Advanced Techniques for Chinese Chunk Segmentation and the Similarity Measure of Chinese Sentences

by

Wang Rongbo

A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Electronic and Information Engineering

The Hong Kong Polytechnic University

October 2005

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

________________________________(Signed)

_____Wang Rongbo (王荣波)_____(Name of student)

Abstract

This thesis addresses two important problems in Chinese information processing, namely Chinese chunk segmentation and the similarity measure of Chinese sentences. The three main contributions reported in this thesis are: (1) a novel Chinese chunk segmentation technique using a statistical model combined with correction rules generated using an error-correction mechanism; (2) a novel similarity measure of Chinese sentences using both word/chunk sequences and POS (Part of Speech) tag sequences of Chinese sentences; and (3) the optimization of parameters used in the combined similarity measure approach by applying a relevance feedback technique and a neural network model.

In the first investigation, a statistical model combined with correction rules generated by an error-correction mechanism is proposed for Chinese chunk segmentation. Chunk segmentation of Chinese sentences in the training corpus was carried out manually to provide a ground rule for training the statistical model with which preliminary chunk segmentation results will be obtained. The chunk segmentation result (correctly and incorrectly segmented chunks) from the statistical model is utilized to generate a set of correction rules for refining the segmentation result. This set of correction rules is generated by an error-correction mechanism in which a comparison between the preliminary segmentation result and the manually segmented result is performed. The statistical model and the learned correction rules can then be used to perform

Chinese chunk segmentation of unseen sentences.

In the second investigation, novel similarity measures of Chinese sentences are proposed by using word/chunk sequences and POS tag sequences of Chinese sentences. The sentence similarity measure is one of very important components in example-based machine translation (EBMT). For Chinese sentences there is no delimiter between any two words, which is different from English sentences. Hence, Chinese word/chunk delimitation should be performed first before a sentence similarity measure can be computed. Both word/chunk sequence feature and POS tag sequence feature used in our proposed similarity measures are based on word/chunk segmentation. Sentence structure information is partially reflected in the POS tag sequence. For the proposed word-sequence-matching-based (WSMB) method, we take into consideration three factors between two sentences: the number of identical word sequences, the length of each identical word sequence, and the average weighting (AW) of each identical word sequence. In computing AW, we weight every POS tag according to its importance. The POS-tag-sequence-matching-based (PTSMB) method is to measure the similarity of Chinese sentences in terms of their structures. If the constituents in two Chinese sentences are similar, then we can judge that these two Chinese sentences are similar in structure. The main idea of this similarity measure is that we perform matching between the POS's of two Chinese sentences using directed graphs. The POS weighting is also utilized in the process.

In the third investigation, we propose a human-computer interaction approach to optimize parameters used in the combined similarity measure of Chinese sentences based on a relevance feedback scheme and a neural network model. In the relevance feedback process, users' intentions and preferences to rank the candidate sentences are captured and used to modify parameters in the similarity measure. For the parameter optimization research, a web-based questionnaire was designed to collect users' feedback data. In this pioneering study, we constructed 50 groups of sentences. There is one source sentence and ten sentences to be retrieved for every group. The ten test sentences are shown in descending order of similarity to the source sentence. The user is asked to provide a new rank according to his or her judgment if he/she does not agree with the ranking done by the computer. The new rank is converted to a set of numerals and stored in a database for the parameter optimization using a neural network model. One clear advantage of this approach is its ability to fine-tune the measure to reflect the user's or users' preferences in matching Chinese sentences. Experimental results show a visible improvement of the similarity measure performance.

In addition to the theoretical and experimental studies in Chinese chunk segmentation and the similarity measure of Chinese sentences, we also implemented them into an EBMT prototype in which we also addressed other issues such as data structure, sentence indexing, and user-friendly interface design.

List of Publications

Journal Papers

[1] R. Wang and Z. Chi, “A Similarity Measure of Chinese Sentence Structures (in Chinese)”, Chinese Information Processing, Vol. 19, No.1, pp. 21-29, 2005..

[2] R. Wang, Z. Chi, B. Chang and X. Bai, “Similarity Measure of Chinese Sentences (in Chinese),” Computer Engineering, Vol.31, No.13, pp. 142-144, 2005.

