AUTOMATIC TEXT ALIGNMENT USING RECURSIVE HAPAX-BASED CUT-THROUGH FRAGMENTATION

NG PEK KUAN

UNIVERSITI SAINS MALAYSIA

2012

AUTOMATIC TEXT ALIGNMENT USING RECURSIVE HAPAX-BASED CUT-THROUGH FRAGMENTATION

by

NG PEK KUAN

Thesis submitted in fulfillment of requirements

for the degree of

Master of Science

March 2012

ACKNOWLEDGEMENT

It has been a long way from the moment I started my journey of doing my Masters research both in Universiti Sains Malaysia (USM), Penang and Paris, France. Now that it is all coming to the end, firstly, I would like to show my utmost gratitude to my family for their everlasting love, support and encouragement.

I would like to show my heartfelt gratitude to Dr. Cheah Yu-N, my main supervisor and Dr. Bali Ranaivo, my co-supervisor for their invaluable help and knowledgeable advice. Without them, I would never know the fun of doing research and the interesting side of Natural Language Processing (NLP). Besides, I would like to thank Prof. Fathi Debili who is my field supervisor in France for giving me the opportunity to spend a year doing my Masters research as a trainee in LLACAN laboratory in Paris and two months working with his students in Tunis, Tunisia. He is the source of inspiration and the idea of hapax-based alignment. In addition, Dr. Dickson Lukose, my supervisor from MIMOS has given me much guidance and constructive advice. Thank you for being so helpful and encouraging.

Moreover, I am lucky enough to have a group of intelligent and generous colleagues and friends around me. Ye Hong Hoe provided me the annotated English-Malay parallel texts. Lim Lian Tze shared her work on the English-Malay bilingual dictionaries. Tan Kar Huan inspired me to build an automatic aligner by giving much feedback on the manual

alignment system that I have built earlier. Chong Chai, Kar Chuan, Mee Mee, Eric and Muk Moi in Penang, Sidnei, Nthatisi, Unore, Sokhna, Cleonice, Nan Nan in Paris, Zied and Sofiene in Tunis, thank you for being my friends and source of support.

Next, I would like to thank the School of Computer Sciences, USM, in particular what was formerly the Computer Aided Translation Unit (UTMK), and LLACAN laboratory in Paris, for allowing me to further my studies at such perfect places for research with the desired environment, needed facilities, perfect research culture as well as friendly and helpful professors and staff such as Mdm. Rohana Omar and Mr. Tan Ewe Hoe in UTMK, Dr. Martin Vanhove, Mme. Magali Diraison, Mme. Jeanne Zerner, Mme. Sylvie and Dr. Catherine Reigner in LLACAN.

Last but not least, I thank MIMOS Berhad and the Embasssy of France for the scholarship to further my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MT	Machine Translation
NLP	Natural Language Processing
POS	Part of speech
SMT	Statistical Machine Translation
EBMT	Example-based Machine Translation
RBMT	Rule-based Machine Translation
CBMT	Corpus-based Machine Translation
UTMK	Computer aided translation unit
SSTC	Structured String-Tree Correspondence
S-SSTC	Synchronous Structured String-Tree Correspondence

PEMADANAN TEKS AUTOMATIK MENGGUNAKAN FRAGMENTASI PERSILANGAN BERULANG BERDASARKAN HAPAX

ABSTRAK

Komunikasi melalui Internet telah menjadi keperluan asas kehidupan. Sistem penterjemahan berkomputer berbilang bahasa telah dihasilkan untuk menyokong komunikasi sebegini. Salah satu daripada kaedah yang digunakan secara meluas ialah pendekatan berdasarkan contoh yang memerlukan banyak contoh sebagai rujukan. Contoh-contoh ini disediakan dengan memadankan teks selari secara manual atau secara separa-automatik dengan campur tangan manusia. Ini memerlukan usaha dan masa yang banyak memandangkan banyak contoh diperlukan untuk menjamin mutu penterjemahan. Tambahan pula, manusia boleh melakukan kesalahan dan mempunyai pilihan tersendiri. Ini menyebabkan masalah ketekalan. Oleh itu, terdapat keperluan yang mendesak untuk membina sebuah sistem pemadanan automatik.

Bahasa Melayu merupakan bahasa yang mempunyai sumber yang kurang daripada segi peralatan dan data linguistik. Usaha membina sebuah sistem pemadanan yang dapat menandingi ketepatan sistem pemadanan bahasa lain yang mempunyai banyak sumber merupakan suatu cabaran oleh kerana kekangan sumber. Objektif penyelidikan ini ialah

untuk mereka kaedah pemadanan baru dengan menggunakan sumber linguistik yang minimum tetapi masih dapat mencapai ketepatan yang berpatutan untuk mengautomatikkan tugas pemadanan dan dengan itu mengurangkan usaha dan masa.

