Automatic Construction of English/Chinese Parallel Corpus

LI Kar Wing

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master Of Philosophy

in

Department of Systems Engineering and Engineering Management

June 2001


香港中文大學
圖書館系統
22 APR 2002
THE CHINESE UNIVERSITY OF HONG KONG

Automatic Construction of English/Chinese Parallel Corpus

Final Version

By

Li Kar Wing

Abstract of thesis entitled:

Automatic Construction of English/Chinese Parallel Corpus

Submitted by LI Kar Wing

for the degree of Master of Philosophy

at The Chinese University of Hong Kong in June, 2001

The increasing need of access in global information has made multilingual corpora to become a valuable linguistic resource for many natural language processing applications. The general-purpose dictionary is less sensitive in genre and domain. As it can be impractical to manually construct tailored bilingual dictionaries or sophisticated multilingual thesauri for large applications, corpus-based approaches provide a statistical translation model to cross the language boundary.

Many domain-specific parallel or comparable corpora are employed in machine translation and cross-lingual information retrieval. Since the number of Asian/Indo-European corpus, especially English/Chinese corpus is relatively deficient, our aim is to automatically construct Chinese/English parallel corpus.

Several research projects for automatic construction of parallel corpus from the World Wide Web are discussed. To identify two texts is mutual translation of each other, alignment is needed. Many different alignment models are studied to facilitate our construction.

This dissertation will present an alignment method relied on dynamic programming to identify the one-to-one Chinese and English title pairs and construct a parallel corpus by downloading the texts accordingly. The method uses the fact that a title is a representation of a text.

The method includes alignment at title level, word level and character level. The longest common subsequence (LCS) is applied to find the most reliable Chinese translation of an English word. As one word for a language may translate into two or more words in another language, deletion, an edit operation is used to reduce overlapping. After reviewing many score functions in different alignment models, a score function is proposed to find the optimal title pairs.

A system based on the method is implemented to test its effectiveness. The system automatically constructs a parallel corpus by downloading the Hong Kong government press release daily articles. The precision of the result is estimated at 99.8% and recall at 63.7%.

The economic monthly reports, press release articles and speech articles published by Hang Seng Bank are also used to test our method. The precisions are at 100%, 96.52% and 100% respectively. The recalls are at 89.36%, 60.3% and 100% respectively. In addition, the results generated by some other automatic parallel corpus construction systems are also surveyed. Finally, the dissertation concludes with an assessment of the present state of the field and the potential extension of the model.

摘要

因為對全世界資訊的需求不斷上升，令多種語言的文集對自然語言處理系統顯得非常重要。普通的字典並不包含一些特定類型和範圍的專用名詞，而以人手去編寫較大的雙語字庫或多語辭典又非常費時， 因此以文集為基礎的統計翻譯方法卻可以彌補這些不足。

很多專用的對照或相關文集被用在機器翻譯和雙語搜索方面。因為亞洲/印歐文集比較少，尤其是中/英文文集更為少見。我們的目的是建立一個自動化收集中/英文對照文集系統。

我們參考了不少有關利用互聯網自動收集對照文集的系統。當要証明兩篇文章是互譯的時候，我們需要使用列隊。我們也對不同的隊列方法做了深入的研究。

本論文將展示以動態規劃為基礎的列隊方法，去找出一對一中/英文標題，然後將與這對標題相連的文章載入文集裏。 這個方法是利用標題可以概括一篇文章的好處。

我們的列隊方法包括標題、詞、字等方面的列隊。最長共同子序列被應用在尋找一個英文詞最適當的中文翻譯。因為一個詞可能翻譯成幾個詞，所以重複的問題會出現。而一種編輯程序　　刪除，被用來解決重複的問題。在參考很多有關計算分數的方法之後，我們研究出一個新的計分方法去找出最適當的標題對。

我們的列隊系統當用於收集香港政府新聞公布文章時，準確率達到99.8%，而回起率達到 63.7%。這種方法也用於收集登在恆生銀行網頁上的經濟月報、新聞稿和演講詞，準確率分別是 100%， 96.52% 和 100%，而回起率分別是 89.36%，60.3%和 100%。最後，本論文總結了在翻譯方面的最新科技和我們將來的發展方向。

Acknowledgements

I would like to express my hearty gratitude to my supervisor, Dr. Christopher C. C. Yang for his valuable comments and suggestions on my research and thesis. I also would like to express my hearty gratitude to my parents for their care so that I can concentrate on my work. Finally, I want to thank all my friends who accompany and support me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The increasing need of access in global information has made multilingual corpora to become a valuable linguistic resource for many natural language processing applications. The general-purpose dictionary is less sensitive in genre and domain. As it can be impractical to manually construct tailored bilingual dictionaries or sophisticated multilingual thesauri for large applications, corpus-based approaches provide a statistical translation model to cross the language boundary.

Apart from their use in machine translation (e.g. [BCD90], [BLM91]), corpus-based models are applied in machine-assisted translation (e.g.[FIP96]), cross-lingual information retrieval (e.g.[DD95], [SB96], [NSI99]) and bilingual lexicography (e.g. [KT90]).

A parallel corpus contains many text pairs. Two texts in each pair are mutual translation of each other. The problems connected with parallel corpus are often usage fees, licensing restrictions and out of dated texts. An idea for overcoming the difficulties is to continually download the parallel articles from different bilingual web sites. Since the articles are dynamic resource, they can complement the gap between the introduction of a new term and its incorporation into a standard reference work such as dictionary.

As a result, **the objective of our research** is to build a system which can automatically construct a parallel corpus.

1.1 Application of Bilingual corpora

One of the main problems in human communication is the presence of a huge variety of written and spoken languages in the world. In the time of world-wide communication systems and international relations, it is becoming more and more important to find ways to support the connection and communication of people from different ethnic parts of the world. To keep pace with the development of new documents, one tries to speed up the translation processes by using corpora.

The corpus-based techniques can be widely used into a variety of applications such as text retrieval, system evaluation, machine translation and speech recognition. The corpus-based techniques emphasize statistical analysis over linguistic theory which led some successes (e.g. [GC91]). Corpus-based approaches can be viewed as a type of automatic thesaurus construction techniques where the relationship between terms is obtained from statistics of term usage ([OD96]). The approaches involve the use of term co-occurrence statistics across large document collections for the construction of domain-specific translation techniques.

Corpus-based approaches are commonly used in machine translation and cross-lingual information retrieval. In this section, we will review the applications of bilingual corpora in machine translation(Section 1.1.1) and cross-lingual information retrieval(Section 1.1.2).

1.1.1 Machine translation

The term machine translation (MT) is normally taken in its restricted and precise meaning of fully automatic translation. However, by considering the integration of other language processing techniques and resources with MT, Maegaard et al.

([MBD99]) define Machine Translation to include any computer-based process that transforms (or helps a user to transform) written text from one human language into another. In addition, Maegaard et al. ([MBD99]) define Fully Automated Machine Translation (FAMT) to be MT performed without the intervention of a human being during the process. Human-Assisted Machine Translation (HAMT) is the style of translation in which a computer system does most of the translation, appealing in case of difficulty to a (mono- or bilingual) human for help(e.g. [FIP96]). Machine-Aided Translation (MAT) is the style of translation in which a human does most of the work but uses one of more computer systems, mainly as resources such as dictionaries and spelling checkers, as assistants(e.g. [FM97], [Mel96aa]).

There are two main directions in producing machine translation systems: *linguistic* and *statistical.*

1.1.1.1 Linguistic

The first approach is to develop analysis and translation tools by defining grammatical, rule-based structures with the use of expert knowledge and theoretical research from the past in connection with available general multilingual dictionaries. Although much research is done in analyzing syntactic and semantic structures of many languages(especially European languages), it is very difficult with the present methods to fully describe natural languages with the general formal language definitions. The problems like the context dependency and ambiguity of natural languages normally exits ([SFI92], [As99], [FM97]). Furthermore, natural languages represent evolutionary systems. There exists no fixed prescriptive regulations for the usage of today's spoken languages. Every collection of rules and word descriptions should be considered as a collection of usage recommendations in this language. Additional problems appear in the context of translating expressions from one language into another. Because of different ambiguous expressions in different languages, it is even harder to decide which expression in the target language corresponds to the source expression in this context.

1.1.1.2 Statistical

The second approach is to analyze available parallel texts and to extract information from these text corpora which have been translated previously by hand. Not only multilingual analysis but also monolingual research can be based on available text corpora. The advantage of processing a text corpus is to obtain context specific information about syntactic structures and usage of words in a given language. In the case of parallel corpora, one can obtain context-specific correlations between these languages, which are usually much less ambiguous than general collections. Resulting data from these corpus analysis processes can be used to develop context-specific tools for translation and to standardize the usage of structures and word sets for future multilingual document production.

Furthermore, combinations from these two research directions represent a reasonable approach. Combinations of formal, general language descriptions, and data resulting from a corpus analysis process represent a good working base for machine translation systems([UIY94], [KT96]).

1.1.1.3 Lexicon construction

The component of a machine translation system which encodes domain knowledge is typically referred to as a lexicon ([OD96]). One prerequisite in statistical machine translation is the availability of multilingual lexica with domain specific data. In the last few years, many projects in the field of bilingual lexicon extraction from parallel corpora have been initiated at different places. The majority of these projects concentrate on the compilation of general bilingual lexica from large parallel corpora. Many different techniques in the area of statistics and linguistic analysis are used to extract lexical data from different kinds of corpora. There are approaches for

processing clean or noisy parallel texts in historically related languages as well, as in less related language pairs like Asian-European pairs([UIY94], [FM97]).

Lexicon is a linguistic resource for many natural language processing systems. They include the vocabulary that the system can handle, both individual lexical items and multi-word phrases, with associated morphological, syntactic, semantic and pragmatic information([PCC99]).

There is increased recognition of the vital role played by lexicons (word lists with associated information), when fine tuning general systems to particular domains(e.g. [RM97]). In addition, as large-scale general lexicons with different layers of encoded information (morphological, syntactic, semantic, etc.) are created, lexicons will still need to be fine-tuned for use in specific applications.

General and domain-specific lexicons are mutually interdependent([UIY94]). Due to the ever-changing nature of language, no general lexicon can be adequate. Integration of different types of linguistic resources , e.g. bilingual corpora, can provide enhanced capability and coverage for general lexicon. This view sees the two as complementary in a more comprehensive perspective. Also, large corpora represent the apparently 'irregular' facts (evidenced by corpus analysis), and provides the divergences of actual usage([PCC99]). In the past few years, steps towards this objective have been taken by many research efforts which aimed at acquiring linguistic and, more specifically, lexical, information from corpora ([UIY94], [CKJ99], [KT96]).

1.1.2 Cross-lingual Information Retrieval

Cross-lingual Information Retrieval (CLIR) refers to the ability to process a query for information in one language, search a collection of objects, including text, images, audio files, etc., and return the most relevant objects, translated if necessary into the user's language([Oar97], [KHF99]).

There are currently two approaches used in cross-lingual text retrieval: controlled vocabulary and free text.

1.1.2.1 Controlled vocabulary

The first approach required that the documents be manually indexed using a predetermined vocabulary. The user is required to express the query using terms drawn from the same vocabulary. This is referred to as a controlled vocabulary approach ([Oar97]). Systems exploiting such approach use a multilingual thesaurus to relate the selected terms from each language to a common set of language-independent concept identifiers. Document selection is based on concept identifier matching.

For a skilled user, who is familiar with controlled vocabulary search techniques, can effectively retrieve the relevant documents. However, the requirement to manually index the document collection makes controlled vocabulary text retrieval techniques unsuitable for large-volume applications. Large volume applications such as World Wide Web search engines in which the documents are generated from diverse sources that are not easily standardized. In addition, an untrained user is hard to select relevant search term in controlled vocabulary text retrieval. The limitations

have motivated the search for approaches which are amenable to less well structured situations.

1.1.2.2 Free text

The alternative to controlled vocabulary is referred as free text retrieval. It uses the words that appear in the documents themselves as the vocabulary. Two approaches of cross-language free text retrieval have been emerged: corpus-based approaches and knowledge-based approaches. These two approaches are not mutually exclusive and the trend in cross-language free text retrieval research is to combine two approaches to maximize retrieval effectiveness. Figure 1 ([Oar97]) illustrates the taxonomy about different corpus-based and knowledge-based approaches.

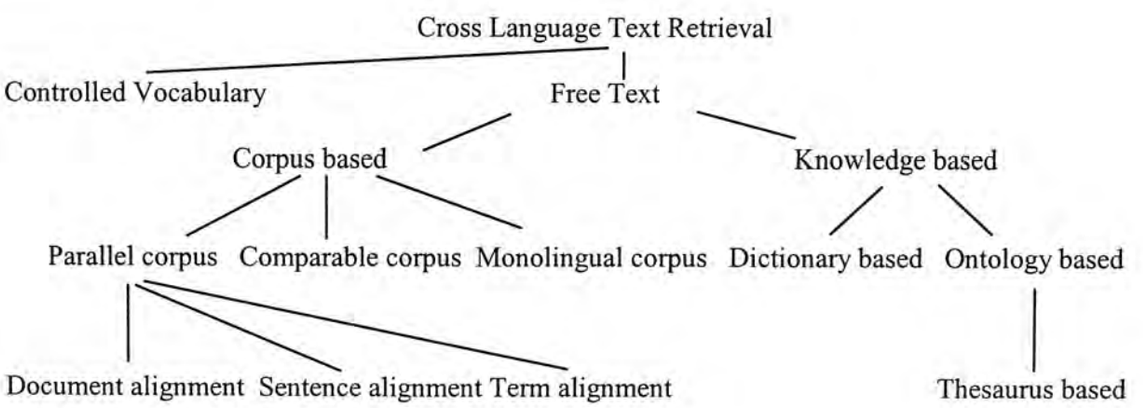

Figure 1: Cross-lingual Text Retrieval Approaches

Unlike controlled vocabulary which requires both the documents and the queries to translate into a common language, free text retrieval applications mostly translate only the query or only the documents. In most applications, only the queries are translated because queries are normally short and it would be more efficiently to deal with.

An ontology can be viewed as an inventory of concepts, organized under some internal structuring principle. Ontologies contain the semantic information that enables language processing systems to deliver higher quality performance. They help with a large variety of tasks, including word sense disambiguation, machine translation([PCC99]).

Using multilingual thesauri in controlled vocabulary text retrieval is referred as one knowledge-based approach in free text retrieval. A multilingual thesaurus, is one type of ontology, organizes terminology from more than one language. Complex thesauri, which encode syntactic and semantic information about terms, used as a concept index in automatic text retrieval systems ([OD96]).

Dictionary-based approaches extend the basic idea of a multilingual thesaurus by using bilingual dictionaries in translation process. Bilingual dictionaries, which have the breadth of coverage required by many cross language free text applications, are a common available cross-lingual knowledge structure. Thus, dictionary-based retrieval is widely used in the linguistic-based cross-language text retrieval ([OD96], [Sal70]).

Dictionary-based cross-language text retrieval approaches replace each term in the query with an appropriate term or set of terms in the preferred language. This method causes two limitations for text retrieval. Firstly, many words do not have a unique translation, and sometimes the alternate translations have very different meanings. Monolingual text retrieval systems face similar challenges from homonomy and polysemy. Polysemous words are words which have more than one meaning. The problem is significantly aggravated by translation ambiguity. If every possible translation is used, it can greatly expand the set of possible meanings because some of those translations will introduce additional homonomous or polysemous word senses in the second language. However, the query expansion would make the system or even a human hard to determine the intended meaning from the available context([Oar97]).

The second important concern for a dictionary-based approach is that the dictionary may lack some terms that are essential for a correct interpretation of the query. The reason for this is that either the query deals with a technical topic which is outside the scope of the dictionary or the user has entered some form of abbreviation or slang which is not included in the dictionary. Even though dictionaries specifically designed for query translation are developed, the effect of this limitation is unlikely to be eliminated completely because usage of language is a creative activity. New terms enter the lexicon all the time. There will naturally be a lag between the introduction of a term and its incorporation into a standard reference work such as a dictionary. The lag is also applied to controlled vocabulary systems based on multilingual thesauri as the introduction of a new term may induce the need to describe a new concept that did not previously appear in the document collection.

1.1.2.3 Application corpus-based approach in CLIR

Corpus-based approaches search to overcome the limitations of knowledge-based techniques by constructing appropriate query translation methods used in a domain-specific application. As it can be impractical to construct tailored bilingual dictionaries or sophisticated multilingual thesauri manually for large applications, corpus-based approaches provide a statistical translation model for cross-lingual information retrieval. The approaches involve the use of term co-occurrence statistics across large document collections for the construction of domain-specific translation techniques.

However, purely statistical approaches introduce errors because many corpus-based techniques have emphasized statistical analysis and made little use of linguistic theory. This introduces errors because the statistical approaches use co-occurrence statistics of terms. Words which are not the translation of each other may sometimes

display the same patterns of co-occurrence as words which are translation of each other.

Therefore, there is a trend of combination of corpus-based approaches and knowledge-based approaches for maximizing retrieval effectiveness. Oard ([Oar96]) has developed a technique based on term-level alignment which also offers the potential for integration of dictionary-based and corpus-based techniques. The basic idea is to estimate the domain-specific probability distribution on the possible translations of each term based on the observed frequency of terms in a parallel document collection.

1.2 Overview of linguistic resources

Apart from corpus, there are many different linguistic resources to facilitate language processing. The linguistic resource includes *written* and spoken corpora, lexical databases, grammars and terminologies. The term "linguistic resources" refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems([GZ95]).

An increasing awareness of the potential economic and social impact of natural language and speech systems has attracted attention. Support is gained from national and international funding authorities. A key technical factor in the demand for lexicons and corpora is the statistical techniques such as hidden Markov models (HMM) (e.g.[FIP96]) and neural networks (e.g.[LC96]), which learn from large data sets organized in terms of many variables with many possible values.