[3] R. Wang, Z. Chi, B. Chang and X. Bai, “An Improved Similarity Measure of Chinese Sentences,” International Journal of Information, Vol. 8, No. 1, pp. 139-145, January 2005, Japan.

[4] R. Wang, Z. Chi, and C. Zhou, “An English-to-Chinese Machine Translation Approach Based on Combining and Mapping Rules (in Chinese),” Computer Engineering and Applications, Vol. 40, No. 30, pp. 97-101 & 135, December 2004.

[5] R. Wang and Z. Chi, “Automatic Chinese Chunk Segmentation Using a Neural Network (in Chinese),” Computer Engineering, Vol. 30, No. 20, pp.133-135, October 2004.

Conference Papers

[1] R. Wang, Z. Chi, and C. Zhou, “An English-to-Chinese Machine Translation Method Based on Combining and Mapping Rules,” Proceedings of Asian Symposium on Natural Language Processing to Overcome Language Barriers, pp.97-102, Sanya, China, March 2004.

[2] R. Wang and Z. Chi, “Automatic Segmentation of Chinese Chunks using a Neural Network,” International Conference on Neural Networks and Signal Processing (ICNNSP2003), Vol. I, pp. 96-99, Nanjing, Jiangsu, China, December 12-15, 2003.

Submitted Journal Papers

[1] R. Wang and Z. Chi, “A Machine Learning Technique for Improving Similarity Measure of Chinese Sentences,” Submitted to International Journal of Computational Linguistics and Chinese Language Processing.

[2] R. Wang and Z. Chi, “Chunk Segmentation of Chinese Sentences Using a Combined Statistical and Rule-based Approach,” Submitted to International Journal of Computer Processing of Oriental Languages.

Acknowledgements

I would like to express my sincere thanks to Dr. Zheru Chi for his great patience and careful, supportive guidance. It is beyond words to express my gratitude to him. He has enriched my academic experience and his friendly personality has influenced me greatly, which will surely benefit me in my future life.

I also appreciate the help from my colleagues in the Center of Multimedia Signal Processing, Department of Electronic and Information Engineering, Hong Kong Polytechnic University. They made my life more enjoyable during my study.

Particularly, I want to give my great thanks to Wang Zhiyong, Hong Anxiang, Li Junli, Wang Xiuying, Song Jiatao, Fu Hong, Chen Sirong, Xie Xudong, Zhao Yi, Chang Baobao, Zhan Weidong, Li Sujian, and Bai Xiaojing. I should also like to give my pure-hearted thanks to so many friends who cannot be listed here. Without their help and encouragement I would not have been able to complete my research.

Finally, I would like to thank my dear parents, brother, and other relatives for their undying support.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AW	Average Weighting
CIP	Chinese Information Processing
CSRA	Combined Statistical and Rule-based Approach
EBMT	Example-based Machine Translation
IAR	Incorrect Adjustment Rate
IR	Information Retrieval
KR	Knowledge Representation
MM	Maximum Matching Method
MT	Machine Translation
MTS	Machine Translation System
NLP	Natural Language Processing
OCR	Optical Character Recognition
OM	Optimal Matching Method
OPT-RF	Optimal Learning Relevance Feedback
POS	Part of Speech

RBMT	Rule-based Machine Translation
RDBMS	Relational Database Management System
RF	Reference Feedback
RIM	Relatively independent meaning
RMM	Reverse Directional Maximum Matching
SDK	Software Development Kit
SM	Statistical Model
SMCS	Similarity Measure of Chinese Sentences
SOM	Self-Organization Map
SVM	Support Vector Machine
TM	Translation Memory

Chapter 1

Introduction

The great advances in science and technology, especially technologies in computers, information, telecommunication and transportation, have encouraged more and more exchanges among people from all around the world. It has also been well recognized that we are in an information explosion era, especially due to the rapid development of Internet technology. Everyone receives an enormous amount of information forwardly or passively. Selection of the information received is therefore becoming a very important issue. People only want to receive information favorable to themselves, including that is helpful for his or her health, work and life. It comes in a variety of formats (including text, graphics, audio and video) and a variety of languages. Naturally, text information processing, natural language processing (NLP), and machine translation have become important research topics.

