to determine the overall best measure among all similarity measures.

- *Establishment of Gold Standard Data for comparison purposes*
  As has been experienced in this study, efforts to make comparisons involving a set of systems will be daunting in view of the data available to researchers. In most likelihood, the data may differ in many aspects, such as density, form, and structure. Hence, to overcome this barrier of comparison efforts, a set of gold standard data for comparison work should be established to streamline all activities into a similar, equivalent setting.

- *Word hypernymy, polysemy, hyponymy, and synonymy*
  This study , in part, addressed the hypernymy and polysemy problems

using the techniques adapted in a series of experiments. The important findings reinforce the applicability of WSD as a novel, effective method for bilingual lexicon extraction tasks. Word hyponymy and synonymy, although have not been specifically addressed in this study, are expected to bring in their fair share of influences in bilingual lexicon extraction tasks. This assertion is not too far-fetched as Somers (2001) notes that the assumption of 1:1 word correspondence is too naive since polysemy, homonym and inflectional problems do occur much or less. Hence, the issues related to these two word categories would be feasible candidates for future research.

- *Automated error analysis*

  Error analysis is a very important step to check incorrect pairs; but, this analysis is very tedious and time-consuming. Thus, an automated error analysis tool for bilingual lexicon extraction, which is able to help identify the errors in the output of data, to visualize the patterns and to suggest the error types would be highly favourable. The error types may include the semantically related (synonymy) and the semantically unrelated (but contextually related) cases. This tool should be able to discern the correct and incorrect target equivalents, where both equivalents appear in the same context window (i.e., hypernymy or hyponymy). In addition to both correct and incorrect target equivalents appearing within the same context window, they might be separated such that a distance keeps them apart (i.e., domain stop word). Likewise, both correct and incorrect target equivalents would appear within the same document or corpora (i.e., general noise word), inflection or morphological error, and unidentified error.

- *Other related analysis*

  Several CCA-based models, which were successfully tested in a series of experiments in this study, are effective in bilingual lexicon extraction tasks. It would be interesting to explore the potentials of other analyses in this field, such as Latent Dirichlet Analysis (LDA) and Conditional

Random Field (CRF). However, implementing these analyses would entail major linguistic pre-processing, thus demanding considerable efforts. To confront such a situation, the author suggests the use of automated linguistic tools (though they are rarely used) to perform minimally supervised works. If these tools are not used, the assumption regarding the availability of major pre-processing tools needs to be factored in any study of this kind.

- *The needs for methods with faster computation*
  One of the major limitations of the current methods is the time that they take to process relevant data. In other words, the current methods require excessive amount of working hours in handling a large corpus of text. For example, the LLR method would typically process to completion 500,000 words in a monthan intolerable time constraint for any practical use. Thus, efficient computing methods should be the main focus in future research to produce faster systems in the bilingual lexicon extraction. Ultimately, more research is needed to develop more systems that are both effective (i.e., in generating precise translations) and efficient (i.e., in taking less time in processing data).

References

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N., and Yarowsky, D. (1999). Statistical machine translation. *Technical report Center for Language and Speech Processing John Hopkins University*. 2, 3

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D., and Yamada, K. (2000). Translating with scarce resources. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. 38

Alajmi, A., E.M., S., and R.R., D. (2012). Towards an Arabic stop-words list generation. *International Journal of Computer Applications*, 46(8). 123

Andrade, D., Nasukawa, T., and Tsujii, J. (2010). Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 19–27, Beijing, China. 48, 81, 83, 95, 129, 130, 200

Baroni, M. and Bernadini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316. 115, 209

Brown, P. F., Cocke, J., Pietra, S. D., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2). 3, 30, 32

Callison-Burch, C. and Osborne, M. (2003). Bootstrapping parallel corpora. In *Proceedings of Workshop on Building and Using Parallel Texts: Data*

Driven Machine Translation and Beyond, HLT-NAACL 2003, pages 44–49. 38, 39

Callison-Burch, C., Talbot, D., and Osborne, M. (2004). Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, pages 175–182, Barcelona, Spain. 14, 29

Cancedda, N., Dejean, H., Gaussier, E., Renders, J. M., and Vinokourov, A. (2004). Report on CLEF-2003 experiments: two ways of extracting multilingual resources from corpora. *Lecture Notes in Computer Science (LNCS)*. 92