However, many approaches based on parallel document collections are required to obtain a suitable document collection before their statistical model can be applied. Such collections are often difficult to find, and very expensive to prepare ([GZ95]).

So the collections are mostly available in highly specialized application domains, e.g. Canadian Parliament debates([LDC]), United Nation documents ([LDC]), European Community documents ([ELD01]), news articles ([LDC]) and religious texts ([ROD99])..

With the effort of the Linguistic Data Consortium (LDC) in the U.S., the European Language Resources Association (ELRA) in Europe, together with many organizations and projects([PCC99]), the problems of building, collecting and disseminating reliable multilingual resources are relieved.

In Japan, the Electronic Dictionary Research (EDR) Institute has created a large monolingual Japanese and English dictionaries, together with bilingual links, a large concept dictionary and associated text corpora ([EDR]).

Chinese counts one-fourth of the world's population. The population of Chinese information in the world is unimaginably large. The increasing need of Chinese/English bilingual corpus for research purposes has made a few Chinese/English corpora available, such as HKUST English/Chinese Bilingual corpus ([Wu94]). However, the HKUST English/Chinese Bilingual corpus only contains the bilingual documents of Hong Kong Legislative Council debates. The parallel documents of Legislative Council debates are analogous to the bilingual texts of the Canadian Hansard([GC91]). In addition, the documents needed to be converted into machine readable form by special(both manual and automatic) arrangement before putting into HKUST English/Chinese Bilingual corpus([Wu94]).

In Hong Kong, even there is no any organization responsible for collecting and disseminating corpora (spoken and written) and other information to research projects and companies, many Web sites hosted in Hong Kong provide documents in both English and Chinese.

However, like other linguistic resources, many parallel web pages and corpora are often encumbered by fees or licensing restrictions which are not able to freely access. The Hong Kong government SAR has facilitated the construction of language processing resources by publishing many governmental, legal and financial documents on the Web for public access, such as press release articles. These documents are written in both Chinese and English, which are good sources of corpus. Among these press release documents, the bilingual parliamentary proceedings of Hong Kong Legislative Council are included.

Since there are many press release parallel documents published in the Hong Kong government web sites, these parallel documents are used to test the proposed automatic corpus construction method.

In addition, on-line bilingual newswire articles provide a continuous large amount of data for relieving the lag between the introduction of a new term and its incorporation into a standard reference work such as dictionary. This kind of dynamic resource has been used by many researchers in their language processing works. Some of them have used newswire bilingual texts in their lexicon construction ([TSB97], [MM98]).

We have also applied the proposed corpus construction method on economic monthly reports, press release articles and speech articles published by Hang Seng Bank. A detailed discussion of the results will be shown at the Chapter 5.

1.3 Written language corpora

Written Language corpora, collections of text in electronic form, are being collected for research and commercial applications in natural language processing (NLP)([EC95]). Written Language Corpora have been used to improve spelling correctors, hyphenation routines and grammar checkers, which are being integrated

into commercial wording packages. Lexicographers have used corpora to study word use and to associate uses with meanings. Statistical methods have been used to find interesting associations among words(collocations). Terminologists are using corpora to build glossaries to assure consistent and correct translations of difficult terms such as dialog box, which is translated as finestra 'window' in Italian and as boite 'box' in French)([EC95]).

Written language corpora provide a spectrum of resources for language processing, ranging from the corpora themselves to computational grammars and lexicons. Between these two extremes are intermediate resources like annotated corpora (also called tagged corpora in which words are tagged with part of speech tags and other information)([Mel98]), tree banks (in which sentences are analyzed syntactically) ([CCF99]) and part-of-speech taggers.

An example of tree banks was created by Penn TreeBank Project ([MSM93]). The project created and distributes syntactic parse trees for approximate one million English sentences of various genres.

1.3.1 Types of corpora

Different types of corpora have been constructed for language processing. They reflect the criteria according to which they are designed and the purpose for which they are created.

In the field of translation, one type of corpus is the multi-source-language monolingual "comparable" corpus, consisting of two sets of texts, one originally written in language A and one of similar texts translated into language A from a variety of different languages ([Zan98]). Its value is mainly theoretical, what is investigated is the linguistic nature of translated text, independently of the source language.

Another kind of corpus is the bilingual (or multilingual) corpus. Language pairs are put together either on the basis of "parallelism" or "comparability." Parallel bilingual corpora consist of texts in language A and their translation into language B. The relationship between texts is directional, i.e. it goes from one text to the other text([Zan98]). This type of relationship is known as **overt translation**([Ros81]). Texts in parallel bilingual corpus is called bitext ([Mel97], [Mel97b]). A bitext comprises two versions of a text, such as a text in two different languages. Translators create a bitext each time they translate a text.

In addition, one can image that a bilingual document pair expresses the same content in two different languages, e.g. commentaries on a sports event broadcast live in several languages. A collection containing such bilingual document pairs is also qualified as a parallel corpus([Ebe98]). The type of relationship between the bilingual documents is called **covert translation**([Leo00]). House defines covert translation as "...a translation which enjoys or enjoyed the status of an original ST (source text) in the target culture"([Gut00]). She calls this type of translation 'covert' because ".. it is not marked pragmatically as a TT(translated text) of an ST(source text) but may, conceivably, have been created in its own right"([Gut00]). In covert translation, the documents are not bound to a specific culture. It is the same as there are a single text in two or more languages([Ros81]).

Comparable bilingual corpora consist of texts in the languages involved, which share similar criteria of composition, genre, and topic.

Bilingual corpora have been used for lexical acquisition (e.g. [GC91]), machine-aided translation (e.g. [FM97]) and cross-lingual information retrieval(e.g. [DD95], [SB96], [NSI99]). Much recent research in Machine Translation aims not so much to create a system able to perform the job of translating a given text automatically, but to implement computerised tools to assist human translators in their work([Zan98]). Parallel corpora can also be treated as "translation memories," from

which translators can retrieve chunks of translated language in order to speed up their work and ensure accurate and consistent translations([IDF93]). Parallel and comparable bilingual corpora have also been used for language learning and the training of translators ([Ba96]).

As regards parallel corpora, it has been observed that translated texts cannot represent the full range of linguistic possibilities of the target language and that they may reflect the stylistic idiosyncrasies of the source language and of individual translators ([PP97b]). However, the comparison between large numbers of texts and their acknowledged translations can show how equivalence has been established by translators under certain circumstances and provide examples of translation strategies. If such corpora are sufficiently varied and large, looking at recurring linguistic choices made by translators allows general patterns to be perceived. Learners can thus notice the translation pattern and generalize from the aggregation of sets of individual instances.

The other type of bilingual corpus is the comparable corpus which can be defined as a collection of texts composed independently in the respective languages and put together on the basis of similarity of content, domain and communicative function([Oar97]). Criteria for creating comparable corpora depend on the homogeneity of texts, both with and across languages, in terms of features such as subject domain, author-reader relationship, text origin and constitution(i.e. "single" or "joint" texts), factuality, technicality and intended outcome(i.e. communicative function). The practice of collecting texts in different languages on the basis of similarity of type, content or function was common in translation research([Zan98]). Zanettin([Zan98]) has used newspaper articles, medical literature and tourist brochures and guides to create some small comparable corpora.

1.3.2 Limitations of comparable corpora

Although there are more comparable documents than parallel documents, it takes time to design the criteria for constructing comparable corpora. In addition, a limit of comparable corpus is the difficulty of generating hypotheses of possible translations. We must rely on known or suspected equivalences as heuristics to retrieve similar contexts in another language, providing a specification which is both sufficiently general to recall a range of possibilities, and sufficiently precise to limit the number of spurious hits ([As99]).

We must then verify that the citations retrieved are in fact sufficiently similar to those in the source language. These procedures are both time-consuming and error-prone: an expression in the one text may occur in a similar context to one in another language text but in fact mean something different. An example is shown in Aston ([As99]). In the example, someone without deep medical knowledge initially assumed that the phrase *loop ileostomy* in a medical research article is equivalent to *ileostomia su bacchetta* in Italian. In fact, they are not equivalent.

Greater certainty as to the equivalence of particular expressions can be obtained by using parallel corpora, consisting of original texts and their translations. If the corpus is aligned, the user can locate all the occurrences of any expression along with the corresponding sentences in the other language. Since parallel concordances provide translations of a word, risks of misinterpretation associated with comparable corpora are diminished. As a result, we mainly focus on translation equivalence in the dissertation. We extended the proposed English/Chinese alignment algorithm in comparable corpus construction in the future.

1.4 Outline of the dissertation

In the chapter, Section 1.1 contains the analysis of corpus-based techniques in different language processing work. The goal is to identify the importance of corpus in language processing. We also describe the two major applications of bilingual corpora: machine translation and cross-lingual information retrieval (Section 1.1.1 and 1.1.2, Chapter 1).

In addition, the chapter includes a description of different kinds of linguistic resources and the development of the linguistic resources in various countries(Section 1.2, Chapter 1). As we emphasis on written parallel corpus, a brief introduction of written corpora and types of written corpus are presented(Section 1.3, Chapter 1).

In the reminder of the dissertation, we divide our research into four parts.
1) The first part contains the study of various methods for automatic corpus construction(Section 2.1, Chapter 2). The chapter also includes a description of related research work such as translation alignment and alignment of sequences. As the translation alignment(Section 2.2, Chapter 2) is strongly related to a number of other sequence comparison problems, with applications in various domains, different alignment techniques relied on dynamic programming algorithm (DPA) are presented(Section 2.3, Chapter 2).

2) The second part includes a description of the related techniques used in the Chinese/English alignment model. The characteristics of translation together problems during alignment are analyzed to facilitate the alignment process(Section 3.1, Chapter 3). For efficient bilingual corpus construction, we concentrate the alignment on title pairs. We start the alignment from character level to word level(Section 3.2, Chapter 3). Since the model relies on dynamic programming algorithm (DPA) to find the optimal alignment of two sequences, the notion of

longest common subsequences and edit distance is introduced to facilitate the optimal alignment(Section 3.3, Chapter 3).

3) The third part contains the English/Chinese alignment model at title level. After comparing with other score functions in the field of translation alignment(Section 4.1, Chapter 4), the proposed score function is presented(Section 4.2, Chapter 4).

4) The last part includes the experimental results and direction of future work(Chapter 5). A conclusion and contribution are given in Chapter 6.

Chapter 2

Literature Review

Before corpora are suitable for language processing, it is necessary for them to be created and prepared (or "annotated")([PCC99]). The term "corpus annotation" can be divided into three broad categories:

1. *identification and markup of logical structure*, usually signaled by typography, such as section breaks, paragraph breaks, footnotes, titles and headings, etc. Mostly, corpora are either created from scratch (via OCR scanning)(e.g. [KR99]) or are obtained from publishers and other bodies who have already rendered the text (e.g. [LDC]) or texts in electronic form. In the latter case, the texts are typically encoded for typographic format in a word processor format. Otherwise, a process is required to render the materials in an encoding format suitable for use for language processing. In both cases, processing is required to introduce markup that identifies logical structure.

2. *identification of more specific elements* present in the text, sometimes signaled typographically, but which usually require some additional processing to identify as well as human intervention for verification. Such elements include sentences, quotations and sub-paragraph elements such as names, dates, etc.

3. *addition of analytical information* in the text, such as part of speech tags, alignment information, prosody markup, syntax, discourse elements, etc. Among the three categories of markup, this is the most costly in terms of processing time and effort.

In order to enable more efficient and effective creation of corpora for language processing, it is essential to understand the nature of each of these phases and establish mechanisms to accomplish each category. Category 1 can be nearly fully automated, but categories 2 and 3 may require more processing overhead as well as human intervention(e.g. [RM97]). Algorithms and methods are needed for automating these two categories. Category 3 has received more attention, since algorithms for identifying complex linguistic elements have typically been viewed as a more legitimate area of research. One of the most important directions for corpora annotation is determining a richer level of annotation that includes word senses, predicate argument structure, noun-phrase semantic categories, and coreference ([PCC99], [Mel97a],[Mel97b]).

In this chapter, some algorithms in automatic construction parallel corpus from category (1) to (3) are discussed (Section 2.1). As category (3) has received more attention and it is important for automatic corpus construction, the chapter also includes a description different translation alignment methods(Section 2.2). Since the multilingual alignment problem is related to a number of other sequence comparison problems in various domains, in particular, sequences comparison in DNA molecules ([Sim99]). The methods used to attack these problems are very similar to those used in translation alignment, and rely largely on dynamic programming. An introduction to these problems is presented (Section 2.3).

2.1 Research in automatic corpus construction

To automatic construction of bilingual corpus, Resnik ([Res98]) has described a technique to explore the World Wide Web for parallel corpus resources called Structural Translation Recognition for Acquiring Natural Data (STRAND).

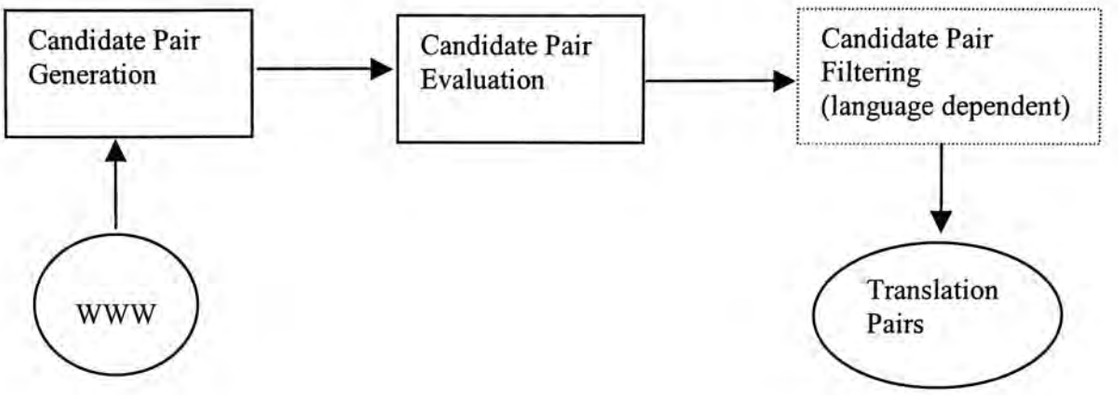

Figure 2: The STRAND architecture

The architecture of STRAND is organized as a pipeline, beginning from a candidate generation module, a candidate evaluation module to candidate pair filtering module. After exploiting a candidate generation module to identify World Wide Web pages that may contain parallel translations, a language independent candidate evaluation module was used to keep those candidate pairs that are likely to be translations. An optional candidate pair module applies to add filtering criteria on language-specific resources. The end result is a set of candidate pairs that can reliably be added to a parallel corpus.

STRAND: Candidate generation module

Resnik([Res99]) noticed that if a web page has been written in many languages, the parent page of the web page will contain the links to different versions of the web page. For example, in a web page, there are two anchor texts such as a1 and a2. a1 is linked to the language 1 version of the web page and a2 is linked to the language 2 version of the web page. It may contain other information as well ([Res98]). An "**anchor**" is a piece of HTML document that encodes a hypertext link. It typically includes the URL of the page being to and the user can click on to it and go there([Res98]).

In addition, a1 and a2 are no more than 10 lines apart in the parent page. The distance criterion captures the fact that for most web pages that point to parallel texts, the anchor texts appear close together. This phenomenon is used as in candidate generation module in STRAND as the selection criterion.

The another criterion is that if a web page got a translation, there is usually a link from the page to its translation in another language and the link in the web page often indicates the language of the linked document. An example, in an English document, if there is a link like "Click here for Spanish version" , "Español" or a Spanish flag, then this link points to the Spanish verison of the document. On the opposite direction, in the Spanish document, there is also a link anchored by "Click here for English version" or "anglaise" which links the Spanish document to its English translation. The pages are called "**sibling**" pages. As a result, if in a web site, there are documents containing links to a document in the same site with one of the anchors in both direction, then the two documents are considered as a candidate parallel pair.

To generate a large parallel corpus, Altavista's "Advance Search" is exploited. By given a particular pair of languages of interest, a candidate generation module uses the pair of languages as a query to the Altavista search engine, asking for the web pages to meet the selection criteria above. One advantage of using Altavista is the large size of its index.

STRAND: Candidate evaluation module

This module tries to determine whether or not two pages should be considered parallel translations. Since the candidate generation module provides a list of candidate parallel pages, the candidate evaluation module filters out bad candidate pairs by employing a structural recognition algorithm. The algorithm assumes that web pages in parallel translation are similar in the way they are structured.

The structural recognition algorithm first runs the bilingual document pairs through a transducer that reduces each text to a linear sequence of tokens corresponding to HTML markup elements. For example, the transducer would replace the HTML source text <TITLE>ACL'99 Conference Home Page</TITLE> with the three tokens [BEGIN:TITLE], [Chunk:24], and [END:TITLE]. The number inside the chunk token is the length of the text chunk, not counting white space.

Since the algorithm assumes that there is a linear relationship in the lengths of text translations ([GC91]), it works by using pieces of identical markup as reliable points of correspondence and computing a best alignment of markup and non-markup chunks between the two documents by applying dynamic programming. It then calculates the correlation for the lengths of the non-markup text chunks by exploiting Pearson correlation coefficient. A test for the significance of this correlation is used to decide whether or not a candidate pair should be identified as a parallel text pair.

As the STRAND seeks to be fully language independent, scalable and automatic creation of parallel corpus, the candidate pair filtering module is treated as language-dependent process and the module is initially avoided. The precision rate up to the candidate evaluation stage is at 88.2% and recall rate at 62.5% which showed that the technique could efficiently create a parallel corpus without human intervention.