In natural language processing, sentence analysis is a key problem. Due to its complexity and difficulty, much research is needed to be done in this field. For Chinese information processing (CIP), the word/chunk segmentation and tagging is a fundamental task that has attracted many researchers. On the other hand, the similarity measure of Chinese sentences is also an important research topic in CIP, especially in example-based machine translation (EBMT). Chinese has become a more and more important language in the world because the largest population uses the language and the country has enjoyed an astounding economic growth for the past two decades and will continue to do so for many years to come. CIP is a very challenging research topic, and it has attracted many researchers all over the world to develop techniques to tackle various problems in CIP.

1.1 Motivation

The main difference between Chinese and a phonetic language, e.g., English, is that there is no delimiter between Chinese words in a sentence. Therefore, word segmentation is a fundamental task in Chinese information processing (CIP). Many techniques for Chinese word segmentation have been proposed. However, there is still no mature theory or reliable technique for a deep Chinese sentence analysis. A promising alternative is to conduct a shallow analysis of Chinese sentences instead of a deep analysis, meaning we perform sentence analysis at the phrase or chunk level but not at the word level. In addition to sentence analysis on which a sentence similarity measure relies heavily, the formulation of sentence similarity measure is also a very important task which can be applied in EBMT research (Watanabe, 1992). POS tag sequence information and word/chunk sequence information are two types of important information that can be utilized in sentence similarity measure.

By considering the above two issues, my PhD study had the following three main objectives. The first objective of my study is to investigate chunk segmentation of Chinese sentences. Although a lot of research work has already been carried out in this area for both Chinese and English sentences, there is still no mature technique that can perform chunk segmentation robustly. Hence, there is still a great need for improving chunk segmentation that can significantly contribute to the performance of a machine translation system. The second objective of my study is to explore novel similarity measures based on word/chunk sequence and the POS tag sequence of Chinese sentences. In addition, the similarity measure should be adaptive so that it can accommodate different user group's preferences. Finally, investigating techniques for incorporating users' feedback information to refine the similarity measure of Chinese sentences forms the third objective of my study.

1.2 Statements of Originality

In this thesis two essential issues, chunk segmentation and the similarity measure of Chinese sentences, are explored. The work described in this thesis was carried out at the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, between July 2002 and July 2005, under the supervision of Dr. Zheru Chi.

The thesis consists of six chapters and one appendix. The work described in this thesis was originated by the author except where acknowledged and referenced, or where the results are widely known. The following is the statement of original contributions:

(1) An improved chunk segmentation of Chinese sentences based on a statistical model combined with correction rules generated by using an error-correction learning algorithm is the work of the author. We also proposed an improved definition of Chinese chunks based on relatively independent meaning (RIM) which is different from other existing definitions. Using RIM to segment Chinese chunks has advantages for Chinese information processing such as EBMT.

(2) An improved similarity measure of Chinese sentences based on word/chunk sequence information is the work of the author. In this measure, word segmentation and tagging is carried out first. We take into consideration three factors of two sentences to be compared, the number of identical word sequences, the lengths of identical word sequences, and the average weighting (AW) of identical word sequences. Every Chinese POS tag is weighed according to its importance commonly agreed. An enhanced similarity measure by including chunk segmentation information is also proposed.

(3) An improved similarity measure of Chinese sentences based on tag sequence information is the work of the author. The objective of this

method is to measure the similarity of Chinese sentences' structures. Similar to word-sequence-matching based method, the POS tag weighting is also utilized in this measure. We use a directed graph to model the POS tag sequence of a sentence where the tag of POS is represented using a node and a directed weighted link is used to connect two neighbor nodes.

(4) Parameter optimization of the similarity measure of Chinese sentences using a relevance feedback (RF) scheme and a neural network model is the work of the author. In this scheme, the parameters used in combining the two measures mentioned in (2) and (3) and the weighting factors adopted in each individual measure are optimized according to the preference of a user group. Experimental results show a visible improvement to the performance of sentence similarity measure.

1.3 Outline of the Thesis

The thesis is organized into six chapters and one appendix. They are outlined as follows:

Chapter 2 reviews Chinese information processing (CIP) and important issues addressed in example-based Chinese-to-English machine translation.