Kaedah penyelidikan ini melibatkan dua fasa. Fasa pertama melibatkan penyediaan sumber and perbaikan peralatan. Fasa kedua melibatkan perekaan algoritma pemadanan. Algoritma yang dicadangkan menggunakan konsep hapax, persilangan, fragmentasi dan perulangan. Ia memotong teks kepada fragmen yang lebih pendek secara berulang berdasarkan pemadanan hapax tanpa menghiraukan sempadan logik teks dan menghasilkan pemadanan baru pada setiap ulangan. Pendekatan persilangan telah digunakan untuk menyingkirkan pemadanan yang salah untuk memaksimumkan jumlah fragmen and pemadanan perkataan.

Keputusan menunjukkan bahawa algoritma berdasarkan hapax yang dicadangkan mencapai keputusan yang baik dengan kepersisan 92.62% dan ingat kembali 74.53% untuk pemadanan perkataan walaupun dengan kekangan sumber. Dengan ketepatan sebegini, campur tangan manusia dapat dikurangkan. Dengan itu, ketekalan dapat ditingkatkan dan usaha, masa serta kesilapan dapat dikurangkan.

AUTOMATIC TEXT ALIGNMENT USING RECURSIVE HAPAX-BASED CUT-THROUGH FRAGMENTATION

ABSTRACT

Communication over the Internet becomes the necessity of life. Multi-lingual machine translation systems are developed to support such communication. One of the most commonly used approaches is the example-based approach which requires a large set of examples as reference. These examples are prepared by aligning the parallel texts either manually or semi-automatically with human intervention. This requires much effort and is time-consuming considering the large number of examples needed to ensure the quality of the translation. Moreover, the fact that humans make mistakes and has preferences raises the consistency issue. Hence, there is an urgent need to develop an automatic aligner.

Malay is a language with scarce resources in terms of linguistic tools and data. It becomes a challenge to develop an alignment system that achieves the same level of accuracy as those for the resourceful languages in view of the resource constraint. The objective of this research is to design a novel alignment method with minimum linguistic resources but still achieves a reasonable level of accuracy to automate the alignment task and thus minimizing the effort and time consumed.

The methodology of the research involves two phases. The first phase involves linguistic resource preparation and tool enhancement. The second phase involves the design of the alignment algorithm. Our proposed algorithm adapts the concept of hapax, cut-through, fragmentation and recursivity. It recursively cuts the text into smaller fragments based on the hapax alignment regardless of the logical boundaries of a text and generates new alignments in each iteration. The cut-through approach is used to eliminate wrong alignments in order to maximize the number of fragments and word alignments.

The results have shown that the proposed hapax-based algorithm performs well with a precision of 92.62% and recall of 74.53% for word alignment regardless of the major resource constraint. With such accuracy, human intervention could be minimized and thus increasing the consistency and at the same time decreasing effort, time and errors.

CHAPTER 1

INTRODUCTION

This chapter introduces the concept of alignment by laying out the close relationship between machine translation (MT) and alignment. The discussion starts with general paradigms of MT and focuses on the approaches and life cycle of EBMT systems. It describes the motivation and problems in alignment and then explores the difficulties encountered that are specific to English-Malay alignment. After the research background, it states the problem statements, research objectives, research scope and contributions. The chapter ends by outlining the organization of the thesis.

1.1 *Research Background*

The topic of this research is under the natural language processing (NLP) domain. NLP explores the problems and the possible use of computers in the automated generation, understanding and manipulation of natural human languages both in text and speech (Chowdhury, 2003). The major applications in NLP are MT, cross-language information retrieval, automatic extraction of bilingual lexica and terminology, word sense disambiguation and compilation of translation memories (Véronis, 2000).

Among the most significant corporations who are involved actively in NLP research are Google and Yahoo. Google has published papers on NLP while Yahoo has a search

technologies team with experts in search, data mining, natural language and data processing. Many major companies and business firms, such as IBM and Microsoft have set up laboratories specifically for NLP research too. NLP is listed as one of the major research areas in IBM while Microsoft has two natural language research groups for natural language computing and natural language processing. Other organizations such as banks have applied NLP technology in their business. All these examples show us the importance of NLP application in the world today.

1.1.1 Machine Translation (MT)

MT is the application of computers to the translation of texts from one natural language into another (Hutchins, 1986). As defined, the text is translated into another natural language. This means that ideally the target text must fulfill the fundamental characteristics of natural language in morphology, syntax, grammar and semantics. This is not an easy task and to-date it still remains as a challenge of MT.