STRAND: Candidate pair filtering module

For the pairs of pages that come from on-line catalogues or other web sites having large numbers of pages with a conventional structure, the evaluation module cannot handle such case. To make sure that two pages are actually written in the languages of interest, linguistic knowledge is needed for automatic language identification and content-based comparison of structurally aligned document segments. Such linguistic knowledge can be gained by using cognate matching or bilingual

dictionary. In the candidate pair filtering module, statistical language identification based on character n-grams was applied to the system([Dun94]).

The statistical language identification process adopts the criterion : given two documents d1 and d2 and supposed that the documents are written in language L1 and L2 respectively, the document pair is kept if the words in d1 look more like L1 than L2 and the words in d2 look more like L2 than L1.

Using this statistical process introduces the need for language-specific training data for the two languages of interest. The disadvantage for this is that the training data for the languages of interest may be not available. Apart from the language of interest, training data for other languages is also needed because the web pages may be written in a language other than the languages of interest. For example, if a page is written in Dutch and our languages of interest are English and French, we need the training data to filter out the Dutch pages from the English and French web pages.

Other automatic parallel corpus construction system

Nie et al. ([NSI99]) developed a similar system to automatically extract parallel texts from the Web. If a document A contained a link to another version of the text B at the same site, and the linked document B also has a link to the document A, then the site is candidate site of parallel texts. For example, an "en français" anchor is used in the English version document to connect the French document, and the French document at the same site has linked the English document through the "in English" anchor. After a list of candidate web sites were extracted, the selection of parallel texts based on the difference of their filenames and the name of directories they were stored. To improve the selection process, the length of texts was taken into account.

2.2 Research in translation alignment

According to Simard et al.([SFI92]), given a text and its translation, an alignment is a segmentation of two texts such that the nth segment of texts is the translation of the nth segment of the other. Empty segments are allowed which can be either correspond to translator's omissions or additions). In other words, alignment is the process of finding relations between a source text and its translation. An alignment may also constitute the basis of deeper automatic analyses of translations. For example, it could be used to flag possible omissions in a translation, or to signal common translation mistakes, such as terminological inconsistencies.

We can view alignments as mathematical relations between linguistic entities([Sim99]) :

Given two texts, A and B, seen as sets of linguistic units: and , a binary alignment X_{AB} is defined as a relation on $A \cup B$:

$$X_{AB} = \{(a_1, b_1), (a_2, b_2), (a_3, b_3), ...\} \quad \text{.....(1)}$$

The interpretation of X_{AB} is: (a, b) belongs to X_{AB} if and only if some translation equivalence exists between a and b, total or partial.

This definition of alignment used by Simard ([Sim99]) was inspired from Kay and Röscheisen([KR93]).

Translation alignment can be viewed at different levels of resolution, from the level of documents to those of structural divisions (chapters, sections, etc.), paragraphs, sentences, words, morphemes and eventually, characters.

As mentioned in Section 2.1, translation alignment technique has been used in automatic corpus construction to align two document segments based on their length

similarity. In the past few years, there has been a growing interest in parallel text alignment techniques. These techniques attempt to map various textual units to their translation and have been proven useful for a wide range of applications and tools.

An example of such a tool is the TransSearch bilingual concordancing system([IDF93], [SFP93]). The system allows translators to submit queries to an archive of existing translations in order to locate ready-made solutions to all sorts of translation problems.

Another example is the system called Champollion([SMH96]) which focuses on the identification of collocations in text corpora and on the automatic production of corresponding translations for a given parallel bilingual corpus. The goal of the system is to compile lexical data above the word level from parallel texts. In Champollion, a collocation compiler called Xtract was used for the task of collocation identification. Referring to the American Heritage® Dictionary of the English Language, a **collocation** is "an arrangement or juxtaposition of words or other elements, especially those that commonly co-occur". In Champollion, French words which were highly correlated with their English collocation were determined using Dice coefficient and the translation formed by identifying combinations of the French words which were highly correlated with the source collocation.

Xtract distinguishes between 'fixed' and 'flexible' collocations. Fixed collocations are rigid noun phrases without interruption between the words like 'United States'. Flexible collocations can be interrupted by other words, or they occur with a different order of their words. An example is the collocation 'make decision' which can appear as 'make a decision' or 'decisions to make' ([Sma93]).

Since automatic corpus construction is closely related to research on alignment of parallel texts and research on acquisition of bilingual lexical information for the use by a translator or translation system, we first present previous work on sentence alignment and show how the notion of sentence alignment is extended to alignment

of segments. Then we turn to discuss alignment at the word and term level, which is very useful for acquisition of lexical information([LSV98]).

2.2.1 Sentence alignment

For aligning sentences of texts, Gale and Church relies on the hypothesis that sentence translations are done using one of six "translation patterns"([SFI92]) : (1)One sentence translates into one. (2) Two consecutive sentences translate into one. (3) One sentence translates into two. (4) Two sentences translate into two. That is : the first sentence of language A and the first sentence of language B are not mutual translations, nor are the second sentence of language A and the second sentence of language B, but together, the first and second sentences of language A constitute a translation of the first and second sentences of language B. (5) A sentence is not translates at all or (6) a new sentence that has no equivalent in the source text is introduced by the translator.

There are two main approaches to sentence alignment, namely text-based and length-based alignment. The text-based approaches use linguistic information in the sentence alignment and the length-based make use of the total number of characters or words in a sentence ([FM97]).

Length-based algorithms make the assumption that the sentences, which are mutual translations in the parallel corpus, are similar in length([GC91]). In Brown et al ([BLM91]), sentence alignment is based solely on the number of words in each sentence. Gale and Church ([GC91]) developed a similar algorithm except that instead of basing alignment on the number of words in sentences, alignment is based on the number of characters in sentences. These approaches based exclusively on sentence lengths work quite well with a clean input, such as the Canadian Hansards and have been widely used by other researchers e.g. Resnik([Res98]), Chen et al([CKJ99]), Wu([Wu94]). However, for cases where sentence boundaries are not

clearly marked, such as OCR input([Chu93]), or where the languages are less parallel, such as Asian-European language pairs([Wu94], [Fun95]), these algorithms do not perform well.

Text-based algorithms use lexical information across the language boundary to align sentences. Warwick-Armstrong and Russell ([WR90]) used a bilingual dictionary to select word pairs in sentences from a parallel corpus and then align the sentences based on the word correspondence information. Wu([Wu94]) modified Gale and Church's ([GC91]) length-based statistical method to the task of aligning English sentences with Chinese sentences with the support of a small lexicon. In Chen's ([Che93]) Estimation-Maximization(EM)-based parameter estimation model, lexical information was used to align sentences. To train the model, Chen used one hundred manual aligned sentence pairs as the initial training data.

Another type of lexical information which is helpful in alignment of European language pairs, is called **cognate**([SFI92]). **Cognates** are pairs of tokens of different languages which share obvious phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. The pairs *generation/génération* constitute typical examples for English and French. Simard et al.([SFI92]) illustrated that cognate provides a reliable source of linguistic knowledge. By combining cognateness and length-based algorithm developed by Gale and Church([GC91]), the result was better than alignments based on length alone.

2.2.2 Word alignment

According to Simard et al.([SFI92]), an alignment that simply puts sentences in correspondence would be considered a "gross" alignment, compared to one that shows word correspondences. Alignment can be done at different levels of resolution, from the level of documents to those of structural divisions (chapters,

sections, etc.), paragraphs, sentences, words, morphemes and eventually, characters([Sim99]). Given a resolution, a correct alignment should be "maximal", i.e. it should be composed of the smallest possible pairs([SFI92]). Many algorithms produce alignment at word or character level. Some of these algorithms use sentence-aligned parallel texts further to compile a bilingual lexicon or tool for translator ([GC91], [Ku93], [WX94], [FM97], [CKJ99]).

To do alignment at word level is a challenging task because one word may represent by morphological or syntactic phenomena rather than other words([KR93]). Also, a good translator does not always translate a word the same way every time it occurs([KR93]). Addition and omission normally appear during translation([SFI92]). In He([He00]), Nida noted that the amount of alteration in translation is affected by two factors: one was the type of texts, which could be different in style and content, and the other was the genius of the translator, who should have a thorough understanding of the text being translated. The alteration of information in translation can be classified in five types:1) modification of meaning; 2) inversion of meaning; 3)omission of meaning; 4) addition of meaning; 5) deviation of meaning([He00]). In light of the lexical and grammatical diversities among different languages, the same information in one language might be conveyed differently in another language. To identify the translation equivalences in parallel texts, many systems and tools were developed.

A system called termight developed by Ido Dagan and Ken Church ([DC94]) is a tool for the identification of technical terms and the support of translation processes. The system is based on part of speech tagging and the word alignment program word_align ([DCG93]). Word_align is based on the program char_align([Chu93]) which combines cognateness with length-based algorithm for aligning parallel texts. Church et al.([Chu93]) report preliminary success in aligning the English and Japanese versions of the AWK manual using char_align. This was possible since the AWK manual happens to contain a large number of examples and technical words that are the same in the English source and Japanese target texts.

Another approach using statistical method has been developed by Pim van der Eijk ([Eij93]). This method concentrates on identifying noun phrase correlations from a previously aligned and tagged parallel corpus. In the preprocessing step, the corpus has to be sentence aligned by using the approach from Gale and Church ([GC91]) and tagged with part-of-speech tags.

For identifying noun phrases, a simple pattern matching algorithm was used. In this algorithm a noun phrase is simply a sequence of zero or more adjectives followed by one or more nouns([Eij93]).

The statistical method for finding correlations is based on the assumption that the translation of a term is more frequent in the subset of target text segments aligned to source text segments containing the source language term than in the entire target language text. The method consists in building a "global" frequency table for all target language terms. To create a "local" frequency table, the frequency ($freq_{local}$(Target term | soruce term)) of the translation of a term in the subset of target text segments aligned to the source text segments containing the source language term is also calculated.

$$\frac{freq_{local}(\text{target term} \mid \text{soruce term})}{freq_{global}(\text{target term})} \quad \ldots\ldots(2)$$

When using this score with low frequent words, problems appear. Therefore, all target language terms with a local frequency below a specific threshold were removed. To improve the results, Eijk ([Eij93]) also proposes a position-sensitivity score function.

Melamed([Mel95]) concentrated on the automatic lexicon evaluation by using several filters. First, all source language and target language words from a sentence

alignment were combined into word pairs. Then, the filters were applied in cascades to find the N-best translations among the translation candidates.

Four different filters were used: 1) part of speech filters; 2)machine-readable bilingual dictionary filters; 3) cognate filters; 4) word alignment filters.

Firstly, the part of speech filter removes every translation candidate with different parts of speech in the source and the target language.

Secondly, if a translation candidate appears in the machine-readable bilingual dictionaries (MRBD), all pairs with the same source language word and a different target language word, and all pairs with the same target language word and a different source language word which occur in the same sentence pair were removed. In other words, the translation of the source language word from the MRBD is assumed to be correct and all other target language words from the same sentence alignment are not considered as possible translations any more.

Thirdly, cognate filters are based on the assumption that there are similarities between the source language word and its translation in related languages. Melamed used the Longest Common Subsequence Ratio (LCSR) to rank the translation candidates by their level of 'cognateness'.

Additionally, the last filter assumes that in related languages information is expressed with a similar word order. If one source language word w1 is aligned to a specific target language word w2, all source words before w1 correspond more likely to the target words before w2 and the source words after w1 correspond more likely to the words after w2.

These automatic evaluation methods can be used to prepare bilingual dictionaries for a human evaluation and to improve machine-readable bilingual dictionaries.

Apart from the methods mentioned above, a statistical approach for bilingual lexicon extraction was developed by Fung and McKeown ([FM94]). Without sentence alignment, lexicon candidates are extract by looking for similarities in the distribution of source and target language word. For this purpose the bilingual text is split in K pieces. After this K-dimensional binary vectors were created for the source and the target language word. If a specific text piece contained the source language word, the corresponding flag in the vector will be set. Then statistical methods can be used to find the similarities between these words.

In the K-vec method the mutual information score has been used:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \quad \text{.....(3)}$$

P(x,y) is the probability that the words x and y occur in corresponding pieces, and P(x) and P(y) are the probabilities that x occurs in a language text and y occurs in the other language text respectively.

The probabilities were estimated by using absolute frequency numbers. Using this mutual information measure, translation candidates can be ranked, and the most likely pair will be chosen as lexicon pair.

Problems with the mutual information score arise with low frequency words. The t-score was used to filter out insignificant values. In order to have a confidence level of 95%, the words have to occur in at least 3 different pieces of text. Another problem is to choose K. If K is too small the mutual information becomes unreliable, and if K is set too high the signal can be lost.

This algorithm was applied to the Canadian Hansards corpus to compile an estimation for a bilingual dictionary. The authors later used the K-vec method for an alignment of parallel texts([FM97]).

In addition, Wu and Xia([WX94]) developed a method for extracting single word translations from a sentence aligned parallel Chinese/English corpora. In this method, the 'estimation-maximization' technique was used together with a significance filtering.

2.3 Research in alignment of sequences

The multilingual alignment problem is related to a number of other sequence comparison problems with applications in various domains([Sim99]), especially sequences alignment of nucleotides in DNA or RNA molecules and of amino acids in proteins ([EGGI93]). The methods used to attack sequence alignment problems are very similar to those used in translation alignment, and rely largely on dynamic programming.

Sequence alignment is an important tool in a wide variety of scientific applications. It shows how the two sequences could be related, i.e. how they can matched, approximately([All99]). A particular model for relating sequences defines a cost or alternatively a score when finding an optimal alignment. In addition, sequence alignment models typically focus on the similarities and differences between characters. To maximize the similarities, longest common subsequence(LCS) is used. To minimize the number of differences, edit distance is applied. Alternatively, some combinations of LCS and edit distance are also used to find the optimal alignment([All99]).

There are many examples of sequence alignment. For instance, the Unix command *diff* f1 f2 finds the difference between files f1 and f2 and updates f1 into f2 by using special edit distance algorithm that is fast([All99]).

To solve a remote screen update problem, sequence alignment is commonly used. If a computer program on machine 1 is being used by someone from machine 2, then machine 1 may need to update the screen on machine 2 as the computation proceeds. One approach is that the program on machine 1 keeps what is currently on the screen of machine 2 and what the screen on machine 2 should become. The differences can be found by an algorithm related to edit-distance and the differences transmitted([All99]).

To solve approximate string matching problem, algorithms related to the edit distance may be used in spelling correctors. If a word w in text is not in the dictionary, the closest word, i.e. one with a few edit operations to w, may be suggested as a correction([SS97b]).

In molecular biology, sequence alignment gives an indication of two strings that are related. Computing a distance between DNA sequences with an alphabet of {A,C,G,T}or protein sequences with an alphabet of 20 amino acids is performed for various purposes, e.g. to find genes or proteins that may have shared functions or properties, or to infer family relationships and evolutionary trees over different organisms([All99]).

In speech recognition systems, sequence alignment is used to find a close match between a new utterance and one in a library of classified utterances.

Chapter 3

Alignment at word level and character level

In this chapter, we will introduce the proposed automatic construction Chinese/English parallel corpus model based on title alignment(Section 3.1). In the section, we also analyze the characteristics of translation together with problems during alignment based on grammatical and lexical features. Then we discuss the proposed alignment model starting from the word level and character level(Section 3.2). In addition, we will present how the system deals with the overlapping ambiguity(Section 3.3) during alignment.

3.1 Title Alignment

According to the Collins Cobuild dictionary, if you align something, you "place it in a certain position in relation to something else, usually along a particular line or parallel to it." A textual alignment usually signifies a representation of two texts which are mutual translations in such a way that the reader can easily see how certain segments in the two languages correspond([MH96]).

Titles of two texts can be treated as the representations of two texts. Referring to He([He00]), the titles present "micro-summaries of texts" that contain "the most important focal information in the whole representation" and as "the most concise statement of the content of a document". In other words, titles function as the condensed summaries of the information and content of the articles. With a few words, a well-written title can provide a reader with enough information to decide

whether to read the article or not. To take the advantage of importance of titles, we use English and Chinese titles to automatically construct the bilingual parallel corpus.

For general bilingual texts, people with a good knowledge of the two languages can identify how two texts are related by looking at the titles of the bilingual texts. To identify the connection between a domain-specific article and its translation, domain knowledge is needed.

A title is also treated as a short sentence in the proposed alignment model and we start the proposed parallel corpus construction at the sentence level. There are several advantages to align titles. Firstly, similar to alignment at sentence level, there is a clear boundary among titles. Secondly, it is easier and faster to identify a title and its translation than to align every sentences in two texts. Thirdly, like sentence alignments which has proven their usefulness in a number of applications([Sim99]), e.g. bilingual lexicography, automatic translation verification and the automatic acquisition of knowledge about translation, the title pairs will use in bilingual lexicon construction in the future. Finally, bilingual titles can be used in analysis of conceptual alteration and the analysis can facilitate the alignment process. (A discussion of conceptual alteration will present later.)

To align two titles, an analysis of the characteristics of title translation pattern for HKSAR government press release titles was performed. Similar to the "translation pattern" suggested by Gale and Church([SFI92]), 1) a title translates into one; 2) a title is not translated at all, e.g. English only or Chinese only articles; 3) a title has no equivalent in another language. For example, a Chinese title "財政司司長將進行首次海外官式訪問"(The Financial Secretary will make his first official overseas visit) was displayed in the HKSAR government Chinese press release web site but in the English press release web site, only the title "The Financial Secretary visits the elderly house" was found. The Chinese title is not equivalent to the English title.

Apart from analysis at title level, the characteristics of translation at word level were also examined where 1)a word is translated one word in another language; 2)many words in one language is translated into one word in another language; 3)some words are not translated; 4)a word is not always translated in the same way; 5)a word is translated into morphological or syntactic phenomena rather than a word.

The characteristics of translation are caused by grammatical and lexical differences between Chinese and English.