Chen, J., Yeh, C. H., and Chau, R. (2004). Identifying parallel web documents by filenames. *Lecture Notes in Computer Science (LNCS)*, 3007. 16, 115, 210

Chiao, Y. C., Sta, J. D., and Zweigenbaum, P. (2004). A novel approach to improve word translations extraction from non-parallel, comparable corpora. In *Proceedings of International Joint Conference on Natural Language Processing*, Hainan, China. 67

Chiao, Y. C. and Zweigenbaum, P. (2002). Looking for French-English translations in comparable medical corpora. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA). Biomedical Informatics: One Discipline*, San Antonio, Texas. 64, 65, 66, 67, 95, 118, 127, 128, 129, 130

Collier, N., Kumano, A., and Hirakawa, H. (2003). An application of local relevance feedback for building comparable corpora from news article matching. *NII*. 116

Dagan, I. and Church, K. (1994). Termight: identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 34–40. 13, 23, 24

Diab, M. and Finch, S. (2000). A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-based multimedia information access (RIAO)*. 18, 34, 74, 75, 95, 118, 119, 126, 128

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74. 28

Fry, J. (2005). Assembling a parallel corpus from RSS news feeds. In *Proceedings of the Workshop on Example-based Machine Translation, MT Summit X*, Phuket, Thailand. 112, 114, 209

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*. 4, 5, 35, 55, 95, 120, 128, 131, 146, 163

Fung, P. (1998). *A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora*. Springer, Berlin. 53, 54, 61, 63, 95, 127, 129, 130, 135

Fung, P. and Cheung, P. (2004a). Mining very non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the Conference on Empirical Method in Natural Language Processing (EMNLP)*, Barcelona, Spain. 15, 124, 137

Fung, P. and Cheung, P. (2004b). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*. 3, 4, 16, 17, 19, 106

Fung, P. and Church, K. W. (1994). K-vec: a new approach for aligning parallel texts. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1096–1102. 14, 25, 26

Fung, P. and McKeown, K. (1994). Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *Proceedings of AMTA 94*, pages 81–88. 14, 26, 27, 106, 126

Fung, P. and Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the COLING-ACL*. 13, 17, 21, 22, 25, 35, 52, 53, 64, 106, 128

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting, Association for Computational Linguistics*, pages 177–184. 24, 25

Garera, N., Callison-Burch, C., and Yarowsky, D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 129–137. 96

Gaussier, E., Renders, J., Matveeva, I., Goutte, C., and Dejean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Conference of the Association for Computational Linguistics (ACL)*. 87, 91, 120, 124, 128, 130, 131, 133

Goeuriot, L., Morin, E., and Daillei, B. (2009). Compilation of specialized comparable corpora in French and Japanese. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpus ACL-IJCNLP*, pages 55–63, Suntec, Singapore. 108, 109, 110

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779. 3, 6, 7, 90, 91, 92, 95, 98, 124, 125, 131, 132, 137, 157, 158, 163, 164, 197

Hao, L. (2008). Automatic identification of stop words in Chinese text classification. In *International Conference of Computer Science and Software Engineering*, Wuhan, China. 123

Haruno, M., Ikehara, S., and Yamazaki, T. (1996). Learning bilingual collocations by word-level sorting. In *Proceedings of the 16th Conference on Computational Linguistics (COLING)*. 208

Hwa, R., Nichols, C., and Simaan, K. (2006). Corpus variations for translation lexicon induction. In *Proceedings of AMTA 06*. 19, 39

Kaji, H., Kida, Y., and Morimoto, Y. (1992). Learning translation templates from bilingual text. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 672–678. 24

Kikui, G. (1998). Term-list translation using mono-lingual word co-occurrence vectors. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*. 70, 71, 72, 73

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *MT Summit*. 137

Koehn, P. and Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. 3, 5, 18, 35, 36, 120

Koehn, P. and Knight, K. (2001). Knowledge sources for word-level translation models. In *Proceedings of the Conference on empirical method in natural language processing (EMNLP)*. 4, 13, 22, 29, 32, 33, 36, 217

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL 2002*, pages 9–16, Philadelphia, USA. 6, 21, 33, 34, 35, 37, 40, 42, 68, 70, 95, 127, 128, 131, 137, 163, 173

Laroche, A. and Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23th International Conference on Computational Linguistics (COLING)*, pages 617 –625. 117, 119, 120, 124, 127, 128, 129, 136