The first attempt in "automating" translation was during the seventeenth century when a German monk invented a mathematical meta-language to represent the meaning of the text so that it could be used for translation using the equations (Freigang, 2001). However, it is still not "machine" translation. The first real MT system was invented by George Artsrouni and Petr Troyanskii when they applied for patents for "translating machines" during the mid 1930s even before the electronic computer was invented in the 1940s. The research on MT started to become more and more popular after the invention

of the electronic calculators. In 1954, the first public demonstration of the feasibility of MT was held at Georgetown University in Washington, D.C. as a product from a collaborative research project between the university and IBM (Hutchins, 2005).

Since then, research efforts were carried out actively for a decade with massive funding for MT projects in the US until the publication of the ALPAC report which is also known as the "Black Book on Machine Translation" in 1966 (Freigang, 2001). It claims that there was no need for further investment in MT and focus should be shifted to the development of machine aids for human translators. Such a widely condemned, biased and short-sighted report has brought a virtual end to MT research and development which led to a dark age of MT for over a decade.

However, research efforts still continued in Canada, France and Germany. The "Universal Translator" in Star Trek which was first seen in the original series episode Metamorphosis in 1967 has proved that MT was not forgotten after all (Okuda et al., 1999). Three years later, the first installation of a MT system, Systran, was done by the US Air Force in 1970 for translating Russian military scientific and technical documentation into English. The accuracy of the system for technical reports was claimed to be over 90%. Following this success was the installation of Systran by the European Commission in 1976 for translating the large volume of documentation from English to French. As it is said in Chinese proverb, good things always come in pairs. In the same year, another successful product, METEO was developed in Canada to translate weather reports (Hutchins, 1999). The MT system has gained attention from the public through

the robot character C-3PO in Star Wars which is equipped with a MT system of more than six millions languages (Melby, 1995). These successful cases efficiently stimulated and recovered the interest and confidence of both the public and the researchers in MT.

Considering the current status of MT development, since these significant successes, research in MT has been carried out actively. Nowadays, MT is widely used for translating scientific documents, technical reports (O'Neill-Brown, 1996), web pages (e.g. Systran), emails (e.g. WorldLingo) and even chat room (Machine Translation News International, 1997). It is also applied in information retrieval, information extraction and text summarization systems. The applications of MT were adopted in various fields such as government, agriculture, banking, commercial, and tourism. Among the most notable MT systems are Georgetown's GAT (Georgetown Automatic Translation), EUROTRA in the Western world from the olden days (Slocum, 1985) and Babelfish on AltaVista, WordLingo and Google Translate in recent years (Hutchins, 2007).

Based on the current performance of the MT systems, although it is still less than efficient to substitute a human translator, it is useful to provide fast and immediate translation as well as producing the first draft for the translator. As a conclusion, MT still has a long way to go but the future outlook is promising.

1.1.1.1 MT Paradigms

There are two main paradigms of MT. The first paradigm is rule-based machine

translation (RBMT). It is categorized by the linguistic rules used in the translation (Carl & Way, 2003). The main concern in this approach is the development of the rules which is extremely time-consuming, labor-intensive, error-prone and expensive.

To overcome the knowledge acquisition problem, the corpus-based machine translation (CBMT) was introduced and it is the focus of the current research. As its name implies, this paradigm relies on a large number of bilingual corpora to carry out the translation task. CBMT could be divided into two main directions, SMT and EBMT.

The SMT approach relies on the probabilistic and statistical models of the translation that were trained on a large number of bilingual corpora. It includes little or no linguistic information. Instead, it relies on the easily-measured features found in the corpus which can be used to predict the translation, such as co-occurrence of the words in the parallel text, the length of the sentences and the relative position of the word in the sentence (Trujillo, 1999).

The EBMT approach retrieves similar examples from the bilingual corpora database, adapting the examples to generate an equivalent translation in the target text (Sumita et al., 1990).

1.1.2 Alignment

Parallel text alignment is the establishment of correspondences between textual units

in one text into their equivalence in the translated or comparable text (Papageorgiou et al., 1994). The alignment could be done in paragraph, sentence, word, character and even byte level.

The concept of alignment is rather new compared to MT. It is first seen following the emergence of the concept of bi-text defined as the texts that are available in two languages in the 1980s (Harris, 1988). As the research in MT (Brown et al., 1990) and bilingual lexicography (Klavans & Tzoukermann, 1990) were getting more and more attention in the early 1990s, the researchers have realized the importance of the alignment system as the first step towards constructing the probabilistic dictionary used in MT and bilingual concordance used in lexicography (Gale & Church, 1991).