3.1.1 Lexcial features

To align two titles, we need to analyze the lexical items(words) of Chinese and English title. A word(a lexical item[Lar98]) in a title may be a concept in a language. According to Larson([He00]), a **concept** "is a recognizable unit of meaning in any given language". For a given language, a concept may not only be represented by a word or words, but may also be represented by a morpheme, by an idiomatic expression, by tone, or by word order([He00]). In translation, a concept represented by a word in one language may be translated into a word, two words, a phrase, or even a sentence in another language.

Since there is no precise definition for word in Chinese([NBR96]), a concept in English may be translated into a word with one or more characters in Chinese and each character in the Chinese word can also treat as a word. For example, the English word "food" can be translated as "食物" in Chinese where "食" stands for "eat" and "物" is "thing" or "object". As a result, abbreviation and morphology are commonly used in both Chinese and English titles. For example, the Chinese title "美利堅合眾國是一個美麗的國家" can be translated as "The U.S. is a beautiful country." By looking up our dictionary, the U.S. can be translated as "美國" in Chinese which is a short-form of "美利堅合眾國". This causes difficulty in the proposed alignment process because the characters of an Chinese word, which

corresponds to an English word, does not appear adjacent to one another in a Chinese title.

Apart from abbreviations, a Chinese word is represented by a morpheme in a Chinese title. For example, the English title "Red flag hoisted" can be manually translated as "海灘懸掛紅旗". Flag can be represented by "旗幟" in Chinese but only "旗"of the whole word "旗幟" appears in the Chinese title. Also, "stamp" can be translated as "郵票" in Chinese and "exhibition" is known as "展覽" but the concept "stamp exhibition" can also be represented by a Chinese word "郵展".

In addition, a concept in one language is broader concept encompassing some narrower concepts, and the translation of such a concept may result in an altered concept in another language. In contrast, a narrower concept in one language may be translated as a broader concept in another language. This is also called **generic-specific** relationship([Lar98]). For example, the counterpart of "環境食物局代表團訪京" published in Hong Kong government press release web site is "EFB delegation on courtesy visit to China". By reading both the Chinese and English titles, the word "China" is modified to be a specific word "京" (Beijing), a city of China.

Omission and addition are also common phenomena. For example, "Closure of Customs Hong Kong Collection Office" is corresponded to "海關香港收款處下月中起停止服務"in Hong Kong government press release web site. "Closure" is represented by "關閉" in our dictionary but in this case, it refers to "停止服務 (stop service)". There is no corresponding words for "下月中起" in the English title.

All these phenomena mentioned are defined as conceptual alteration([He00]). Sager([He00]) provided five types of alteration of information in translation which presented in Section 2.2.2. Nida([He00]) gives an explanation for **conceptual alteration** is that 1) no two languages were completely isomorphic; 2)different

languages might have different domain vocabulary; and 3) some languages were more rhetorical than other languages.

Other than conceptual alteration, morphology causes other difficulties in the alignment process. Since an English word can be translated into a set of Chinese words, some of the Chinese translations wholly appear in a Chinese title and some of them partially appear in the Chinese title. For example, the English title " a beautiful American girl" can be translated as "一個漂亮的美國小女孩". "beautiful" is known as {美麗 , 漂亮} where both "漂亮" and "美" of "美麗" appear in the Chinese title. However, the character "美" of "美麗" wrongly match with the character "美" of "美國". This mismatch problem is classified as overlapping ambiguity in the proposed algorithm because the morpheme "美" of "美麗" wrongly overlaps with the morpheme "美" of "美國".

Also, an English word can be translated into a number of synonyms in Chinese. The longest Chinese word contains the short Chinese words. So the short Chinese words can be viewed as morpheme of the longest Chinese word. For example, referring to our dictionary, "red" is known as {紅, 紅色, 紅色的} in Chinese. If "紅色的" appears in a Chinese title, it stands that "紅" and "紅色" also appear in the Chinese title. This phenomenon is classified as overlapping in the proposed alignment algorithm.

Besides, morpheme of the Chinese words can be useful in the alignment process. If the Chinese synonyms, which corresponds to an English word, share some common characters, the common character can help the proposed system to identify the information related to the English word. For example, "Police Department" is known as "警務處" in Hong Kong. According to our dictionary, "police" is known as{警方, 警察}and "department" is known as {署, 部門, 處 }. The common character for "police" is "警". Since "警" of "警方" overlaps with "警" of "警察", the system

knows that "police" is something related to "警". This phenomenon is also classified as overlapping in the proposed alignment algorithm.

Furthermore, morpheme of a Chinese word appears more than once in a Chinese title and the proposed system is not sure the morpheme corresponds to which English word. For example, the English phrase "airline lounge" can be manually translated as"機場候機室" where the Chinese character "機" appears twice. According to our dictionary, the word "lounge" is known as { 大堂, 餐廳 , 廳} which all do not appear in the Chinese phrase "機場候機室" . The word "airline" is known as { 航機, 航 空公 司} based on our dictionary. The character "機" of "航機" appears in "機場候機室" twice.

3.1.2 Grammatical features

Apart from lexical features, grammatical features such as **chaining** and **redundancy** [Lar98] also cause problems in the proposed alignment process. In chaining, some part of the preceding information is repeated at the beginning of the new unit. For example, consider the phrase "red colour". Based on our dictionary, red is known as { 紅, 紅色, 紅色的, 赤} in Chinese and colour is known as {色, 顏色, 色彩 }. In the Chinese translation set of "red", the meaning of colour has been included.

Redundancy is a repetition([Lar98]). Sometimes the exact words, phrases, or clauses are repeated. More often the exact words are not used. For example, the phrase "red colour" can be manually translated as "赤紅色" in Chinese. But according to our dictionary, red is known as {紅, 紅色, 紅色的, 赤}. "red" can be represented by "赤" as well as "紅色" in Chinese. From the Chinese translation "赤紅色", we know that the meaning of red has been repeated.

The grammatical features are classified as overlapping in the proposed alignment algorithm.

3.1.3 The English/Chinese alignment model

The alteration of information in translation and overlapping of words causes difficulties in the alignment process. To overcome the problems during alignment, sequence comparison methods in the domains other than translation alignment were reviewed. The methods used to attack sequence comparison problems are very similar to those used in translation alignment, and rely largely on dynamic programming([Sim99]).

Dynamic programming "is defined as an algorithmic technique in which an optimization problem is solved by caching subproblem solutions(memoization) rather than recomputing them"(Algorithms and Theory of Computation Handbook, p.1-26). Dynamic programming refers to a very large class of algorithms. Examples of dynamic programming are edit distance and longest common subsequence (LCS). In sequence alignment, longest common subsequence (LCS) is commonly exploited to maximize the number of matches between characters of two sequences. The edit distance is applied to minimize the difference between two sequence.

In the alignment model, a title is viewed as a sequence of words and a words can be treated as a sequence of characters. Both edit distance and longest common subsequence are used to optimize the alignment of two sequences. Edit distance is used to deal with overlapping and longest common subsequence is applied to maximize the number of matches between Chinese titles and English titles. The proposed algorithm includes three sections:1) alignment at word level and character level, 2) reduce overlapping ambiguity, 3) alignment at title level.

3.2 Alignment at word level and character level

3.2.1 Alignment at word level

To start the proposed alignment, we take the advantage of language characteristic of English and align from English titles to Chinese titles. The advantage is that there is a clear delimiter between two English words.

Since Chinese texts do not contain obvious word boundaries. A Chinese text consists of a linear sequence of non-spaced or equally spaced ideographic characters ([WT93]). In contrast, English text can be segmented into words using spaces and punctuations as word delimiters([WT93]). Figure 3 illustrates a typical Chinese text without word separation and its English translation.

很高興今天能夠出席由「基本法推介聯席會議」所舉辦的「基本法頒布十周年研討會」。《中華人民共和國香港特別行政區基本法》這套史無前例的法律，對「一國兩制」的偉大構思及中央政府授予香港特區的高度自治權，均作出了具體的憲制性規定。欣逢《基本法》頒布十周年，我們正好藉這個極具歷史意義的時刻，回顧「一國兩制」的實踐情況，以探討它的意義和未來發展。

It gives me great pleasure to attend the Symposium in Commemoration of the 10th Anniversary of the Promulgation of the Basic Law of the HKSAR organised by the Joint Committee for the Promotion of the Basic Law of Hong Kong. The Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China is the constitutional framework which institutionalises the unprecedented concept of "One Country, Two Systems" and the high degree of autonomy conferred upon the HKSAR by the Central Authorities. This year marks the 10th anniversary of the promulgation of the Basic Law. On this historic occasion, it is timely for us to review implementation of the "One Country, Two Systems", and explore the implications for the future.

Figure 3 Typical Chinese text without word separation

The absence of word boundaries is generally regarded as a big obstacle to the computer processing of Chinese language([WT93]). The obstacle is also applied to extract certain meaningful and content-bearing units in English, such as morphemes, phrases or some combinations. These units do not have natural boundaries separating one another. This is probably an universal phenomenon which is illustrated in relation to Chinese and English in Table 1([WT93]).

Linguistic Unit	Delimiters	
	Chinese	English
Morpheme	No	No
Simple word	No	Yes
Compound word	No	No
Phrase	No	No
Clause	No	No
Sentence	Yes	Yes

Key: Yes means the existence of delimiters, which include spaces and punctuation marks; no stands for the absence of obvious delimiters.

Table 1: The common characteristics of Chinese and English

Since there is a clear delimiter between two English words, the word alignment starts from English title. To obtain the maximum information from the English titles, each English title is segmented into a small unit, a single English word. Each word of an English title is translated into a set of Chinese words through dictionary lookup.

As dictionary provides linguistic information, Resnik([Res99]) has suggested to use bilingual dictionaries in alignment of document segments in his parallel corpus construction. Utsuro et al.([UIY94]) use both bilingual dictionary and statistics to align Japanese and English texts. Klavans et al. ([KT96]) also use dictionary and

statistical method to enhance the lexicon entries. Melamed ([Mel95]) applies machine-readable bilingual dictionaries (MRBDs) to provide linguistic information to rank the translation candidates of a word. In the alignment approach developed by Melamed ([Mel96b]), he suggested when a large translation lexicon is not available, a small hand-constructed translation lexicon can be used to assist text alignment.

3.2.2 Alignment at character level: Longest matching

To facilitate the alignment of two words, alignment at character level is needed to study. To align English words with their French correspondences, Simard et al. ([SFI92]) use cognate. Chen et al. ([CKJ99]) has applied a Japanese morphological analyzer to align the Japanese translation of an English word and a Japanese word. Utsuro et al.([UIY94]) also used morphological analysis in their English/Japanese alignment.

To align French and English texts, Melamed([Mel96b]) applied the longest common subsequence ratio (LCSR) to rank the translation candidates of a word by their level of "cognateness". Whether a pair of words is considered cognate pair depends on the ratio of the length of their longest(not necessarily contiguous) common subsequence to the length of the longer string([Mel95]). For example, "gouvernement" is the French translation of "government". "gouvernement" is 12 letter long. "gouvernement" has 10 letters that appear in the same order in "government". As a result, the LCSR for these two words is 10/12. On the other hand, the LCSR for 'conseil' and "conservative" is only 6/12. The remaining question was what minimum LCSR value should indicate that two words are cognates. The notion of cognate used by Melamed ([Mel96b]) is different from the notion of cognate used by Simard et al.([SFI92]).

Since Chinese language has a rich vocabulary, a concept may often be expressed in multiple ways([NGZ00]). Also, in modern standard Chinese, word normally consists of more than one character although some characters may be regarded as words

too([WT93]). A Chinese character normally corresponds to the morpheme in an English word([WT93]). It is possible that the characters of a Chinese translation of an English word may not ALL appear in a Chinese title. As a result, to align the translation of an English word to the Chinese characters in a title at character level, we applied the longest common subsequence (LCS) to find the longest matching between a Chinese translation of an English word and its correspondence in a Chinese title. There are many different longest matching approaches were exploited in Chinese information retrieval (IR) to match between queries and documents([NBR96]). Different longest matching method were also used in Chinese segmentation([WT93]).

There are many advantages of using a longest matching approach. Firstly, long Chinese words usually describe more precise meaning than short words([NGZ00]). Secondly, in many cases, single-character words usually have quite different meanings, or archaic meanings, in comparison with the meaning of the compound in modern Chinese([NGZ00]).

Longest common subsequence (LCS) has been applied in Resnik's ([Res99]) parallel corpus construction. Resnik([Res99]) has used a program *diff* to align two documents in parallel corpus construction. The program *diff* compares two files by computing the LCS of their lines([CRG96]). Before discussing the usage of longest common subsequence in the proposed algorithm, we would like to introduce the formal definition for longest common subsequence.

3.2.3 Longest common subsequence (LCS)

First of all, the notion of subsequence and common subsequence must be clarified. **Subsequence** is known as: Given a sequence $X=x_1x_2x_3...x_m$, a sequence T is a subsequence of X if it can be obtained by deleting zero or more elements from X([CRG96]).

Common subsequence is known as: By given two sequences $X=x_1x_2x_3...x_m$, and $Y=y_1y_2y_3...y_n$, a sequence Z is a common subsequence of two sequences X and Y if Z is a subsequence of both X and Y.

For example, if X=ywcqpgk and Y=lawyqqkpgka, the sequence 'qpgk' is a common subsequence of both X and Y. But the sequence 'qpgk' is not the longest common subsequence of X and Y. The sequence 'yqpgk' is also a common subsequence of X and Y. The length of 'yqpgk' is longer than the length of 'qpgk'.

The **longest common subsequence** problem is defined as: By given two sequences X and Y and we wish to find a maximum-length common subsequence of X and Y([CLR90], [AD86]). The longest common subsequence (LCS) problem typically asks for generating one or more longest common subsequences of X and Y since X and Y can have several different longest common subsequences. It is possible that there is no common subsequence of X and Y, e.g. there is no common subsequence of X and Y.

A sequence Z is a longest common subsequence of X and Y if the length of Z is maximal. In the proposed English/Chinese algorithm, the set of all the longest common subsequence of X and Y is denoted by LCS(X,Y). The length of the elements of LCS(X,Y) is denoted by |LCS(X,Y)|. In case, there is no common subsequence of X and Y, LCS(X,Y)=∅.

Some authors use string to represent a sequence of characters([RY97], [OH92]). A longest common substring problem is introduced([SS97a]). In our case, we use sequence in our notion to emphasis on the order of characters in sequence.

Memoization

To solve the LCS problem by dynamic programming, we need a recursive solution. The problem with the recursive solution is that the same subproblems get called many times. To solve the recursive problem, instead of recomputing a subproblem, we check whether or not the calculation has been done before. This recursive version of dynamic programming is known as "memoization"([Epp96]).

Example of LCS

Given a sequence X=ywcqpgk and a sequenceY=lawyqqkpgka.

The sequence 'qpgk' is a common subsequence of both X and Y. But the sequence 'qpgk' is not the longest common subsequence of both X and Y.

The sequence 'yqpgk' is also a common subsequence of X and Y. The length of 'yqpgk' is longer than the length of 'qpgk'.

Since there is no common subsequence of length 5 or greater, the sequence 'yqpgk' is a longest common subsequence (LCS) of X and Y. In addition, there is another LCS of X and Y, 'wqpgk'. As a result, LCS(X,Y)={yqpgk,wqpgk}.

3.2.2 Applying LCS in the English/Chinese alignment model

Before explaining how the longest common subsequence(LCS) is applied in the proposed English/Chinese algorithm, we clarify our notions:

$\{a_1, a_2, a_3,\ldots\}$ denotes as a set. Set is "an unordered collection of values where each value occurs at most once"(Dictionary of Algorithms, Data Structures, and Problems, National Institute of Standards and Technology). In the proposed algorithm, each element in a set is unique.

ε denotes as an empty sequence.

$b_1b_2b_3\ldots$ denotes as a sequence where the symbol b_1 can be a word in English or a character in Chinese.

An English title, $E = e_1e_2e_3\ldots e_i\ldots$ where e_i is an English word in E. An English title is formed by a sequence of English words and an English word is formed by a sequence of English letters. Letter is referred as English letter and character is referred to a Chinese character in our notion.

A Chinese title, $C = char_1char_2char_3\ldots char_q\ldots$ where $char_q$ is a Chinese character in C. A Chinese title is formed by a sequence of Chinese characters. Unlike English, the written Chinese text has no delimiters to mark word boundaries and there is no precise definition of word in Chinese([NBR96]).

Translated(e_i) = $\{ T^1_{e_i}, T^2_{e_i}, T^3_{e_i}, \ldots, T^j_{e_i}, \ldots \}$ where $T^j_{e_i}$ is the j^{th} Chinese translation of an English word e_i . A Chinese translation is formed by a sequence of Chinese characters.

The set of the longest-common-subsequence (LCS) of one Chinese translation $T_{e_i}^{j}$ and C is denoted as LCS($T_{e_i}^{j}$, C).

$$\text{The set MatchList}(e_i) = \bigcup_{j} \text{LCS}(T_{e_i}^{j}, C) \quad \text{.....(4)}$$

The set MatchList(e_i) holds all the unique longest common subsequences of $T_{e_i}^{j}$ and C.

The set MatchList(e_i) is an empty set in case there is no common subsequence of $T_{e_i}^{j}$ and C and there is no reliable translation for e_i appearing in C.

$$\text{Count}(e_i) = \begin{cases} 1 & \text{if MatchList}(e_i) \neq \varnothing \\ 0 & \text{Otherwise} \end{cases} \quad \text{.....(5)}$$

If the set MatchList(e_i) for e_i is not empty, then Count(e_i) for e_i is 1.