In Chapter 3, a statistical model combined with correction rules for Chinese chunk segmentation is presented. We first give an improved definition of Chinese chunks based on relatively independent meaning (RIM). Using RIM to segment Chinese chunks has advantages for CIP. Experimental results are also reported and discussed in the chapter.

In Chapter 4, two similarity measures of Chinese sentences, one based on word/chunk sequence information and the other based on POS tag sequence information, are presented. We also proposed a weighted average of two measures for an improved similarity measure performance.

Chapter 5 presents an optimization technique for the parameters used in the combined similarity measure by making use of a reference feedback scheme and a neural network model. The relevance feedback scheme is to capture the users' preferences in ranking similar Chinese sentences. The neural network model is utilized to optimize the model parameters by using the collected feedback data.

Chapter 6 concludes the research work presented in this thesis and provides some directions for future research work in the field.

As an application of our proposed similarity measure of Chinese sentences, the implementation of a simple EBMT prototype for computer-aided sentence translation from Chinese to English is described in Appendix A. The main implementation issues described include (1) a user-friendly interface, (2) integration of word segmentation and tagging programs into the system to preprocess the input sentence(s), (3) construction of a small corpus and corresponding index files for testing the sentence-level translation system; and (4) implementation of our proposed similarity measure of Chinese sentences.

Chapter 2

Chinese Information Processing

2.1 Introduction

Natural languages are the overwhelmingly preferred medium for human information exchange. The term “natural languages” refers to the languages that people use daily, like English and Chinese, as opposed to artificial languages like programming languages used for computers. A computer normally stores and processes information in ways not closely related to human languages. Natural language processing (NLP) has been a popular research topic for many years (Jones, 1996; Jones & Somers, 1997; Manning & Schütze, 1999; Nirenburg, et al., 1992). The utmost goal of NLP is to design and develop software that can analyze, understand, and generate language sentences that humans use naturally. Actually, this is a very challenging goal to achieve. That a computer can understand a sentence means it knows what concept(s) a word or phrase represents and knows how to connect those concepts together in a meaningful way. However, it is very ironic that the symbol system of natural language is the easiest for humans to learn and use but is very difficult for a computer to master. Computers still fail to master the basic meaning of our spoken and written languages although they possess very powerful performance in computation. Up to now many applications of NLP have been developed, such as information retrieval, intelligent search engines, language translations, and automatic summarization. As the Internet becomes more popular and more widely used, these applications will also attain greater importance.

NLP research has attracted many researchers from different disciplines, in particular linguistics, psychology and engineering. Many branches have developed within this domain. Machine translation (MT) is one of important branches in NLP which has been researched for about sixty years (Somers, 1997). MT is the study of the machine translation from one natural language to another by using a computer or several computers automatically. Several machine translation systems have been developed, and some have been successfully applied to the translation of manuals of continuously updated products for some international companies. Other MT systems have also developed for individuals to improve their work efficiency. However, existing MT systems are still far from producing satisfactory translation performance. It is well recognized that in spite of great potential in the application of such technology, it is almost impossible to develop an MT system that can match human performance in the foreseeable future. This is because MT is such a complex scientific task involving almost every aspect of natural language processing and multi domain knowledge including linguistics, mathematics, psychology and artificial intelligence. Each small step of progress in MT depends on the development of all the above fields. For example, the development of psychology and artificial intelligence may present a more robust and applicable knowledge representation method for MT.

In addition, the search efficiency of a huge database needs support from outstanding research in applied mathematics and computer science.

Following the development in language technology, corpus-based MT approaches (statistical approach or example-based approach) have partially succeeded in replacing the traditional rule-based approaches. The main advantage of corpus-based MT systems is that they are self-customizing in the sense that they can learn the translation of terminology and even stylistic phrasing from previously translated examples. This process can avoid rule construction which is always a bottleneck in rule-based machine translation (RBMT).

2.2 Chinese Information Processing

Chinese information processing (CIP) is a very active research field in which the processing of Chinese information including text and speech by computers has been researched. The main research and development topics in CIP include, but are not limited to, word segmentation and tagging, analysis of Chinese syntax and semantics, Chinese input technology and systems, development and application of Chinese corpus, speech input and synthesizers, Chinese character recognition, Chinese sentence understanding and generation, human machine interfaces, machine translation, Chinese information retrieval, automatic text categorization, electronic dictionaries, and automatic information filtering, computer-aided editing, mutually transferrance between speech and text, etc. (Xu, 2001). A fundamental task for CIP is word/chunk segmentation and tagging.