Since then, research on alignment was carried out actively. The initial effort focused on sentence alignment (Brown et al., 1991; Gale & Church 1991; Kay & Roscheisen 1993; Chen 1993). It was a good starting point for research in parallel text alignment with a high accuracy of approximately 96% - 99%. The researchers moved forward to challenge the word alignment (Gale & Church, 1991a) and even proceeded to character alignment (Church et al., 1993). More and more new approaches appeared and new elements were included to improve the quality of the alignment such as cognate (Simard et al., 1992; Ribeiro et al. 2001) and POS tagging (Papageorgiou et al., 1994).

Another crucial breakpoint in the history of alignment was the alignment of disparate languages. In the initial stage, alignment research focused on European language pairs.

However, starting from 1993, there were researchers who raised the concern of generalizing the alignment techniques which were developed for European languages to other language pairs, especially the Asian languages. The first success was achieved in aligning the English and Japanese versions of the AWK manual (Church et al., 1993). The victory continued with the successful attempt of aligning English and Chinese by Wu in 1994 (Wu, 1994) and proceeded further by the introduction of DK-vec algorithm for aligning Asian/Indo-European noisy parallel text (Fung & McKeown, 1994). Since then, researchers from all around the world were trying to adapt the technique to their own languages. Now, not only the popular and commonly used language pairs become the focus of alignment, the other less-used languages were also addressed.

Nowadays, the application of text alignment is widely spread across domains ranging from MT and lexicography to word sense disambiguation (Tufis et al., 2004a), parsing (Hopkins & Kuhn, 2006), question answering (Echihabi & Marcu, 2003), multi-lingual text compression (Conley & Klein, 2006) and identification of idiomatic expression (Villada Moiron & Tiedemann, 2006).

There are a lot of resources, parallel corpora and tools developed. For example, there are a lot of parallel corpora available with more than 200 language pairs such as the Canadian Hansard, the JRC-Acquis Multilingual Parallel Corpus, the Opus project, COMPARA, LILABAR and Europarl. Visualization tools for alignment such as Alpaco, Cairo, I*Link and UPlug were implemented. Some of the most notable alignment tools available are Giza++, BITAM, ATLAS, Twente, Plug PWA and Kvec++. Many projects

on alignment algorithms and tools were carried out such as ARCADE, EGYPT, MULTI-TEXT EAST, PLUG, GIZA++ and BITAM. This intensive development brings us to the conclusion that the research on alignment is notably active currently and will be progressing fast in the near future.

1.1.2.1 Problems in Alignment

During the parallel text alignment process, there are a lot of challenges and difficulties encountered. In order to provide an aligned corpus with the highest possible accuracy, these challenges must be addressed and solved.

First, corpus with noise is a common phenomenon which creates problems for NLP application. This problem is caused by inconsistency and different styles of translating text by human translators. Each human translator has his own style of translating. He might insert a chunk of text to elaborate further on the topic or omit some portion of the text if he feels it is not necessary as he translates the text. Consequently, a portion of text in a source text could be totally missing in the target text or vice-versa. Sometimes, the text could be substituted by fragments of text which is not a corresponding translation (Fung & McKeown, 1994). In other cases, reordering of the sentences or words in the translation text could happen quite frequently.

Secondly, beside text-related problems, formatting of the file could cause difficulties in alignment too. Pictures, figures, tables, bullets, headers, footers and other formatting

details could create a lot of problems. Misplacing or excluding these formatting details could lead to wrong alignments. For instance, many pictures in the English version of the AWK Manual are missing in the Japanese version (Church et al., 1993).

Thirdly, different languages have different sentence and word boundary characteristic. Many alignment approaches take sentence boundaries as the anchor points both in European languages (Brown et al., 1991; Chen, 1993; Kay & Roscheisen 1993) and Asian languages (Wu, 1994). However, location of sentence boundary for European languages and Asian languages might be different. For instance, sentence in English might be expressed as two sentences in Chinese. This causes the one-to-many or many-to-one sentence alignment problem. In fact, the end-of-sentence punctuation itself is ambiguous and different for every language. For instance, a period could denote both the abbreviation and sentence boundaries (Palmer, 1994). Some languages do not even have punctuation to denote sentence boundary, such as Thai (Aroonmanakun, 2007). Word boundaries for Thai, Chinese, and Japanese are explicit too. This has increased the complexity of the alignment task. The sentence and word segmentation is required and the errors in segmentation would be brought to the alignment phase. The problem becomes more complicated when the punctuations are lost in an OCR input.

Next, in the simplest and ideal case, one word in the source text corresponds to a word in the target text. However, this is not always the case. A word in the source text could be aligned to multiple words in the target text. Conversely, multiple word expressions in the source text could be aligned as a whole to a single word in the target

text too (Véronis & Langlais, 2000). These word-to-phrase correspondences always happen in idiomatic expressions and compound words. Their potential discontinuity and reordering will make the alignment task even more difficult.