Contiguous(e_i)={x | x∈ MatchList(e_i) and all the characters of x appear adjacently in C}

.....(6)

If all the characters of the elements of MatchList(e_i) appear adjacently in C, the elements are also the elements of the set Contiguous(e_i). This is based on the hypothesis that if all the characters of a word w1 appear adjacently and the characters of a word w2 do not appear adjacently, the word w1 is more reliable than the word w2. For example, the word "propose" can be translated as "建議" in Chinese. The translation "建議" can wrongly align with "就**建** 築條例的動 **議** 辯

論"(on the "Construction Bill" motion debate). As a result, we use the function Contiguous(e_i) to solve the problem.

If MatchList(e_i) =∅, the Contiguous(e_i) is also an empty set

$$\text{Reliable}(e_i) = \begin{cases} \arg\max_{x \in \text{Contiguous}(e_i)} |x| & \text{if Contiguous}(e_i) \neq \varnothing \\ \arg\max_{x \in \text{MatchList}(e_i)} |x| & \text{Otherwise} \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Reliable (e_i) contains the most reliable translation of e_i with respect to C.

Example

E=A beautiful American girl

e_1=A e_2= beautiful e_3= American e_4=girl

C=一個漂亮的美國小女孩

$char_1$=一 $char_2$=個 $char_3$=漂 $char_4$=亮 $char_5$=的 $char_6$=美 $char_7$=國 $char_8$=小 $char_9$=女 $char_{10}$=孩

Translated (e_2)={ $T^1_{e_2}$, $T^2_{e_2}$, $T^3_{e_2}$ }={漂亮的, 美麗的, 美}

LCS($T^1_{e_2}$, C)={ 漂亮的} LCS($T^2_{e_2}$, C)={ 美, 的} LCS($T^3_{e_2}$, C)={ 美}

MatchList(e_2) ={ 漂亮的,的, 美}

Contiguous(e_2) ={漂亮的,的, 美}

Reliable (e_2) =漂亮的

Count(e_2) =1

3.3 Reduce overlapping ambiguity

As two or more words may translate into one word ([He00]), the translations of the words is overlapped. The overlapping ambiguity is more significant in Chinese because there is no precise definition of word. For example, a Chinese string ABCD can be segmented into bi-gram AB BC CD, or uni-gram A B C D([NBR96]). The overlapping between two words will cause the proposed score function to count the same Chinese character more than once. To resolve the overlapping between 1) a reliable word selected by the function Reliable and the rest of elements in MatchList, 2) two reliable words selected by the function Reliable, the **deletion**, an edit operation is used.

Melamed ([Mel97b]) used an one-to-one assumption to solve one-to-many translation problem. In his algorithm, each word is translated to at most one other word. If two word (w1,w2) are translated into one word w3, either the word pairs(w1,w3) and (w2,null) or (w1,null) and (w2,w3) may appear. One element in a word pair w is a word. Another element is the translation of that word w.

Similar to Melamed([Mel97b]), our aim to reduce overlapping is to make a word in a Chinese title referring to exactly one English word in an English title. An edit operation, deletion, is applied in our case to solve overlapping problem.

3.3.1 Edit distance

Edit distance has been used to minimize the number difference between two sequences([APD99]). Edit distance of two sequences is defined as the minimum number of insertions, deletions and substitutions required to transform one sequence into the other([RY97], [EGGI93]). Substitution is also known as mismatch or

change([All98]). Substitution, deletion and insertion are **edit operations** or **point mutations**([All98]).

Performing an insertion means augmenting a sequence by adding a character. A deletion means removing a character from a sequence. A substitution is the replacement of a character in a sequence by another character([OH92], [RY97], [APD99]). A match represents a character in a sequence matches a character in another sequence([APD99]).

Example of edit distance

Figure 4 shows an example alignment of sequence 'compression and approximate matching' and 'comprehension of appropriate mapping' by using edit operations. Spaces have been replaced by '/' to make them visible. The sequences have been padded out with a special null pseudo-character, denoted by '-', so they have the same length. Matches are emphasized by a vertical bar, '|', between the matches characters. If the first sequence is considered to be the parent, then a column with a '-' in the top row represents an insertion and a column with a '-' in the bottom row represents a deletion. No column may contain two '-'s. A column with different characters in the top and downs represents a change, also known as a mismatch. A match is sometimes called a copy.

```
compre --ssion/and /approximate/matching
||  ||||   |||||   ||| |||    ||||| |   |||
comprehension/of- /appropriate/ma -pping
```

Figure 4 Example of edit distance

The example is inspired by Allison et al.([APD99]).

The example tells that an alignment is one way of editing one sequence into another using point mutations where point is character in a sequence. Each operation can be given a cost or a score, and one can then search for an optimal alignment. Matches are good and are given low costs or high scores. Mutations are bad and are given high costs or low scores([APD99]).

Edit distance has been used in many different applications, especially molecular biology. In molecular biology, the edit distance gives an indication of similarity between two DNA sequences([All99]). A detailed discussion of the applications of edit distance is provided in Section 3.3.

3.3.2 Overlapping in the English/Chinese alignment algorithm

The matched characters between two sequences can be viewed as the overlapped characters in the proposed algorithm. To solve overlapping problem, the proposed algorithm keeps the mismatched characters between two sequences by removing the overlapped characters. The method to solve overlapping problem can be treated as detecting the difference between two sequences in Allison([All99]). Unlike file revision([All99]), remote screen update([All99]) and change detection in information([CRG96]), we do not need to update the sequences by combining the difference and the original sequences or data tree. So substitution and insertion are not needed in the proposed algorithm. To detect the difference between two sequences, longest common subsequence(LCS) is applied to find the set of matched characters. Once the difference is known, the matched characters are deleted from the unreliable sequence. In other words, if the difference is detected, the mismatched characters of the unreliable sequence are inserted into an empty sequence by order.

Dele(X,Y)

We define a function called **Dele(X,Y)** to deal with overlapping problem as follows:

Given two sequences $X=x_1x_2x_3\ldots x_m$ and Y

For i=1 to m

{ assign a boolean value 0 to x_i ; }

For each t ∈ LCS(X,Y)

{ assign the boolean value of each matched character to 1 ;}

For i=1 to m

{ IF the boolean value of x_i is 1

THEN delete x_i from X ;

}

.....(8)

Example of Dele(X,Y)

Let X=wyqqkpgk and Y= Y=ywcqpgk
LCS(X,Y)={yqpgk, wqpgk}
Dele(X,Y)=qk

0 0 0 0 0 0 0 0
wyqqkpgk

0 1 1 0 0 1 1 1
wyqqkpgk
|| |||
- yq - -pgk the mismatch characters in the case: wqk

1 1 1 0 0 1 1 1
wyqqkpgk our program matches the first 'q' in wyqqkpgk
| | ||| because our matching follows sequential order
w-q- - pgk the mismatch characters in the case are yqk

After deleting the characters labeled with 1, the new sequence is qk

Differ(X,Y)

A function called **Differ(X,Y)** is defined as follows:

Given a set X= $\{x_1, x_2, x_3, \ldots, x_m\}$ and each element x_c of X is a sequence

Given a sequence Y

Initialize Differ(X,Y): = $\varnothing$

$$\text{Differ}(X,Y)=\bigcup_{c=1}^{m}\text{Dele}(x_c,Y) \quad \ldots\ldots(9)$$

Example of Differ(X,Y)

IF X={ab, bcd, dbc} and Y=bc

Differ(X,Y)={a,d}

Application of Dele(X,Y) in the English/Chinese alignment algorithm

The application of Dele(X,Y) divides into three levels:

1) It reduces the overlapped characters between two reliable words (e.g. Reliable(e_i) and Reliable(e_t) where i<t) because two words may translate into one word. For example, by using Dele(X,Y), we can remove the Reliable(e_i) from a copy of the Chinese title C (the copy of C is called Remain). If Reliable(e_i) is identical to Reliable(e_t) and Reliable(e_i) appears in C once, then removing Reliable(e_i) from Remain will cause Reliable(e_t) disappearing in Remain. So even though Reliable(e_t) appears in C, it may not appear in Remain. On the other hand, Count(e_t) memorizes that e_t gets a translation in C. By doing so, we do not need to count the same Chinese character twice in the proposed score function.

2) It reduces the overlapped characters between a reliable word (Reliable(e_i)) and the other word words in the MatchList(e_i) . Since a long Chinese word may contain several short words([NGZ00]), the short words in MatchList(e_i) may overlap with the long word Reliable(e_i).

3) It reduces the overlapped characters between a reliable word(Reliable(e_i)) and the words in a MatchList other than MatchList(e_i).

WaitList

WaitList is defined as a set to hold the result after each Differ process.

Initialize WaitList : =∅ and Remain:= C for an E where E is an English title and C is a Chinese title

For each $e_i \in E$

{ IF Reliable(e_i) ≠ ε

THEN

{ WaitList :=Differ(WaitList, Reliable(e_i)) ∪Differ(MatchList(e_i)\{Reliable(e_i)} , Reliable(e_i)) ;

Remain:= Dele(Remain, Reliable(e_i));

}

}.....(10)

Example 1:

E = A beautiful American girl e1=A e2=beautiful e3=American e4=girl
C =一個漂亮的美國小女孩

Initialize WaitList:=∅ and Remain:= C

Translated(e_1)= {一, 一個} MatchList(e_1) ={一, 一個}
Reliable(e_1)= 一個 Count(e_1) =1

WaitList : = Differ(WaitList,Reliable(e_1)) ∪ Differ(MatchList(e_1)\{Reliable(e_1)}, Reliable(e_1));
:= Differ(∅ , 一個) ∪ Differ ({一}, 一個)
:= ∅ ∪ ∅
:= ∅
Remain := Dele (Remain, Reliable(e_1))
:= Dele (一個漂亮的美國小女孩, 一個)
:= 漂亮的美國小女孩

Translated(e_2)={漂亮的, 美麗的 , 美} MatchList(e_2)={漂亮的,的,美}
Reliable(e_2)= 漂亮的 Count(e_2)=1

WaitList :=Differ (∅ , 漂亮的) ∪ Differ ({ 的, 美 }, 漂亮的)
:= {美}
Remain:= Dele(Remain, Reliable(e_2))
:= Dele (漂亮的美國小女孩 , 漂亮的)
:= 美國小女孩

Translated(e_3)={美利堅合眾國的, 美國人,美國的} MatchList(e_3) = {美國}
Reliable(e_3)= 美國 Count(e_3)=1

WaitList:=Differ ({美}, 美國) ∪ Differ(∅ , 美國)
:= ∅
Remain := Dele(美國小女孩 , 美國)
:= 小女孩

Translated(e_4)={小女孩} MatchList(e_4) = {小女孩 }
WaitList :=Differ (∅ , 小女孩) ∪ Differ (∅, 小女孩)
:= ∅
Remain := Dele (小女孩 , 小女孩)
:= ε

Chapter 4

Alignment at title level

Since we have discuss the proposed alignment model at character level and word level, we will go through the proposed alignment at title level. In this chapter, we will present the proposed score function(Section 4.2) to find the optimal solution for aligning titles after a brief review of other score functions in the field of translation alignment(Section 4.1).

4.1 Review of score functions

Alignment algorithms can be used to infer a relationship between sequences when the true relationship is unknown. Simple algorithms use a cost or score function that gives a fixed to each edit operation. These algorithms tend to find optimal alignment([PAD00]). For translation alignment, Gale and Church ([GC91]) used a score function based on a probabilistic model to align two texts at sentence level. In our case, a title can be viewed as a short sentence. Simard et al.([SFI92]) has adopted the same score function together with the notion of cognateness in their English/French text alignment. Chen et al.([CKJ99]) extended Gale and Church's program in their Japanese/English sentence alignment based on sentence length. All Gale and Church([GC91]), Simard et al.([SFI92]) and Chen et al.([CKJ99]) relied on the length similarity to calculate the relationship between two sentences. Since Chinese and English have no history of common development, we cannot reply on the sentence length similarity for alignment([CT95]).

In addition, unlike Gale and Church([GC91]), Simard et al.([SFI92]), Brown et al.([BLM91]), Chen([Che93]) and Kay and Röscheisen([KR93]), we do not have a large corpus and build a statistical/probabilistic model based on the corpus. Our objective is to construct a bilingual corpus. Furthermore, unlike sentence in a text, the titles in one language are independent of one another. To align the bilingual titles, the proposed score function simply combines both the length similarity and matched words in English and Chinese titles.

Before further discussing the proposed score function, we would like to clarify the notion of matched word. The matched words in our case is defined as follows: if the translation of an English word in an English title appears in a Chinese title, both the English word and its translation are referred as the matched words. For example, in the example 1(p.58), since MatchList(e_1) is not empty set, e_1 is an English matched word and the elements in MatchList(e_1) are the Chinese matched words.

4.2 The Score function

4.2.1 (C matches E) and (E matches C)

After removing the overlapped characters, each character of the Chinese translations that matches a character in a Chinese title C is given a fixed score 1. So we get:

$$(C \text{ matches } E) = \sum_{x \in \text{WaitList}} |x| + (|C| - |\text{Remain}|) \quad \ldots\ldots(11)$$

$|x|$ is the cardinality of x

As one word may translate two or more words([He00]), the words in WaitList($\sum_{x \in \text{WaitList}} |x|$) morphologically align with the English words in an English title.

(|C| - |Remain|) tells how many characters in C that matches the translations of English words in an English title E.

To find the optimal alignment, we search the highest (C matches E) score.

Example of (C matches E)

If we use result obtained from the example 1(p.58), we get the following:

E = A beautiful American girl

e_1=A e_2=beautiful e_3=American e_4=girl

C =一個漂亮的美國小女孩

$$(\text{C matches E}) = \sum_{x \in \text{WiatList}} |x| + (|C| - |\text{Remain}|)$$

$$= 0+(|C|-0)$$

$$= |C|$$

$$= 10$$

The score function (C matches E) only tells how well the Chinese title C is similar to E. The counterpart of (C matches E) is (E matches C).

(E matches C)

The score function (E matches C) is defined as

$$(\text{E matches C}) = \sum_i \text{Count}(e_i) \quad \text{.....(12)}$$

(E matches C) tells how many words in an English title E get a translation wholly or morphologically in a Chinese C. (E matches C) can be viewed as counting the word pairs.

Disadvantage of using (E matches C)

A Chinese translation of an English word consists of one or more characters. If only some characters of the Chinese word appear in a Chinese title, error may occur when using (E matches C).

Example 2

An English title E1= Revision of Government fees.
There are three Chinese titles:
C1= 調整政府收費　C2=檢討禮賓府開支　C3=民政事務局削減免費項目

Translated(fees)={費用, 收費, 款項, 開支} Translated(revision)={調整, 檢討}
Translated(Government)={政府} Translated(of)={有關}
(E1 matches C1)=3 (E1 matches C2)=3 (E1 matches C3)=3
(C1 matches E1)=6 (C2 matches E1)=5 (C3 matches E1)=3

The difference among (C1 matches E1), (C2 matches E2) , (C3 matches E1) shows which Chinese title is the counterpart of E1. In contrary, the results for (E1 matches C1), (E1 matches C2) and (E1 matches C3) are equal. We cannot find the optimal alignment based on the results.

From the results shown, the (C matches E) is more reliable in Chinese/English title laignment because the (C matches E) can effectively count both morphological alignment and word alignment. As a result, the alignment algorithm relies on (C matches E).

4.2.2 Length similarity

It is also possible that two or more Chinese titles will get the same (C matches E) score when they align with an English title E. We apply match ratio and absolute difference of the match ratios to make the proposed alignment algorithm more accuracy.

There are two match ratios:

$$1)\quad Ra(\text{E matches C}) = \frac{\sum_{i} Count(e_i)}{|E|} \qquad \text{.....(13)}$$

$$2)\quad Ra(\text{C matches E}) = \frac{\sum_{x \in WaitList} |x| + (|C| + |Remain|)}{|C|} \qquad \text{..... (14)}$$

The nominators of Ra(E matches C) and Ra(C matches E) are (E matches C) and (C matches E) respectively.

The absolute difference of Ra(E matches C) and Ra(C matches E) shows the degree of similarity from a bidirectional point of view.

Example of the match ratios and absolute difference

By using results obtained from both the example 1(p.58) and example of (C matches E) to calculate the match ratios and the absolute difference of the match ratios, we get:

$$Ra(\text{E matches C}) = \frac{\sum_{i} Count(e_i)}{|E|} = \frac{1+1+1+1}{4} = 1$$

$$Ra(\text{C matches E}) = \frac{\sum_{x \in WaitList} |x| + (|C| + |Remain|)}{|C|} = \frac{0 + (|C| - 0)}{|C|} = \frac{|C|}{|C|} = \frac{10}{10} = 1$$

The absolute difference of Ra(E matches C) and Ra(C matches E) is 0.

Disadvantage of using match ratio and absolute difference

Referring to the example 2(p.62),

E1= Revision of Government fees.

C1= 調整政府收費　C2=檢討禮賓府開支　C3=民政事務局削減免費項目

Translated(fees)={費用, 收費, 款項, 開支} Translated(revision)={調整, 檢討}

Translated(Government)={政府} Translated(of)={有關}

(E1 matches C1)=3	(E1 matches C2)=3	(E1 matches C3)=3
Ra(E1 matches C1)=3/4	Ra(E1 matches C2)=3/4	Ra(E1 matches C3)=3/4
(C1 matches E1)=6	(C2 matches E1)=5	(C3 matches E1)=3
Ra(C1 matches E1)=6/6	Ra(C2 matches E1)=5/7	Ra(C3 matches E1)=3/11

The absolute difference for Ra(E1 matches C1)=3/4 and Ra(C1 matches E1)=6/6 is |3/4-6/6|=0.25

The absolute difference for Ra(E1 matches C2)=3/4 and Ra(C2 matches E1)=5/7 is |3/4-5/7|=0.035

The absolute difference for Ra(E1 matches C3)=3/4 and Ra(C3 matches E1)=3/11 is |3/4-3/11|=0.477

From the example, we can see that the minimum absolute difference causes C2 to wrongly align with E1. In addition, the match ratio does not consider the possibility of addition or omission during translation. So the highest value of match ratio may not cause a correct alignment. To improve this, we combine the match ratio and absolute difference with (C matches E).