2.2.1 Word Segmentation and Tagging

Word boundary detection is required in spoken languages and some written languages such as Chinese where no delimiter is used in a sentence. In processing such a language, word segmentation and tagging plays an important role (Furuse & Iida, 1994). Other tasks in Chinese information processing, for example the similarity measure of Chinese sentences, Chinese sentence analysis, and automatic Chinese text classification, etc., are certainly based on segmentation and tagging. Several methods have been proposed for Chinese word segmentation and tagging, for example, the maximum matching method (MM), reverse directional maximum matching (RMM), optimum matching (OM), the neural network approach, and the word-frequency-based method.

2.2.2 Chunk Segmentation

Chinese language understanding is a long-term objective of Chinese information processing. Sentence analysis is a fundamental problem in approaching this objective. As a result, there have been a number of research topics actively pursued in Chinese information processing for a better analysis of Chinese sentences, for example word segmentation and tagging, phrase segmentation and classification, word sense disambiguation, and anaphora resolution, etc. Some of these research topics were proposed for deep sentence parsing which is still far from a mature technology. A promising alternative is to perform shallow analysis of Chinese sentences, which is helpful for many NLP tasks, including information extraction, text summarization, and spoken language understanding. Shallow analysis can also play an important role in example-based machine translation (EBMT) for the following two reasons:

(1) Without sentence analysis, a sentence can hardly be found to match well with the input sentence even if a large sentence database is available.

(2) Performing a deep parsing on sentences violates the essential idea of EBMT which was proposed for avoiding the drawbacks of a rule-based method based on deep sentence analysis.

Shallow parsing, also called partial parsing or chunk parsing, has become a promising alternative to deep parsing (Abney, 1991; James, et al., 2002; Li, et al., 2002b; Liu, et al., 2000; Megyesi, 2002; Molina & Pla, 2002; Osborne, 2002; Ramshaw & Marcus, 1995; Sun & Jurafsky, 2004; Sun & Yu, 2000; Zhang, et al., 2002). The main goal of a shallow parser is to segment a sentence into non-overlapping segments of certain syntactic units without touching deep structures of a sentence (Ramshaw & Marcus, 1995). The main technical problems included in shallow parsing are chunk segmentation and the relationship analysis between identified chunks in a sentence.

A few methods have been proposed for shallow parsing of sentences for processing different languages. These methods can be divided into two main categories: the statistical method and the rule-based method. Either method might not produce satisfactory performance individually, but a combined approach could be more attractive.

2.3 Machine Translation

As soon as the computer was invented, people started to hope that the computer could assist humans in performing some translation tasks by a way of so called machine translation (MT). At present, MT is still one of research focuses in the science and technology domains.

Machine translation is about the translation from one natural language to another by a computer or computers with or without human assistance. It refers mainly to the text translation but not dialogue translation, although spoken language translation is also a popular research topic now (Kitano, 1994; Oi, et al., 1994; Sobashima, et al., 1994). MT is different from computer-based translation tools which only support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, etc. The most important point of MT itself is the automation of the full translation process (Hutchins, 1995; Nomiyama, 1992). MT is a multi-disciplinary, fast growing research domain, which makes use of almost all computational methods known in artificial intelligence. Its development depends greatly on the progresses in linguistics, mathematics, psychology, and artificial intelligence, etc.

It is well recognized that MT presents very challenging system engineering problems and there is still a long way to go to build a satisfactory MT system. Existing MT systems cannot produce satisfactory translations.