Polysemy or multiple possible corresponding words will cause ambiguity in the alignment task. Polysemy refers to the phenomenon when a word has more than one meaning. If a word has two meanings, the problem crops up when both corresponding words for each meaning appears in the target text. Even if a word is monosemous, synonyms or multiple different corresponding words may make the alignment difficult. The situation becomes worse if the same corresponding word appears several times in the target text. In addition to the ambiguities mentioned, difference in POS of the corresponding words and different syntactic structures of the sentences between the source text and target text adds on to the problem. Determining the word with the corresponding translation is not as easy as it seems.

Furthermore, some of the function words, e.g. prepositions and auxiliary verbs, do not have any corresponding translation. There is no clear guideline on how to align these words (Véronis & Langlais, 2000). Current practice aligns these words to null. However, it creates confusion when a word is also aligned to null due to the incomplete dictionaries when it actually should be aligned to a word in the target text.

As a conclusion, full word alignment is almost impossible considering there are still many problems which could not be solved entirely yet.

1.1.2.2 Difficulties in English-Malay Alignment

The main difficulty in aligning English-Malay parallel texts is the lack of resources. No doubt, there are a lot of English resources. However, Malay is a language with scarce resources.

In terms of the data resources, there is no complete bilingual dictionary. The dictionary entries are inconsistent. We do not have a complete part-of-speech (POS) tagging of all the entries in the dictionary. There are not many parallel texts in Malay too. Needless to mention, there are fewer annotated or aligned parallel texts in Malay. There is neither Malay Wordnet nor treebanks.

As for tools, there are not many NLP tools for Malay either. There is no Malay lemmatizer. There is a publication proposing a Malay stemming algorithm but the software is not available. A simple Malay morphological analyzer is available to provide all the possible segmentations. The POS tagger is still on-going research. The Malay tagset is still under construction. Not even the POS annotations in the dictionaries are available. Besides, there is no syntactic parser specific to Malay. The current Malay applications are adapting other parsers, in fact, other NLP tools of other languages, to be used for Malay.

There are many research efforts done for the alignment of languages with scarce resources (Tufiş et al., 2003; Martin et al. 2005; Lopez & Resnik, 2005; Wang et al., 2006). However, effort in automatically aligning Malay parallel text is almost none.

Hence, research in resource preparation and NLP tool adaptation is necessary to bring the NLP research in Malay to the next level.

1.1.3 The Unbreakable Bond between Machine Translation and Alignment

There is a close relationship between alignment and MT, in particular EBMT. In fact, the bond between them is unbreakable. They are inter-dependant.

Firstly, the alignment is crucial in preparing the data and resources needed in EBMT such as the large collection of reference examples. Automatic extraction of lexicon and terminology is often done by aligning parallel corpora. A significant project in automatic acquisition of translation lexicon is the Twenty-One Project. Six bilingual lexicons were derived from the Dutch, French, English and German parallel corpus (Hiemstra, 1998). In addition to the European languages, Dan Tufis from Romania has done a lot of work in automatic translation lexicon extraction for English, Romanian, Slovene, Estonian and Hungarian. He proposed cost-effective, fast and accurate ways to extract translation lexicon automatically from the parallel corpora (Tufis et al., 2004; Tufis, 2002; Tufis & Barbu, 2001; Tufis & Barbu, 2001a; Tufis & Barbu, 2000). Tiedemann extracted bilingual lexicon for English, Swedish and Gerrman from the aligned parallel text too (Tiedemann, 1998). Beside these European languages, translation patterns were extracted from Japanese-English parallel legal corpora. These extracted translation patterns are very useful to both MT and bilingual dictionary compilation (Ohara et al., 2003). Moreover, translation units, both single-word unit and multi-word unit, were extracted from aligned

Chinese-English parallel corpora (Chang et al., 2002). Furthermore, a tool for extracting technical terminology using POS tagging and word alignment, Termight, was developed and used by the translators at AT&T Business Translation Services (Dagan & Church, 1994).

Secondly, the alignment is one of the major components or crucial modules in EBMT. A Chinese-English EBMT system which is based solely on the word alignment information of the Chinese-English parallel corpus was proposed considering the great need in Chinese-English translation and vice-versa due to the Olympic Games held in Beijing, China in 2008 (Yang et al., 2004). As for the SMT, sentence alignment of the parallel corpus is often used in the parameter estimation of the SMT (Brown et al., 1993). However, investigation was carried out to incorporate word-level alignment into the SMT to achieve a significant improvement in the quality of both the alignment and translation (Callison-Burch et al., 2004). In the phrase-based statistical translation, alignment is used in the phrase / rule extraction too (Watanabe et al., 2006). The quality of the alignment is extremely important in a statistical or hybrid MT system. Some research has been done to combine the output from a few alignment systems to improve the performance of the MT system (Ayan et al., 2004).