Generating an alignment matrix

Given a set of Chinese titles TC={C_1, C_2,…,C_k,…} where C_k is a Chinese title

Given a set of English titles TE={E_1, E_2,….,E_p,….} where E_p is an English title

For each $E_p \in$TE

{

For each $C_k \in$ TE

{ calculate (C_k matches E_p);

calculate absolute difference of Ra(C_k matches E_p) and Ra(E_p matches C_k) ;

}

}.....(15)

An alignment matrix is generated.

An example of alignment matrix

E1=Revision of Government fees
E2=Revision of the Housing Department expenditure
C1= 調整政府收費
C2= 政府檢討發展資訊科技的各項開支
C3= 房屋署調整各項開支

Translated(fees)={費用, 收費, 開支} Translated(revision)={調整, 檢討, 修正}
Translated(Housing)={房屋,房子, 收容}
Translated(Department)={署,處,局, 部門}
Translated(expenditure)={開支,費用,開消, 消費}

For E1) (C1 matches E1)= 6
Ra (E1 matches C1) is 3/4 =0.75 Ra(C1 matches E1) is 6/6=1 |3/4 - 6/6| =0.25

(C2 matches E1)= 6
Ra(E1 matches C2) is 3/4=0.75 Ra(C2 matches E1) is 6/15=0.4 |3/4-6/15| = 0.35

(C3 matches E1)= 4
Ra(E1 matches C3) is 2/4=0.5 Ra(C3 matches E1) is 4/9=0.444 |3/5-5/8|=0.056

For E2) (C1 matches E2)=3
Ra(E2 matches C1) is 2/6=0.33 Ra(C1 matches E2) is 3/6=0.5 |2/6-3/6|=0.17

(C2 matches E2)=4
Ra(E2 matches C2) is 2/6=0.333 Ra(C2 matches E2)) is 4/15=0.267 |2/6-4/15|=0.066

(C3 matches E2)=9
Ra(E2 matches C3)) is 4/6=0.666 Ra(C3 matches E2) is 9/9=1 |0.666-1|= 0.333

Title	C1	C2	C3
E1	**6** |3/4 -6/6|=0.25	**6** |3/4 - 6/15|=0.35	4 |3/5-5/8|=0.056
E2	3 |2/6-3/6|=0.17	4 |2/6-4/15|=0.066	**9** |4/6-9/9|=0.333

Combination of the match ratios, absolute difference and (C matches E)

For each E_p row where $E_p \in TE$

{ pick up a C_k with highest (C_k matches E_p) score;

IF there are more than one C_k with the same highest (C_k matches E_p)

THEN

{ select the C_k with the minimum absolute difference;}

Once a C_k is selected, check the C_k column

IF (C_k matches E_p) is the highest score in the column

AND there is no other (C_k matches E_t) with the same highest score $t \neq p$

THEN

{ C_k is the counterpart of E_p ;

}

ELSE IF (C_k matches E_p) is the highest score in the column

AND there is another (C_k matches E_t) with the same highest score

THEN

{ IF the absolute difference for C_k and E_p is the minimum in the column

THEN { C_k is the counterpart of E_p;}

}

}.....(16)

Example 3

Referring to the example of alignment matrix, we get

Title	C1	C2	C3
E1	**6** |3/4 -6/6|=0.25	**6** |3/4 - 6/15|=0.35	4 |3/5-5/8|=0.056
E2	3 |2/6-3/6|=0.17	4 |2/6-4/15|=0.066	**9** |4/6-9/9|=0.333

In the E1 row, both (C1 matches E1)=6 and (C2 matches E1)=6 get the highest score, we pick up the title with minimum absolute difference. ∵ C1 is selected

As C1 is selected, we look at the C1 column. In the C1 column, (C1 matches E1)=6 is the highest. As a result, C1 is the counterpart of E1

C3 is the counterpart of E2 because (C3 matches E2) is the highest score in E2 row.

A title together with the whole article may not be translated in another language. It is the case for C2.

Chapter 5

Experimental results

We have used the Hong Kong government press articles(Section 5.1), Hang Seng Bank economic monthly reports(Section 5.2), Hang Seng Bank press release articles(Section 5.3) and Hang Seng Bank speech articles(Section 5.4) to test the proposed title alignment algorithm. In the chapter, we will present the alignment results for each collection. Finally, we will discuss the quality of these four collections and our future work(Section 5.5).

5.1 Hong Kong government press release articles

As we had talked about the daily press release articles published by the Hong Kong government in Section 1.2, these newswire documents are written in both Chinese and English which are a good source of corpus.

In the governmental web site, some newswire documents are only available in one language. In some days, no press release article is published. Before discussing the title alignment method to overcome the difficulties, we introduce the structure of the web site.

The government web site with bilingual press release articles are arranged similar to the third principle mentioned by Resnik([Res99]): The web site contains a completely separate monolingual subtree for each language. Only the single top-level homepage pointing off to the root page of single-language version of the site.

For retrieving the bilingual press release articles, there are two top-level homepages: one for English (http://www.info.gov.hk/eindex.htm) and another for Chinese (http://www.info.gov.hk/). Each top-level homepage gets an anchor at the right hand corner to link to another top-level homepage.

The press articles for two languages are placed in one subdirectory(e.g. http://www.info.gov.hk/gia/general/200101/16/). For one language, there is a title homepage page. The page lists all the press release titles of a day. Each title in the title homepage is anchored to its article. For example, the address http://www.info.gov.hk/gia/general/200101/16.htm brings the system to the English title homepage for 16th January, 2001. Similar to Chinese, there is also a title homepage for 16th January, 2001. It is http://www.info.gov.hk/gia/general/200101/16c.htm.

The English/Chinese alignment system firstly gets access to the title homepage of a day and then start the title alignment. By using the proposed title alignment algorithm, we collects 8898 press release document pairs (total number of documents are 17796 =8898*2) starting from 1st May,1999 to 31st January, 2001. To evaluate the document pairs, the performance metrics of the proposed alignment system are based on recall and precision.

Recall = number of relevant documents the system retrieved / number of relevant documents.....(17)
Precision = number of relevant documents the system retrieved / number of documents the system retrieved.....(18)

Recall and precision are designed to assess the effectiveness from the users' perspective as they compare the system output against some predefined standard output([LSV98],[OD96]).

In our case, the recall and precision are defined as:

Recall = number of relevant title pairs the title alignment system retrieved/ number of relevant title pairs are published on the web

Precision = number of relevant title pairs the title alignment system retrieved/ number of title pairs the system retrieved

To get the recall daily, we manually count how many title pairs (defined as Rexist) are published on the web everyday. The number of document pairs generated everyday is defined as Rget.

So we get:

$$\frac{\sum_{i=1}^{640} Rget_i / Rexist_i}{640} \quad \text{.....(19)}$$

$Rget_i$ is the number of title pairs the system retrieved in day i. $Rexist_i$ is the number of title pairs published on the web in day i. From 1st May,1999 to 31st January, 2001, 640 day government press release documents are available for evaluation. So we compute the precision and recall for the 640 days.

For precision, we define as follows:

$$\frac{\sum_{j=1}^{640} Precise_j / Pget_j}{640} \quad \text{.....(20)}$$

where $Precise_j$ is the number of accurate title pairs the system retrieved in day j and $Pget_j$ is the number of title pairs the system retrieved in day j.

The content of documents are also read to ensure the quality of our evaluation,. Since our dictionary lacks many proper nouns or technical terms, such as API (空氣

污染指數, air pollution index), Tsing Yi (青衣, name of a place in Hong Kong) and SMEs(中小型企業, small and medium enterprises), we use a threshold to overcome the weakness. We set the threshold to be 4 which means if the Chinese translations of 4 or more English words in a English title appear in a Chinese title, then the English title is considered as an element in the alignment matrix. (E matches C) is a good tool for the threshold, e.g. $\sum_i \text{Count}(e_i) \geq 4$. The advantage of this threshold mainly prevents the English titles wrongly aligning with Chinese titles.

Since some documents are only written in one language either English or Chinese, the threshold can effectively ensure an English title aligning to a correct Chinese title. In addition, in some cases, the variation between two short English titles is a proper noun. For example, Kolwoon salt water stoppage and Shatin salt water stoppage. The threshold can relieve the variation problem for short titles.

The precision and recall from 1st May,1999 to 31st January, 2001 based on threshold=4 are 99.8% and recall at 63.7% respectively. The high precision rate is mainly contributed by two reasons:1) The titles are separated by date. So there are few titles talking about the same topic. 2) Most of the titles published in one day are distinct.

One reason for the low recall rate is that in one title homepage(the title homepage lists all the titles of press release article of a day), it is possible that two or more titles are identical, e.g. several weather report titles were published in a day, or an amended version of an article together with its earlier version were published in one day. Content of the articles are different but their titles are identical. To solve the problem, we take a title T1 from the top of title homepage. If there are any following titles in the title homepage that are identical to T1, these titles will be ignored and only T1 is considered in the title alignment algorithm. The advantage of the method is divided into two parts:

1) The press release articles are published based on time and their filenames are the time they generated. As a result, the titles at the top of title homepage is the latest version of the article.

2) These articles are normally weather reports, fire report or air pollution index reports. Also, the content of these articles are very similar. The large number of these articles will not be useful in our corpus-based techniques in the future.

The other reason for low recall rate is because our dictionary lacks many proper nouns or technical terms.

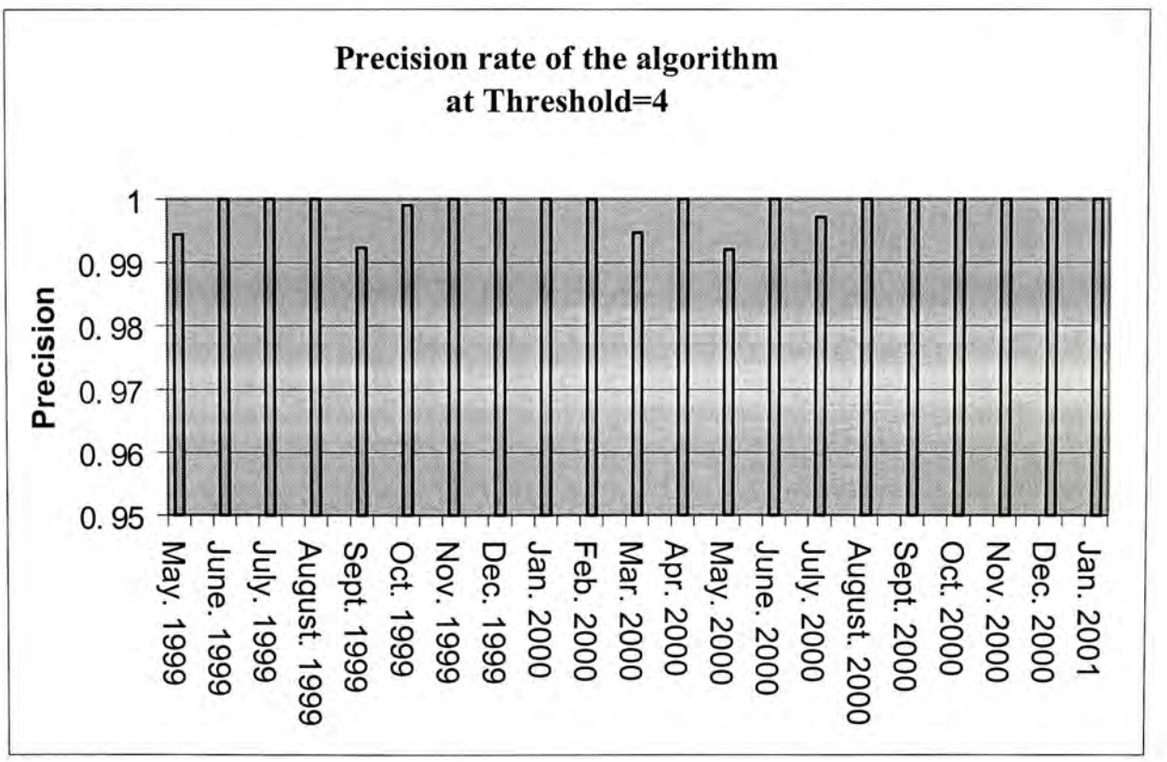

Figure 5 The precision rate for Hong Kong government press release articles based on the threshold 4 .

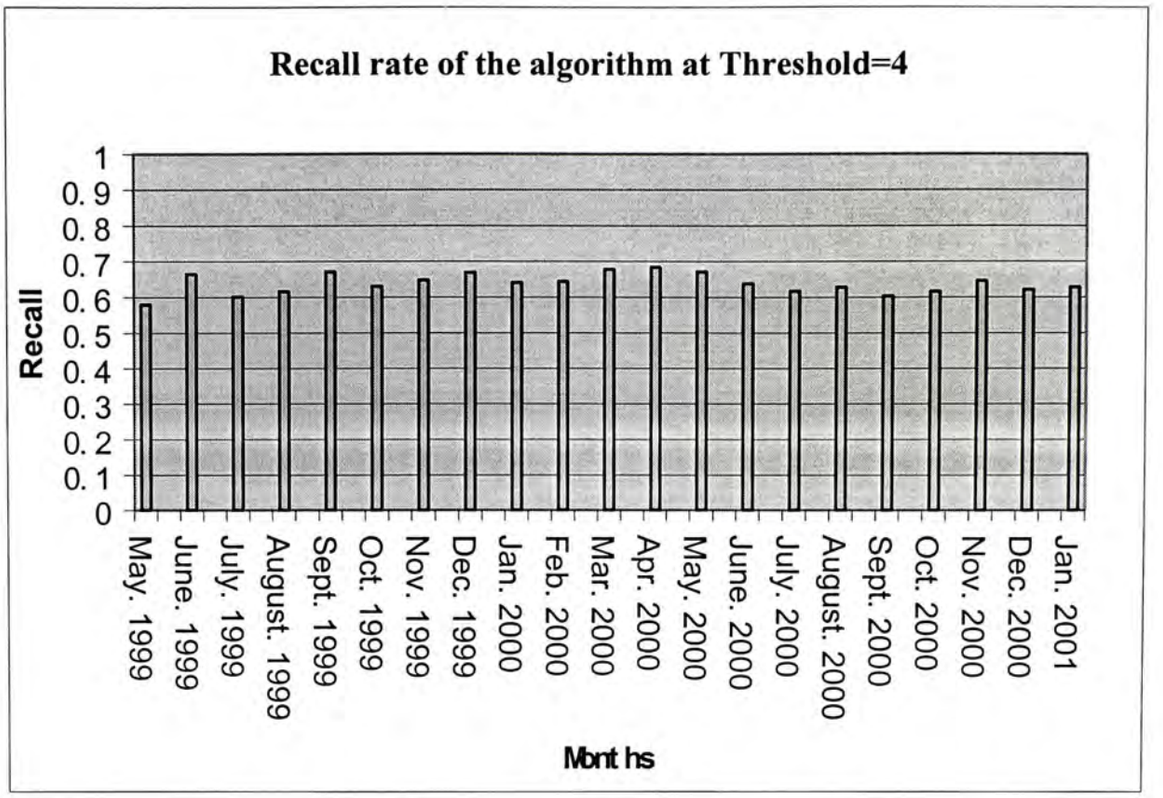

Figure 6 The recall rate for Hong Kong government press release articles based on the threshold 4.

Figure 5 and Figure 6 show that the title alignment algorithm can effectively construct a parallel corpus based on the bilingual titles.

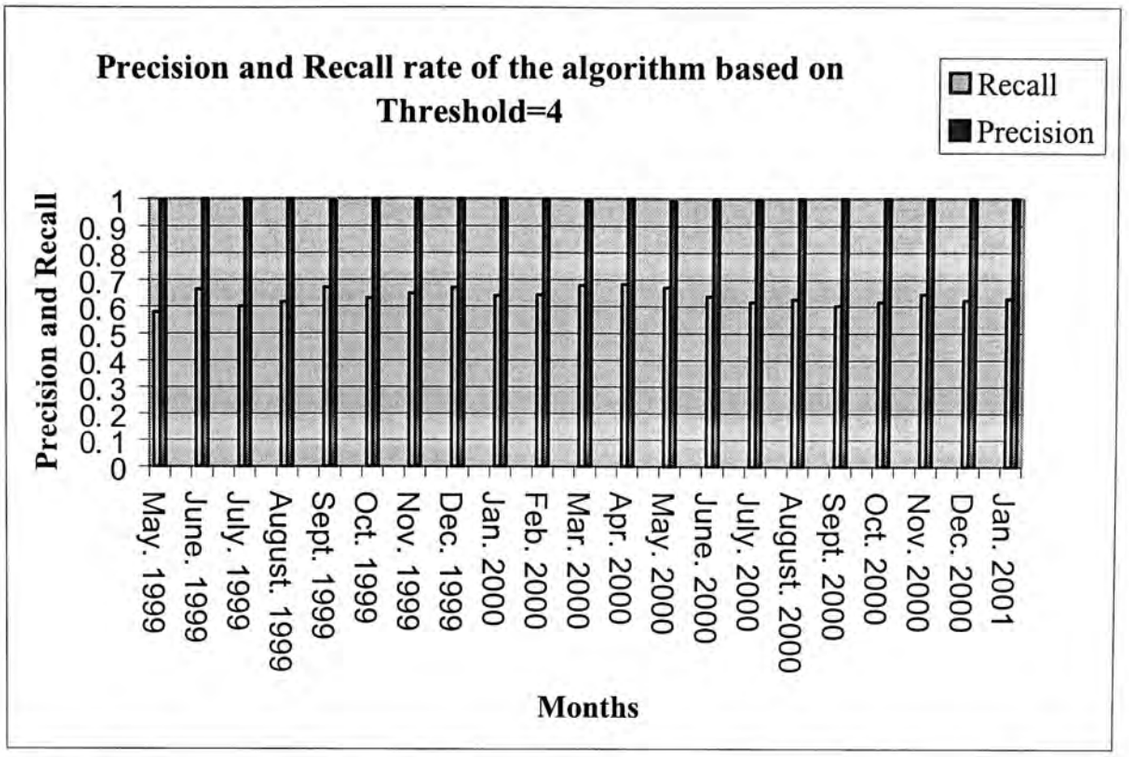

Fig. 7 Precision and recall rate of the title alignment algorithm based on threshold=4

An example of the retrieved result on 1st May, 1999 is shown in the Figure 8.