In different periods of MT research, researchers emphasized different methodologies along with the improvement of understanding on MT (Nomiyama, 1992). The dominant framework of MT research until the end of the 1980's was based on essentially linguistic rules of various kinds: rules for syntactic analysis and lexical analysis, rules for lexical transfer, rules for syntactic generation, rules for morphology, etc. However, it is very difficult to retrieve the knowledge of natural languages and to formalize them in rules. Since the end of the 1980's, the dominance of the rule-based approach has been broken by "corpus-based" or "example-based" methods (Hutchins, 1995; Juola, 1994; Nirenburg, et al., 1994; Sumita, et al., 1990). In 1988, a research group from IBM published the results of experiments on a system based purely on statistical methods. The effectiveness of the method was a considerable surprise to many researchers and has inspired others to explore various kinds of statistical methods in subsequent years (Brown, et al., 1990; Brown, et al., 1992; Brown, et al., 1993; Brown, et al., 1988; Manning & Schütze, 1999). At the almost same time, Prof. Nagao led a Japanese group to publish preliminary results on machine translation based on corpora of translation examples, i.e., using the approach now generally called "example-based" translation (Nagao, 1984). For both statistical approaches, the principal feature is that no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents. To avoid the thorough syntactic and semantic analysis of sentences, a super-function based machine translation was proposed in which the super-function showed the correspondence between original language sentence patterns and target language sentence patterns (Ren, 1999).

2.3.1 Example-Based Machine Translation (EBMT)

Example-based machine translation is a reasonably well-established paradigm for machine translation which emerged in the end of the 1980's as an alternative to rule-based MT systems. EBMT retrieves similar examples (pairs of source phrases, sentences, or texts, and their translations) from a database of examples, adapting the examples to translate a new input sentence (Jones, 1992; Sumita & Iida, 1991). The basis of EBMT is the existence of a large number of translated parallel bilingual texts, including pairs of bilingual phrases and sentences. The retrieval process is done by measuring the distance of the input sentence to each of examples in the database. The smaller a distance is, the more similar the example is to the input text. However, developing a good distance metric of similarity is a key problem of EBMT. From similar examples and their translations, the best possible translation is generated by transfer and replacement operations to the translation of the examples according to the differences between the input and the examples. The EBMT method can overcome difficulties introduced in constructing dictionaries and rules in a rule-based machine translation (RBMT) method in which it is very challenging to attain the language knowledge of the text and formalize them in rules. It is also difficult to improve translation performance because the effect of adding a new rule is difficult to anticipate, and because translation using a large-scale rule-based system is very time-consuming.

EBMT has no rule, and the use of examples is relatively localized. Simply inputting more appropriate examples into the database can improve the system's performance significantly. Different from RBMT, EBMT is easily upgraded and the more examples the system has, the better the performance of an EBMT system can achieve.

One clear advantage of EBMT is that deep semantic analysis can be avoided because it is assumed that translations appropriate for given domain can be obtained using domain-specific examples (pairs of source and target expressions). An EBMT system directly returns the translation without using of rules. Moreover, in EBMT, the reliability factor is assured by the translation result based on the differences between the input text and the similar examples found. In addition to this, retrieved examples that are similar to the input text convince users that the translation is relatively accurate compare to the results of using rule-based method (Sumita & Iida, 1991). However, the searching and matching efficiency of EBMT with a huge number of examples is an important issue to be addressed. The problem can be partially solved by using indexing or/and parallel computing techniques.

In summary, EBMT can be characterized as follows:

(1) It is easily upgraded by simply inputting appropriate examples to the database;

(2) It can assign a reliability factor to the translation result;

(3) It is accelerated effectively by both indexing and parallel computing techniques;

(4) It is robust because of best-match scheme. Based on the sentence similarity measure techniques, the top most similar sentences to the input one can be retrieved which can be considered as the best match scheme.

(5) It well utilizes translators' expertise. The translators' expertise is stored in database in sentence sample style which can be used to generate the new translation of the inputs.

There are three key issues that need to be addressed in EBMT:

(1) Establishment of correspondence between units in a bi/multi-lingual text at the sentence, phrase and word levels;

(2) A mechanism for retrieving from the database the unit that best matches the input;

(3) Exploit the retrieved translation example to produce the actual translation of the input sentence.

One of the most important problems in EBMT is how to measure similarity between a sentence or its fragments and a set of stored examples. Generally, an EBMT system consists of two databases: an aligned bilingual examples database and a thesaurus; and three translation modules: examples retrieval, example-based transfer, and new translation generation. The thesaurus can be used in calculating the semantic distance between the content words in the input and those in the examples.