Last but not least, alignment is used after the MT too. It is used to evaluate a MT system. Ding Liu and Daniel Gildea have proposed a metric based on stochastic iterative alignment to evaluate the performance of the MT system (Liu & Gildea, 2006).

As a conclusion, the alignment is used before, during and even after the MT phase. Hence, this implies that alignment and MT are closely related and inter-dependant.

1.1.4 The need for automatic aligner

As discussed in section 1.1.3, text alignment plays a major role in MT, i.e. prior, during and post MT phase. In particular, it is crucial in the example acquisition stage of the EBMT life cycle and used by the compiled EBMT approach in the translation knowledge acquisition and extraction module. In fact, it is one of the most straight-forward, easiest and least expensive way to gather a large number of examples in view of the free availability of large and multi-lingual parallel corpus compared to generating the examples by linguists. Beside this, the collected examples are more robust and have a wider coverage of different domains and genres of text. In addition, it minimizes the effect of translation preference and style which happens when the examples are generated by a specific group of linguists.

The free availability of parallel texts in large numbers and the fact that a large number of examples are needed to ensure the quality of the EBMT systems have driven the effort to align parallel texts in a large scale in order to gather as many examples as possible. This increases the coverage and robustness of the EBMT system and ensures the consistency. However, large scale text alignment is expensive in terms of time and effort and almost impractical to be done manually. Hence, an automatic aligner is in urgent need to support the large scale text alignment for example extraction in an EBMT system.

1.2 *Research Overview*

In this section, we provide an overview of the research by laying out the problem statements, research objectives, research scope and the contributions of the research.

1.2.1 Problem Statements

As mentioned in the previous section, alignment of parallel texts plays a major role in the example preparation which serves as the core reference of an EBMT system. The first problem is that the alignment task requires much effort, is time-consuming and error-prone due to the current practice of manual and semi-automatic alignment. The second problem to be tackled in this research is the lack of resources both in linguistic tools and linguistic data for Malay language which makes the alignment of Malay parallel texts a major challenge. Here, the challenge is to achieve the same (if not better) level of accuracy as the other resourceful languages.

1.2.2 Research Objectives

The first objective of this research is to propose an approach to automate the word and fragment alignment task and minimize human intervention, thus increase consistency and at the same time decrease the effort, time and errors. The second and main objective of the research is to design a novel alignment algorithm for English-Malay parallel texts that uses minimum available linguistic resources but at the same time achieve a reasonable level of accuracy. Besides, the preparation and construction of language

resources as well as the enhancement of the existing language tools are essential for future research in computational Malay language processing.

1.2.3 Research Scope

The research focuses on parallel text alignment of English-Malay language pairs. Other languages are not considered. Although fragment alignments are provided as a by-product of the system, the focus of the research is on the word-level alignment alone. Other levels of alignment such as sentence, phrase and character alignment are not part of the research scope. Moreover, the research is done purely based on the available Malay resources. No adaptation of tools from other languages is used for Malay in this research. Tool enhancement and construction of linguistic resources are done if necessary.

1.2.4 Contributions

The main contributions of the research are:

a) The design of a simple but novel alignment algorithm for scarce resources.
    - The algorithm is novel based on the hybrid of hapax, cut-through calculation and recursivity.
    - Only basic language resources and tools are needed, which is a stemmer and a bilingual dictionary. Most languages, including the scarce and rare languages, have these tools.
    - The algorithm is simple but effective. It provides reasonably high performance

as complex alignment systems with lots of resources.

b) The development of an automatic alignment system
    - It provides automatic word and fragment alignments.
    - It allows interactive annotation where human intervention is possible to check and rectify the suggested alignment results if necessary.

c) The enhancement of existing Malay linguistic tools as well as preparation and construction of Malay linguistic data.
    - Enhancement of the Malay morphological analyzer to act as a stemmer.
    - Construction of a complete root word dictionary through root word extraction from Kamus Dewan.
    - Synchronization of two bilingual dictionaries (English-Malay and Malay-English bilingual dictionary) into a single source of reference.

1.3 *Thesis Organization*

The thesis is organized into six chapters to present the details of the research. The overview of each chapter is as follows:

CHAPTER 1 sets the stage for the research by laying out the research background, problem statements, research objectives, research scope and contributions.

CHAPTER 2 gives an overview of text alignment in terms of level of alignment and category of approach. It reviews the existing approaches for each level of alignment and ends with the factors to be considered in the text alignment task.

CHAPTER 3 discusses the preparation of the resources including the enchancement of the linguistic tools as well as the construction and synchronization of the linguistic data. It also presents the analysis and results of the resources.