1 $100,000 reward for information on murder in Tuen Mun
1 警方懸賞十萬元呼籲市民提供屯門謀殺案消息

2 Services for employers and employees enhanced
2 勞工處加強對僱員和僱主服務

3 Trade Director to attend APEC Senior Officials Meeting
3 貿易署署長出席亞太經合組織高級官員會議

4 Public holiday clinic service
4 公眾假期門診服務

5 Announcement on the Air Pollution Index (API) by EPD
5 環境保護署公布空氣污染指數

6 34 men arrested for unlawful assembly
6 三十四名男子涉嫌非法集會被捕

7 Stamp distribution network expanded for added convenience
7 香港郵政擴大郵票小冊銷售網絡為市民提供額外方便

Figure 8 An example of the government press release retrieved result for 1st May, 1999

We also use the F-measure([LSV98]) which combines recall and precision in a single efficiency measure:

$$F = 2\frac{(\text{recall} \times \text{precision})}{(\text{recall} + \text{precision})}$$

The F measure in this case is

$$F = 2\frac{(63.7\% \times 99.8\%)}{(63.7\% + 99.8\%)} = 0.778$$

5.2 Hang Seng Bank economic monthly report

We also test the proposed title alignment algorithm on Hang Seng Bank economic monthly report. Unlike the government press release articles where the articles were separated by date, all the monthly report titles for one language are displayed in one title homepage. The title homepage for one language is anchored with the title homepage in another language. All the economic monthly reports are put into http://main.hangseng.com/econ/mon/. In addition, an economic monthly report in one language is anchored to its counterpart in another language.

To evaluate the economic reports, it is easier than evaluating the government press release articles because Chinese version of an economic report gets a "c" at the end of the filename and English version of that report gets a "e" at the end of the filename. For example, the Chinese file "m0201c.html" is the counterpart of the English file "m0201e.html". In addition, we set the threshold to be 0. The reason for this is that unlike some government press release articles that may only be written in one language, all the economic reports are written in two languages.

In the Chinese title page (http://main.hangseng.com/econ/mon/monc.html), there are 47 Chinese economic monthly report titles available for retrieval starting from August, 1996 to February, 2001. There are also 47 English economic monthly report

titles available for retrieval in the English title page (http://main.hangseng.com/econ/mon/mone.html),

According to the equation 17 and 18, the precision and recall in this case are:
Precision= 42 / 42= 100%
Recall= 42 / 47= 89.36 %

The system retrieved 42 title pairs and these 42 title pairs are correctly aligned. The high precision rate is contributed partially by the date. Date is printed at the end of each title, e.g. the title "The Future of the Renminbi (August 2000)". The date helps the English/Chinese alignment system to determine the correct alignment even though many proper noun, such as Renminbi(currency of China), are not appeared in our dictionary.

There are totally 47 title pairs listed in the title homepages but only 42 pairs were retrieved by the English/Chinese alignment system. The reasons for this are divided into two folds:

1) Our dictionary lacks many proper nouns, e.g. euro(歐 羅).

2) A significant conceptual alteration between an English title and its counterpart. For example, the counterpart of the English title "Economic Stability in the Year of Political Transition (November,1996)" is "一九九七年香港經濟展望 (一 九 九 六 年 十 一 月)"(Outlook of Hong Kong in1997).

Translated(Economic)={經濟} Translated(Stability)={ 穩定}
Translated(Year)= {年} Translated(Political)={政治}
Translated(Transition)={ 過渡}

Without reading their contents, the two titles looks like talking about two different things.

1 A Winding Road to Recovery (May 1999)
1 待上坦途的香港經濟(一九九九年五月)

2 The Asian Currency Turmoil - Impact on Hong Kong's Exports in 1998 (January/February 1998)
2 亞洲貨幣風潮對一九九八年香港出口之影響(一九九八年一月/二月)

3 Housing the Unaffordable: a Policy Dilemma (February 2001)
3 公屋居屋進退維谷(二零零一年二月)

4 Hong Kong's Export Performance - Short-term Trends and Long-term Outlook (August 1996)
4 香港出口表現一短期動向及長期展望(一九九六年八月)

5 The Future of the Renminbi (August 2000)
5 人民幣的前景(二零零零年八月)

6 Falling Residential Property Values and Repercussions on the Economy (May 1998)
6 住宅物業價格下跌對經濟體系的影響(一九九八年五月)

7 The Office Property Market Adjustment (August 1998)
7 寫字樓物業市場的調整(一九九八年八月)

8 Real Dynamics of Competition (February 2000)
8 競爭動力的真諦(二零零零年二月)

9 Advancing Hong Kong's Bond Market (March 2000)
9 推動香港的債券市場(二零零零年三月)

Figure 9 Some Hang Seng bank economic report title pairs retrieved by the English/Chinese alignment system

The F measure in this case is

$$F = 2\frac{(89.36\% \times 100\%)}{(89.36\% + 100\%)} = 0.9438$$

5.3 Hang Seng Bank press release articles

In addition, we test the title alignment algorithm on Hang Seng Bank press release articles. Same as the web site arrangement of economic monthly report, all the press release article titles for one language are displayed in one title homepage. The title homepage for one language is anchored with the title homepage in another language.

All the press release articles for both languages are put into http://main.hangseng.com/new/rel/ . A press release article in one language is anchored to its counterpart in another language.

As easy as evaluating the economic reports, Chinese version of a press release article gets a "c" at the end of the filename and English version of that article gets a "e" at the end of the filename. For example, the Chinese file "a043099c.html" is the counterpart of the English file "a043099e.html". All press release articles are written in two languages. So we set the threshold to be 0.

In the Chinese title page (http://main.hangseng.com/new/rel/relc.html), there are 184 Chinese press release titles available for retrieval starting from 5th August, 1996 to 9th February, 2001. There are also 184 English press release titles available for retrieval in the English title page (http://main.hangseng.com/new/rel/rele.html),

According to the equation 17 and 18, the precision and recall in this case are:
Precision= 111 /115 = 96.52%

Recall= 111/184 = 60.3%

The English/Chinese alignment system retrieves 115 titles pairs but only 111 title pairs are correct. The reason for low precision rate is that many titles are highly correlated to one another. These articles are talking about the same topic in different time. This causes the English/Chinese alignment system wrongly align the titles. For example, the system wrongly align the English title "Hong Kong Paralympic Medallists Reap Cash Awards For Outstanding Performance (14 November 2000)" with the Chinese title "香港傷殘奧運健兒以傑出成績贏得現金獎勵(二零零零年十一月一日)(Hong Kong Paralympic athletes receive cash awards for outstanding performance(1st Nov. 2000))" where the right Chinese title should be "香港傷殘人士奧運獎牌得主喜獲現金獎勵(二零零零年十一月十

四 日)(Hong Kong Paralympic Medallists gladly receive their cash awards(14 Nov.,2000)".

Translated(Hong)= {香} Translated(Kong)= {港}

Translated(Paralympic) ={傷 殘 人 士 奧 運 會}

Translated(Medallists)={ 獎 牌 得 主} Translated(Reap)= {獲得, 獲, 得}

Translated(Cash) ={現 金} Translated(Awards)= {獎 勵}

Translated(Outstanding)= {傑 出} Translated(Performance)={成 績}

There are totally 184 title pairs on the web site and the English/Chinese alignment system successfully retrieves 111 pairs.

One cause for low recall is that many technical terms in the financial domain are not available in our dictionary, such as "E-shopping"(shopping through the internet), "Mondex"(a kind of electronic cash card), "Supercash"(a kind of loans in Hong Kong), "CU Link"(a multi-function card for the staff and students of the Chinese University of Hong Kong).

The F measure in this case is

$$F = 2\frac{(60.3\% \times 96.52\%)}{(60.3\% + 96.52\%)} = 0.7423$$

1 Hang Seng Reduces Prime Rate (30 April 1999)
1 恒生銀行調低最優惠利率(一九九九年四月三十日)

2 Hang Seng Bank to Open 10 Branches at Airport Express and Tung Chung Line Stations in July (6 July 1998)
2 恒生銀行七月在機場快線及東涌線開設十間分行(一九九八年七月六日)

3 Technology Experts Provide Investment Insight At Hang Seng Bank Seminar (10 November 1999)
3 科技專家於恒生銀行研討會提供投資錦囊(一九九九年十一月一日)

4 New Appointment To Hang Seng Bank Board (16 August 1999)
4 恒生銀行委任新董事(一九九九年八月十六日)

5 Hang Seng Credit Card Launches New Membership Rewards Programme (5 February 2001)
5 恒生信用卡推出全新會員獎賞計劃(二零零一年二月五日)

6 Hang Seng Bank Opens Causeway Bay Branch at Redeveloped Building (18 August 1998)
6 恒生銀行在重建物業開設銅鑼灣分行(一九九八年八月十八日)

7 Hang Seng Bank Limited 1996 Interim Results - Highlights (5 August 1996)/ul
7 恒生銀行有限公司一九九六年中期業績摘要(一九九六年八月五日)

8 Hang Seng Bank And IBM Offer Business Solutions To Customers (5 October 1999)
8 恒生銀行與IBM合作為客戶提供商務錦囊(一九九九年十月五日)

Figure 10 Some Hang Seng bank press release title pairs retrieved by the English/Chinese alignment system

5.4 Hang Seng Bank speech articles

Same as the web site arrangement of economic monthly report, all the speech article titles for one language are displayed in one title homepage. The title homepage for one language is anchored with the title homepage in another language. All the press release articles for both languages are put into http://main.hangseng.com/new/spee . A speech article in one language is anchored to its counterpart in another language.

Same as other articles published by Hang Seng Bank, Chinese version of a speech article gets a "c" at the end of the filename and English version of that article gets a "e" at the end of the filename.

In the Chinese title page (http://main.hangseng.com/new/spee/speec.html), there are 11 Chinese press release titles available for retrieval starting from 18th May, 1998 to 9th November, 2000. There are also 11 English press release titles available for retrieval in the English title page (http://main.hangseng.com/new/spee/speee.html),

The threshold is 0 for this case because all the speech articles in one language has a counterpart in another language.

According to the equation 17 and 18, the precision and recall rates are:
Precision= 11/11=100%
The system retrieves 11 title pairs and 11 pairs are correctly aligned.
Recall=11/11=100%
There are 11 title pairs but the English/Chinese alignment system simply retrieves 11 pairs.

1 Speech by Mr Vincent H C Cheng, Vice-Chairman and Acting Chief Executive of Hang Seng Bank, at the CLSA Investors' Forum Asia 98 (18 May 1998)
1恒生銀行副董事長兼署理行政總裁鄭海泉先生在「里昂證券亞洲投資者研討會98」上的演辭全文(一九九八年五月十八日)

2 Statement by Mr Vincent H C Cheng, Vice-Chairman and Chief Executive of Hang Seng Bank, at the Bank's 1998 Interim Results Announcement (3 August 1998)
2恒生銀行副董事長兼行政總裁鄭海泉先生在該行「一九九八年公佈中期業績記者會」上的講辭全文(一九九八年八月三日)

3 Speech by Mr Vincent H C Cheng, Vice-Chairman and Chief Executive of Hang Seng Bank, at the 2000 WESTERN FORUM OF CHINA (21 October 2000)
3恒生銀行副董事長兼行政總裁鄭海泉先生於「中國西部論壇」的講辭全文(二零零零年十月二十一日)

4 Speech by Mr Vincent H C Cheng, Vice-Chairman and Chief Executive of Hang Seng Bank, at the Credit Suisse First Boston Asian Investment Conference (24 March 1999)
4恒生銀行副董事長兼行政總裁鄭海泉先生在「瑞士信貸第一波士頓投資會議」上的演辭全文(一九九九年三月二十四日)

5 Speech by Mr Vincent H C Cheng, Vice-Chairman and Chief Executive of Hang Seng Bank, at the Ceremony of Honorary Visiting Professor Appointment by Zhejiang University (16 September 2000)
5恒生銀行副董事長兼行政總裁鄭海泉先生於浙江大學授予客座教授儀式的講辭全文(二零零零年九月十六日)

Figure 11. Some Hang Seng bank speech title pairs retrieved by the English/Chinese alignment system

The F measure in this case is

$$F = 2\frac{(100\% \times 100\%)}{(100\% + 100\%)} = 1$$

Even though there are many proper noun in each title, we still can retrieve high precision and recall rate. The reason for this is divided into two folds: 1) There are only 11 title pairs. 2) Retrieving the titles mainly base on date.

5.5. Quality of the collections and future work

Name of corpus	Precision	Recall	F-measure
Hong Kong government press release article	99.8%	63.7%	0.778
Hang Seng Bank economic monthly report	100%	89.36%	0.9438
Hang Seng Bank press release article	96.52%	60.3%	0.7423
Hang Seng Bank speech article	100%	100%	1

Table 2 The precision and recall rates for each corpus

Table 2 shows that the title alignment algorithm can effectively align the Chinese and English titles based on the same dictionary.

Name of corpus	Total number of Chinese articles available	Total number of English articles available	Total number of Chinese/ English pairs available	Total number of correct Chinese/ English pairs retrieved by the system
Hong Kong government press release article	16165	15732	13543	8898
Hang Seng Bank economic monthly report	47	47	47	42
Hang Seng Bank press release article	184	184	184	111
Hang Seng Bank speech article	11	11	11	11

Table 3 The number of articles in each corpus

The low recall rate reveals a weakness of the title alignment algorithm. The English/Chinese title alignment algorithm is designed to ensure that an English title aligns with its best Chinese counterpart. If an English title E1 finds out that an Chinese title C1 is its counterpart. However, an English title E2 also finds out that the Chinese title C1 is its counterpart and the score for E2 and C1 is higher than the score for E1 and C1. Then the title alignment algorithm will decide that E2 and C1 should be aligned. To ensure the high precision rate, E1 will not align with its second best counterpart.

All the articles and their titles in four different collections can be classified as dissemination class of translation([MBD99]). **Dissemination** refers to the class of translation in which an individual or organization wants to broadcast his or her own material, written in one language, in a variety of language to the world. The other class is **assimilation** which refers to the class of translation in which an individual or organization wants to gather material written by others in a variety of languages and convert them all into his or her own language. According to Maegaard et al.([MBD99]), the quality of documents in dissemination class is high. Statistical or probabilistic methods based on high quality corpus are expected to produce better outputs.

Future work

To improve the performance, we will try to apply the title pairs to construct a bilingual lexicon by using statistical method. Also, we will use the corpora in machine translation and cross-lingual information retrieval.

Furthermore, since a Hong Kong government press release article may cause other media to publish some articles related to it, we will extend the English/Chinese alignment system to construct a large comparable corpus by aligning the bilingual

titles from different media with similar topics. The reason for this is that while a corpus-based technique developed from a parallel document collection can in principle be used for unrelated applications as well, significant reductions in retrieval effectiveness should be expected. Techniques based on comparable document collections may overcome this limitation([Oar97]). During the alignment of Hang Seng Bank press release titles, the result shows that the title alignment algorithm can also be used in aligning comparable titles because for the four wrongly aligning pairs, all the pairs are highly related to each other.

Also, since the low recall rate is mainly caused by technical terms or proper noun, we will try to find out a solution for this in the future.

Finally, as mentioned in Resnik([Res99]), we will try to develop a Web crawler to mine the web sites for candidate title pairs.

Chapter 6

Conclusion

The increasing need of access to global information has made multilingual corpora to become a valuable linguistic resource for many natural language processing applications. The general-purpose dictionary is less sensitive in genre and domain. As it can be impractical to manually construct tailored bilingual dictionaries or sophisticated multilingual thesauri for large applications, corpus-based approaches provide a statistical translation model to cross the language boundary.

Many domain-specific parallel or comparable corpora are employed in machine translation and cross-lingual information retrieval. Since the number of Asian/Indo-European corpus, especially English/Chinese corpus is relatively deficient, we had presented a title alignment method relied on dynamic programming to identify the one-to-one Chinese and English title pairs and construct a parallel corpus by downloading the texts accordingly. The method uses the fact that a title is a representation of a text. We has applied the title alignment method to automatically construct four Chinese/English parallel corpora.

The English/Chinese title alignment method includes alignment at title level, word level and character level. The longest common subsequence (LCS) is applied to find the most reliable Chinese translation of an English word. As one word may translate into two or more words, deletion, an edit operation is used to reduce overlapping. Also, a score function has proposed to find the optimal title pairs.

Experimental results show that the title alignment method obtains over 95%of precision and over 60% of recall and over 0.74 of F-measure.

Bibliography

[All99] Allison, L. "Dynamic Programming Algorithm (DPA) for Edit-Distance". In *Algorithms and Data Structures Research & Reference Material.* School of Computer Science and Software Engineering, Monash University, Australia 3168, c1999.
http://www.csse.monash.edu.au/~lloyd/tildeAlgDS/Dynamic/Edit.html

[APD99] Allison, L., Powell, D., Dix, T. I. "Compression and approximate matching". In *The Computer Journal*, Volume 42, Issue 1, pp. 1-10. 1999

[All98] Allison, L. "Information-Theoretic Sequence Alignment". Technical Report 98/14 School of Computer Science and Software Engineering, Monash, University, June 1998

[AED98] Allison, L., Edgoose, T., Dix, T. I. "Compression of Strings with Approximate Repeats". In *Intelligent Systems in Molecular Biology* (ISMB'98), pp8-16, Montreal, 28 June - 1 July 1998

[AD86] Allison, L., Dix, T. I. "A bit-string longest common subsequence algorithm". In *Information Processing Letters*, 23(6), pp305-310, 1986

[As99] Aston, G. "Corpus use and learning to translate". In *Textus* 12: 289-314. 1999.