Manning, C. and Schütze, H. (2002). *Foundations of statistical natural language processing*. MIT Press, Cambridge MA. 3, 19, 20, 21, 47, 119, 129, 131, 132

Masuichi, H., Flournoy, R., Kaufmann, S., and Peters, S. (2000). A bootstrapping method for extracting bilingual text pairs. 104

Melamed, I. D. (1995). Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *The Third Workshop on Very Large Corpora (WVLC3)*. 14, 26, 28, 43

Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*. 25

Melamed, I. D. (1997). A word-to-word model of translational equivalence. *CoRR*. 28, 29, 129, 131

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. In *Proceedings of Conference of the Association for Computational Linguistics (ACL)*, page 107130. 24

Melamed, I. D. (2000). Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249. 14, 29

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the ACL 2006*. 4, 19, 107, 108

Otero, P. G. and Campos, J. R. P. (2005). An approach to acquire word translations from non-parallel texts. In *EPIA*, pages 600–610. 36, 37, 76, 77, 78, 79, 95

Otero, P. G. and Campos, J. R. P. (2008). Learning Spanish-Galician translation equivalents using a comparable corpus and a bilingual dictionary. *Lecture Notes in Computer Science (LNCS)*. 96, 120, 124

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95, Portland, Oregon. 210

Pekar, V., Mitkov, R., Blagoev, D., and Mulloni, A. (2006). Finding translations for low-frequency words in comparable corpora. *Journal of Machine Translation*, 20(4). 126

Pietra, S. D. and Pietra, V. J. D. (1994). Candide: a statistical machine translation system. In *Human Language Technology (HLT)*. 13, 29, 31

Prochasson, E., Morin, E., and Kageura, K. (2009). Anchor points for bilingual lexicon extraction from small comparable corpora. In *Machine Translation Summit XII*, pages 284–291. 79, 95, 98, 117, 127, 128, 129, 130, 200

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd ACL*, pages 320–322. 4, 22, 34, 44, 57, 58, 106, 129, 162

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th ACL*, pages 519–526. 14, 23, 34, 40, 45, 51, 59, 60, 61, 68, 74, 95, 120, 124, 127, 128, 129, 130, 131, 163, 164, 215

Rapp, R., Sharoff, S., and Babych, B. (2012). Identifying word translations from comparable documents without a seed lexicon. In *Proc. the Eighth Language Resources and Evaluation Conference, LREC 2012*, Istanbul. 136

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. 18

Resnik, P. and Smith, N. A. (2003). Web as a parallel corpora. *Computational Linguistic*, 29(3):349–380. 102, 103, 111, 113, 210

Sahlgren, M. (2006). *The word-space model.* Ph.D thesis, Stockholm University. 49, 51, 52

Sharoff, S., Babych, B., and Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. In *Proc. of International Confenrence on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006*, pages 739–746, Sydney. 34

Shezaf, D. and Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 98 –107, Uppsala, Sweden. 84, 95, 119, 125, 129, 130, 131

Simard, M., Foster, G. F., and Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation.* 24, 25

Somers, H. (2001). Bilingual parallel corpora and language engineering. In *Anglo-Indian Workshop Language Engineering for South-Asian Languages (LESAL).* 15, 16, 17, 18, 102, 111, 236

Tanaka, K. (2002). Measuring the similarity between compound nouns in different languages using non-parallel corpora. 96

Tanaka, K. and Iwasaki, H. (1996). Extraction of lexical translations from non-aligned corpora. In *Proceedings of Conference of Computational Linguistic (COLING).* 84

Tanaka, K. and Matsuo, Y. (1999). Extraction of translation equivalents from non-parallel corpora. 96

TszWai, R., He, B., and Ounis, I. (2005). Automatically building a stopword list for an information retrieval system. In *5th Dutch-Belgium Information Retrieval Workshop*, Utrecht, The Netherland. 123

Weeds, J. and Weir, D. (2003). A general framework for distributional similarity. In *Proceedings of EMNLP 03*, pages 81–88. 48

Wong, W. Y. (2009). *Learning lightweight ontologies from text across different domains using the web as background knowledge*. PhD thesis, University of Western Australia. 191

Yamamoto, K., Matsumoto, Y., and Kitamura, M. (2001). A comparative study on translation units for bilingual lexicon extraction. In *Proceedings of the Workshop on Data-Driven Machine Translation, the 39th ACL*, pages 87–94, Toulouse, France. 208