CHAPTER 4 describes the proposed methodology and algorithm which adopt the concept of hapax, cut-through, fragmentation and recursivity. It presents details of each step of the algorithm and provides examples.

CHAPTER 5 presents the experiments and results of three main categories of analysis which are the first round analysis, all round analysis and overall performance. The interpretation of the result is discussed in detail with causal analysis.

CHAPTER 6 sums up the research by revisiting the research objectives and contributions. The thesis concludes with some suggestions for future work.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews the state-of-the-art of text alignment. It starts with the discussion of the initiatives and development of parallel corpus from the view point of multi-linguality, genre and reference alignment. The different levels of alignment and existing approaches are discussed in detail.

2.1 *Parallel Corpus*

Initial research in text alignment focuses on parallel text. There are many initiaves in collecting and building parallel corpora taking into consideration multi-linguality, genre and reference alignment.

In terms of multi-linguility, the ARCADE project, which focuses on text alignment research such as parallel corpus collection, alignment techniques, evaluation metrics and alignment system evaluation, produces a bilingual sentence-aligned French-English corpus (Langlais et al., 1998). While ARCADE I only focuses on bilingual parallel text, ARCADE II extends the corpus collection further, leading to multilingual parallel texts which include 10 language pairs with both western European languages (English, German, Italian and Spanish) as well as distant languages using non-Latin scripts (Arabic, Chinese, Greek, Japanese, Persian and Russian). French was used as the pivot language

(Chiao et al., 2006). The JRC-Acquis corpus, which claimed to be the largest available parallel corpora world-wide to-date with the latest version 3.0, is available in 22 languages with 231 language pairs (http://langtech.jrc.it/index.html#). The Europarl corpus has collected parallel texts from proceedings of the European Parliament in 11 European languages with 110 language pairs (Koehn, 2005). The OPUS corpus is a collection of translated open source documents freely available from the web. It started from the OpenOffice.org documentation in OPUS v0.1, KDE and PHP manuals in OPUS v0.2 and movie subtitles from OpenSubtitles as well as biomedical data from the European Medicines Agency (EMEA) in the latest version of the OPUS corpus (Tiedemann & Nygaard, 2003; Tiedemann & Nygaard, 2004; Tiedemann, 2009). To-date, some of the OPUS sub-corpus is available in up to 92 languages. Other parallel corpora include COMPARA – an English-Portuguese parallel corpus (Frankenberg-Garcia & Santos, 2003), Nunavut Hansard – an Inuktitut-English parallel corpus (Martin et al, 2003), Parasol – a Slavic language parallel corpus (Waldenfels, 2006), a Japanese-English patent parallel corpus (Utiyama & Isahara, 2007) as well as an Indonesian-English parallel corpus built from the ANTARA News, BPPTPANL and BTEC-ATR (Budiono & Hakim, 2009).

Besides multi-linguality, the genre of the corpora plays a major role in the text alignment. The ARCADE I project produces a corpus composed of four main genres – technical manuals, scientific articles, literature and institutional texts which are further divided into three subgenres, i.e. direct translation, indirect translation and speech transcription. The institutional texts are easy to gather and align. They are highly sought

after in alignment tasks. Technical manuals are usually close to the original text but the large proportion of technical terms and greater structural complexity in technical manuals increase the difficulty of the alignment task. Scientific articles have the similar difficulties but with more linear prose and freedom in translation. The last but not least, literature such as novels have the most linear prose and uses free style in translation where dropped, merged or split segments are always observed (Véronis et al., 2000). The ARCADE II includes the MD corpus which collects news articles from the French monthly newspaper Le Monde Diplomatique (Chiao et al., 2006). OPUS corpus includes a sub-corpus of movie subtitles and a sub-corpus on biomedical data (Tiedemann, 2009). Utimaya and Isahara focused the corpus collection on patents (Utimaya & Isahara, 2007) while Resnik et al. use the Bible as the parallel corpus (Resnik et al., 1999). Moreover, research efforts are active in using the Web as parallel corpus and mining it for bilingual text (Resnik & Smith, 2003; Zhang et al., 2006; Shi et al., 2006).

In addition to the multi-linguality and genre, most of the parallel corpora provide reference alignment at different levels to ease the research work. The JRC-Acquis corpus (Steinberger et al., 2006) is paragraph aligned while ARCADE I corpus (Véronis et al., 2000) and Europarl corpus (Koehn, 2005) are sentence aligned. In the later years, researchers tend to develop corpora which provide alignments of smaller units such as phrases and words since the alignment research has moved towards this direction. For instance, the ARCADE II (Chiao et al., 2006) corpus provides not only paragraph and sentence alignments but also named entity phrase alignment. Another example is the OPUS corpus (Tiedemann, 2009) which provides both sentence and word alignments.