[At99] Atallah, M. J. *Algorithms and Theory of Computation Handbook*, CRC Press LLC, 1999. ISBN: 0849326494

[Ba96] Barlow, M. "Parallel Texts in Language Teaching". In S. Botley, J. Glass, T. McEnery and A. Wilson (Eds), *Proceedings of Teaching and Language Corpora* 1996. UCREL Technical Papers, 9, Lancaster, UCREL, pp. 45-56. 1996

[BLM91] Brown, P., Lai, J., and Mercer, R. "Aligning sentences in parallel corpora". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, USA, 1991

[BCD90] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. "A statistical approach to machine translation". In *Computational Linuistics*, 16(2):79-85, 1990

[CCF99] Calzolari, N., Choukri, K., Fellbaum, C., Hovy, E., Ide, N. "Multilingual Resources". In *Multilingual Information Management: Current Levels and Future Abilities*, Chapter 1 Multilingual Resources, pp.8-26, 1999
http://www.cs.cmu.edu/~ref/mlim/

[CT95] Collier, N. and Takahashi, K. "Sentence alignment in parallel corpora: The asahi corpus of newspaper editorials". Technical Report 95/11, Centre for Computational Linguistics, UMIST, Manchester, October 1995.

[CLR90] Cormen, T. H., Leiserson, C. E., Rivest, R. L. *Introduction to Algorithms*. MIT Press, ISBN: 0262031418, pp314-319, 1990.

[CRG96] Chawathe, S., Rajaraman, A., Garcia-Molina, H., Widom, J.. "Change detection in hierarchically structured information". In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 493--504, Montr'eal, Qu'ebec, June 1996.

[Che93] Chen, S. F. "Aligning Sentences in Bilingual Corpora using Lexical Information". In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 9-16, 1993.

[CKJ99] Chen, A., Kishida, K., Jiang, H., Liang, Q., Gey, F. "Automatic Construction of a Japanese-English Lexicon and its Application in Cross-Language Information Retrieval". In *Proceedings of the Multilingual Information Discovery And Access workshop of the ACM SIGIR'99 Conference*, August 14, 1999

[Chu93] Church, K. W. "Char_align: A Program for Aligning Parallel Texts at the Character Level". In *Proceedings of ACL-93*, Columbus OH,1993

[CDG93] Church, K. W., Dagan, I., Gale, W., Fung, P., Helfman, J., Satish, B. "Aligning Parallel Texts: Do Methods Developed for English-French Generalize to Asian Languages?". In *Proceedings of Pacific Asia Conference on Formal and Computational Linguistics*, pp.1-12,1993

[DD95] Davis, M., and Dunning, T. "Query translation using evolutionary programming for multilingual information retrieval". In *Proceedings of the Fourth Annual Conference on Evolutionary Programming*,1995

[DC94] Dagan, I. and Church, K. W. "Termight: Identifying and translating technical terminology". In *Proceedings of the Fourth Conference on Applied Natural LanguageProcessing*, 34--40, Stuttgart,1994

[DCG93] Dagan, I., Church, K. W., Gale, W. "Robust Bilingual Word Alignment for Machine Aided Translation". In *Proceedings of the Workshop on Very Large Corpora*, Columbus, Ohio, 1993, pages 1—8

[Dun94] Dunning, T. "Statistical identification of language". In *Computing Research Laboratory technical memo MCCS 94-273*, New Mexico State University, Las Cruces, New Mexico,1994

[Ebe98] Ebeling, Jarlie. "Contrastive Linguistics, Translation, and Parallel Corpora". In *Meta*, Vol 43, Issue 4, pp.602-615, 1998.

[Eij93] Eijk, P. "Automating the acquisition of bilingual terminology". In *Proceedings of Meeting of the European Chapter of the Association for Computational Linguistics*, 113-119. 21-23 April, Utrecht,1993.

[EC95] Ejerhed, E., and Church, K. "Language Resources". In *Survey of the State of the Art in Human Language Technology*, Chapter 12 Language Resources, Cambridge University Press ISBN 0-521-59277-1. pp.441- 474,1999

[EDR] Electronic Dictionary Research (EDR) http://www.iijnet.or.jp/edr/

[ELD01] ELDA. "Multilingual and Parallel Corpora", In European Language Resources Association (ELRA) home page, 2001: http://www.elda.fr/cata/tabtext.html#multilex, W0023, MLCC, Multilingual and Parallel Corpora

[EGGI93] Eppstein, D., Galil, Z., Giancarlo, R., Italiano, G. F. "Efficient algorithms for sequence analysis". International Advanced Workshop on Sequences, Positano, Italy, 1991. In *Sequences II: Methods in Communication, Security, and Computer Science*, R.M. Capocelli, A. De Santis, and U. Vaccaro, eds., Springer-Verlag, 1993, pp. 225-244.

[Epp96] Eppstein, D. "Longest Common Subsequence" . *ICS 161: Design and Analysis of Algorithms, Lecture Notes*, Winter 1996. http://www.ics.uci.edu/~eppstein/161/960229.html

[FIP96] Foster, G., Isabelle, P., Plamondon, P. "Word completion: A first step toward target-text mediated IMT". In *Proceedings of COLING-96*, Copenhagen, Denmark, 394-399,1996

[Fun95] Fung, P. "A Pattern Matching Method for Finding Noun and Proper Noun Translations from noisy Parallel Corpora". In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Boston, MA, 1995.

[FM94] Fung, P. and McKeown, K. "K-vec: A New Approach for Aligning Parallel Texts". In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto/Japan, 1994.

[FM97] Fung, P. and McKeown, K. " A technical word- and term-translation aid using noisy parallel corpora across language groups". In *Machine Translation* 12: 53-87, 1997

[GC91] Gale, W. A., and Church, K.W. "Identifying word correspondences in parallel texts". In *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language*, Asilomar, California,1991

[Gut00] Gutt, Ernst-August. *Translation and Relevance*. Cognition and Context. St. Jerome Publishing, ISBN 1-900650-29-0.pp.45-65, April, 2000.

[GZ95] Godfrey, J. J. and Zampolli, Antonio "Language Resources". In *Survey of the State of the Art in Human Language Technology,* Chapter 12 Language Resources, Cambridge University Press ISBN 0-521-59277-1. pp.441- 474, 1995

[He00] He, S. "Translingual Alteration of Conceptual Information in Medical Translation: A Cross-Language Analysis between English and Chinese". In *Journal of the American Society for Information Science*, Vol. 51, No. 11, pp.1047-1060,2000.

[IDF93] Isabelle, P., Dymetman, M., Foster, G., Jutras, J-M., Macklovitch, E., Perrault, F., Ren, X., Simard, M. "Translation Analysis and Translation Automation". In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan,1993

[KR99] Kanungo, T. and Resnik, P. "The Bible, Truth, and Multilingual OCR Evaluation". In *Proceeding of SPIE Conference on Document Recognition and Retrieval (VI)*, San Jose, CA, 27-28 January, 1999.

[KR93] Kay, M., & Röscheisen, M."Text-translation alignment". In *Computational Linguistics*, 19(3), 121-142,1993

[KT90] Klavans, J., and Tzoukermann, E. "The BICORD System: Combining lexical information from bilingual corpora and machine readable dictionaries". In *Proceedings of the Thirteenth International Conference on Computational Linguistics (COLING-90)*, pp.174-179, 1990

[KT96] Klavans, J. and Tzoukermann, E. "Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons". In *Machine Translation*, 1996.

[KHF99] Klavans, J., Hovy, E., Fluhr, C., Frederking, R. E., Oard, D., Okumura, A., Ishhikawa, K., and Satoh, K. "Multilingual (or Cross-lingual) Information Retrieval". In *Multilingual Information Management: Current Levels and Future Abilities*, Chapter 2 Multilingual (or Cross-lingual) Information Retrieval, 1999.

[Ku93] Kupiec, J. "Algorithm for finding noun phrase correspondences in bilingual corpora". In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, Columbus, Ohio, pp.17-22, 1993

[Kur01] Kurth, Frank. "Comparison of Melodies with LCS-Algorithms". The MiDiLiB project. http://leon.cs.uni-bonn.de/forschungprojekte/midilib/english/notes2lcs.html

[Lar98] Larson, M. L. *Meaning-based translation: A guide to cross-language equivalence.* Lanham, MD: University Press of American

[Leo00] Leonardi, Vanessa. "Equivalence in Translation: Between Myth and Reality". In *Translation Journal*, Vol. 4, No.4, October, 2000.

[LSV98] Langlais, P., Simard, M., Véronis, J. "Methods and Practical Issues in Evaluating Alignment Techniques". In *Proceedings of COLING-ACL 98*, Montréal, Québec, 1998.

[LDC] LDC Linguistic Data Consortium (LDC) home page. World Wide Web page. http://www.ldc.upenn.edu/

[LC96] Lin, C. and Chen, H. "An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) Documents". In *IEEE Transactions on Systems, Man, and Cybernatics - Part B: C Cybernetics*,VOL.26, NO.1 February , 1996

[MH96] Macklovitch, E., Hannan, Marie-Louise "Line'Em Up: Advances In Alignment Technology And Their Impact on Translation Support Tools". In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*, Montréal, Québec,1996.

[MBD99] Maegaard, B., Bel, N., Dorr, B., Hovy, E., Knight, K., Iida, H., Boitet, C., Maegaard, B., Wilks, Y. "Machine Translation". In *Multilingual Information Management: Current Levels and Future Abilities,* Chapter 4 Machine Translation. 1999.

[MSM93] Marcus, M. P., Santorini, B., Marcinkiewicz, M. A."Building a large annotated corpus of English: the Penn Treebank". In *Computational Linguistics*, vol. 19, 1993.

[MM98] Melamed I. D., Marcus M. P. "Automatic Construction of Chinese-English Translation Lexicons", IRCS Technical Report #98-28, 1998.

[Mel95] Melamed I. D. "Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons". In *Proceedings of the 3rd Workshop on Very Large Corpora*, Boston/Massachusetts, 1995.

[Mel96a] Melamed, I. D. "Automatic Construction of Clean Broad-Coverage Translation Lexicons". In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA'96),* Montreal, Canada, 1996

[Mel96aa] Melamed, I. D. "Automatic Detection of Omissions in Translations". In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, 1996.

[Mel96b] Melamed, I. D. "A Geometric Approach to Mapping Bitext Correspondence". IRCS Technical Report #96-22. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96),* Philadelphia, PA, May,1996

[Mel97] Melamed, I. D. " A Portable Algorithm for Mapping Bitext Correspondence". In *Proceedings of 35th Conference of the Association for Computational Linguistics (ACL'97*), Madrid, Spain,1997

[Mel97b] Melamed, I. D. " A Word-to-Word Model of Translational Equivalence". In *Proceedings of 35th Conference of the Association for Computational Linguistics (ACL'97),* Madrid, Spain,1997

[Mel98] Melamed I. D. "Manual Annotation of Translational Equivalence: The Blinker Project", IRCS Technical Report #98-07, 1998.

[NBR96] Nie, J.Y., Brisebois, M., Ren, X. "On Chinese text retrieval". In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zürich, Switzerland, August 1996.

[NSI99] Nie, J.Y., Simard M., Isabelle, P., Durand, R. "Cross-language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web". In *ACM SIGIR'99 8/99* Berkley, CA USA,1999

[NGZ00] Nie, J.Y., Gao J., Zhang J., Zhou M. "On the use of Words and N-grams for Chinese Information Retrieval". In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, Hong Kong, 2000.

[Oar96] Oard, D. W. "Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications". Ph.D. Dissertation, University of Maryland, College Park,1996.

[OD96] Oard, D. W. and Dorr, B. J. *A Survey of Multilingual Text Retrieval.* UMIACS-TR-96-19 CS-TR-3815,1996.

[Oar97] Oard, D. W. "Alternative approaches for cross-language text retrieval". In Hull D, Oard D,(Eds.) ,*1997 AAAI Symposium in Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, March 1997.

[OH92] Orpen, K. S. and Huron, D. "The Measurement of Similarity in Music: A Quantitative Approach for Non-parametric Representations". In *Computers in Music Research*, Vol. 4 (1992): pp1-44.

[PCC99] Palmer, M., Calzolari, N., Choukri, K., Fellbaum, C., Hovy, E., Ide., N. "Multilingual Resources". In *Multilingual Information Management: Current levels and Future Abilities*, Chapter 1 Multilingual Resources , April, 1999.

[Pen] The Penn Treebank Project http://www.cis.upenn.edu/~treebank/home.html

[PP97a] Peters C. and Picchi, E. "Using linguistic tools and resources in cross-language retrieval". In *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, March 1997.

[PP97b] Picchi, E. and Peters C. " Reference Corpora and Lexicons for Translators and Translation Studies". In A. Trosberg (Ed.), *Text Typology in Translation Studies*, Amsterdam and Philadelphia, John Benjamins Publishing Company,1997

[P93] Pim van der Eijk. "Automating the Acquisition of Bilingual Terminology". In *Proceedings of the 6th Conference of the European Chapter of the ACL*, Utrecht/The Netherlands, 1993. Association for Computational Linguistics.

[PAD00] Powell, D. R., Allison L., Dix, T. I. "Fast, optimal alignment of three sequences using linear gap costs". In *Journal of Theoretical Biology*, Vol. 207, No. 3, Dec 2000, pp. 325-336.

[Res98] Resnik P. "Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text". In Farwell D., Gerber L., and Hovy E. (eds.), *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98)*, Langhorne, PA, Lecture Notes in Artificial Intelligence 1529, Springer, October, 1998.

[Res99] Resnik P. "Mining the Web for Bilingual Text". In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland, June 1999.

[RM97] Resnik, P., Melamed, I D. "Semi-Automatic Acquisition of Domain-Specific Translation Lexicons". In *Proceedings of the 5th ANLP Conference*, 1997.

[ROD99] Resnik P., Olsen M. B., Diab M. "The Bible as a Parallel Corpus: Annotating the `Book of 2000 Tongues'" , In *Computers and the Humanities*, 33(1-2), pp.129-153, 1999.

[Ros81] Rose, Marilyn Gaddis. "Translation Types and Conventions". In *Translation Spectrum: Essays in Theory and Practice*, Marilyn Gaddis Rose, Ed., State University of New York Press, pp.31-33, 1981,

[RY97] Ristad, E. S., and Yianilos, P. N. "Learning string edit distance". In *Machine Learning: Proceedings of the Fourteenth International Conference* (San Francisco, July 8--11 1997), D. Fisher, Ed., Morgan Kaufmann, pp. 287—295, 1997

[Sal70] Salton,G. "Automatic processing of foreign language documents". In *Journal of the American Society for Information Science*, 21(3):187-194, May, 1970.

[SB96] Sheridan, P. and Ballerini, J.P. "Experiments in multilingual information retrieval using the SPIDER system", In *Proceedings of the 19th ACM SIGIR Conference*, 58- 65., 1996.

[SS97a] Skiena, S. S., Skiena, S.(1997a) "Longest common substring". In *The Algorithm Design Manual* 1997 ISBN 0-387-94860-0

[SS97b] Skiena, S. S., Skiena, S.(1997) "Approximate String Matching". In *The Algorithm Design Manual* 1997 ISBN 0-387-94860-0

[Sim99] Simard, M. "Text-translation Alignment: Three Languages Are Better Than Two". In *Proceedings of EMNLP/VLC-99*. College Park, MD.1999

[SFP93] Simard, M., Foster, G. F., Perrault F. "TransSearch: A Bilingual Concordance Tool", Centre d'innovation en technologies de l'information, Laval, Canada. (1993). http://www-rali.iro.umontreal.ca/Publications.fr.html

[SFI92] Simard, M., Foster, G., Isabelle P. "Using Cognates to Align Sentences in Bilingual Corpora". In *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, Montreal, Canada.1992

[Sma93] Smadja F. "Retrieving Collocations from Text: XTRACT". In *Computational Linguistics*, 1993.

[SMH96] Smadja, F., McKeown, K., Hatzsivassiloglou, V. "Translating collocations for bilingual lexicon: A statistical approach". In *Computational Linguistics* 22, 1-38, 1996

[TSB97] Takahashi, Y., Shirai S., Bond, F. "A method of automatically aligning Japanese and English newspaper articles". In *Natural Language Processing Pacific Rim Symposium '97*: NLPRS-97, 657-660. 1997.

[WR90] Warwick-Armstrong, S. and Russell, G. "Bilingual Concordancing and Bilingual Lexicography", Euralex, 1990

[Wu94] Wu, D. "Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria". In *32nd Annual Conference of the Association for Computational Linguistics*, Las Cruces, New Mexico, pp80-87, 1994

[WX94] Wu, D. and Xia X. "Learning an English-Chinese Lexicon from a Parallel Corpus". In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, Columbia/Maryland, Columbia, Maryland, 1994. pp.206-213

[WT93] Wu, Z. and Tseng G. "Chinese text segmentation for text retrieval: Achievements and problems". In *Journal of The American Society for Information Science*, 44(9):532--542.

[UIY94] Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y., Nagao, M. "Bilingual text matching using bilingual dictionary and statistics". In *COLING-94*, 15th International Conference, Kyoto, Japan, volume 2,1994

[Zan98] Zanettin, F. "Bilingual comparable corpora and the training of translators". In Laviosa, Sara. (ed.) *META, 43:4, Special Issue. The corpus-based approach: a new paradigm in translation studies*: 616-630.