2.2 *Two Alignment Approaches*

In general, there are two main alignment approaches which are statistical approach and linguistic approach. This section will only discuss briefly these approaches in terms of their strengths and weaknesses. The details of the methods that follow each approach will be discussed in the next section in their respective levels of alignment.

2.2.1 Statistical Approach

The statistical alignment approach requires no prior knowledge. It is solely based on probability. It chooses the alignment that maximizes the probability over all possible alignments. Dynamic programming algorithm and weighted sum are used in the selection of the most probable alignments. For instance, length-based approach in sentence alignment is one of the statistical approaches.

This approach is robust. Context-dependent usage of words could be extracted and it works well regardless of the word segmentation problem (Haruno & Yamazaki, 1996). Hence, it is language independent. However, full alignment is hard to achieve since the number of word correspondences acquired is limited.

2.2.2 Heuristic Approach

The heuristic approach uses associative measures derived from corpus and external sources such as linguistic resources. Dictionary-based approach is one of the popular

heuristic approaches. It uses intensively the linguistic information and knowledge such as lexicon. The efficiency of the approach depends greatly on the quality of the linguistic resource used. Context-dependent keywords are hardly extracted and it performs poorly with wrong segmentations.

2.3 *Levels of Alignment*

There are four levels of text alignment which are paragraph alignment, sentence alignment, word alignment and character alignment. Among all the levels of alignment, sentence alignment has gained the most attention during the early days of alignment research while word alignment has become the focus of alignment research recently in view of its wide application in the natural language processing domain. Research on other levels of alignment is rare.

2.3.1 Paragraph Alignment

There is not much research which focuses on paragraph alignment since it is often discussed together with sentence alignment. Most of the sentence alignment techniques could be used to align paragraphs. For instance, the baseline alignment method uses the relative position of the paragraph based on paragraph counts or word counts (Gelbukh et al., 2006). The proposed sentence alignment approach based on character length was adopted in paragraph alignment (Gale et al., 1991).

Recently, Gelbukh et al. have proposed a hybrid paragraph alignment method based on dictionary-based similarity measure and anchor point technique to align English-Spanish parallel corpus (Gelbukh et al., 2006). They further formalize the paragraph alignment task as an optimization problem of a bipartite hypergraph using lexical-based distance measure and dynamic programming algorithm (Gelbukh & Sidorov, 2006; Gelbukh et al., 2007). Esteva and Xu have proposed a paragraph alignment method based on local cosine similarity to retrieve archival bond and find stories in a digital text achive (Xu & Esteva, 2011).

2.3.2 Sentence Alignment

Compared to paragraph alignment, sentence alignment has gained more attention among the parallel text research community. Sentence alignment approaches could be divided into six main categories.

Firstly, the length-based approach is based on the idea that the sentence length of the source text is likely to be similar to the corresponding sentence length in the target text. It is purely statistical and knowledge-poor where the word identity and meaning are not taken into consideration. It is more computationally efficient and language independent (Aswani & Gaizauskas, 2005). Brown, Lai and Mercer use word count as sentence length (Brown et al., 1991) while Gale and Church use character count (Gale & Church, 1991).

Secondly, the lexical-based approach uses lexical information and it is based on the idea that a word in the source text is likely to have a corresponding translation in the target text. It is resource hungry and knowledge-rich but computationally expensive. It performs better on noisy texts (Aswani & Gaizauskas, 2005). The word correspondence-based approach is popular among others in this category. For instance, Kay's model (Kay, 1991; Kay & Roscheisen, 1993), IBM Model-1 (Brown et al., 1993), and Melamed's model (Melamed, 1996) use geometric correspondence while Haruno and Yamazaki's model uses both bilingual dictionary and POS tags (Haruno & Yamazaki, 1996).

Thirdly, anchor-based approach finds and matches anchors by different means to align parallel texts. Simard et al. first uses cognates as anchors in aligning sentences (Simard et al., 1992). Cognates are words which are similar across languages in terms of orthography and meaning due to borrowing or common inheritance from the same linguistic ancestor. Wu has proposed an improved version of Gale and Church's model with domain-specific lexical cues as anchors (Wu, 1994). Toth et al. proposed a hybrid alignment method of Hungarian-English corpus that uses named entity as anchors (Toth et al., 2008) while Vosoughpour and Faili use sentence bead as anchors to align English-Persian parallel corpus (Vosoughpour & Faili, 2010).

Next, content-based features such as HTML tag strutures are used in sentence alignment too. Shi et al. proposed a Document Object Model (DOM) tree alignment method which exploits HTML tag structure similarity between parallel web pages (Shi et