2.3.2 Similarity Measures in EBMT

Some main problems underlying EBMT will be reviewed further in this section. They include (1) bilingual sentence database construction; (2) methods for matching new inputs with the examples in the database; and (3) what to do with the examples once they have been identified. For points (1) and (3), we will only give a simple introduction because they are not closely related to the core discussion of this thesis. For point (2), we will review it in detail by investigating the problems behind it and existing approaches which lead to our proposed research, that is, Chinese chunk segmentation and similarity measures for Chinese sentences.

(1) Bilingual Sentences Corpora

Since EBMT is a corpus-based method, the first thing to do is to construct the bilingual aligned sentences corpora (Gale & Church, 1993; Grishman, 1994; Meyers, et al., 1996; Somers, 1998). At the same time, corpus alignment at the sentence level, phrase level or word level should be preformed in bilingual corpus construction.

EBMT systems are often felt to be best suited to a sublanguage approach, and an existing corpus of translations can often serve to define implicitly the sublanguage which the system can handle. Researchers may build up their own parallel corpus or may locate such corpora in the public domain.

The first step is to collect example sentences from various sources, especially from the Internet. The more examples covering different language styles, the better performance an EBMT system can achieve. Once a large number of bilingual sentences are collected, there remains a problem of aligning them. The alignment problem can of course be circumvented by building the example database manually, as is sometimes done for Translation Machines (TMs), when sentences and their translations are added to the memory.

The obvious and intuitive “grain-size” for examples seems to be the sentence, though evidence from translation studies suggests that human translators work with smaller units such as the chunk or phrase. Furthermore, the sentence as a unit appears to offer some obvious practical advantages – sentence boundaries are for the most part easy to determine, and in experimental systems and in certain domains, sentences are simple. However, in the real world, the sentence provides a grain-size which is too large for practical purposes, and the matching and recombination process needs to be able to extract smaller units from the examples and yet still work with them in an appropriate manner. This in turn suggests a need for parallel text alignment at a sub-sentence level, or that examples are represented in a structured fashion.

(2) Similarity Measure

Similarity measure is an important problem in many engineering domains, for example, image retrieval (Belongie, et al., 2002; El-Naqa, et al., 2000; Gudivada & Raghavan, 1995; Guo, et al., 1998; Lim, et al., 2001; Patrice & Konik, 2000; Stricker & Orengo, 1995). The most important task in an EBMT system is to find the best matched example (or a set of examples) of the source-language input sentence (Cranias, et al., 1994; Lim, et al., 2001; Maruyama & Watanabe, 1992; Matsumoto, et al., 1993; McLean, 1992; Nirenburg, et al., 1993). The basis of the EBMT approach is having a large-scale bilingual corpus. Therefore, how to find the most similar sentence or sentences to the input from the corpus affects the quality and processing speed of a machine translation system (Che, et al., 2003; Chen, et al., 2001; Gowda & Diday, 1992; Li, 2002; Li, et al., 2003(a); Mandreoli, et al., 2002; McEnery & Wilson, 1996; Sui & Yu, 1998; Wang & Chi, 2005; Wang, et al., 2005; Zhang & Shasha, 1997). This search problem depends of course on the way the examples are stored. In more conventional EBMT systems, the matching process may be more or less linguistically motivated

(Somers, 1999). The search performance is affected by the similarity measure adopted. There are several approaches proposed to compute sentence similarity. However, there is still much room to improve the performance, due to diverse and complex nature of natural languages.

Feature selection and weighting is a very important problem in this process as addressed in other domains (El-Naqa, et al., 2000; Lu, et al., 2000; Wu, et al., 2004). Different similarity measures may be used for different feature representations of Chinese sentences.

A number of similarity measures have been proposed, such as mutual information (Chen, et al., 2003), Dice coefficient (Lin, 1998), cosine coefficient (Shyu, et al., 2004), distance-based measurements (Shyu, et al., 2004), and feature contrast model (Eidenberger & Breiteneder, 2003). McGill et al surveyed and compared 67 similarity measures used in information retrieval (Lin, 1998). For sentence similarity measure, the matching methods can be categorized into two main groups according to the degree of analysis made to sentences. They are string-matching-based and syntactic/semantic-based. For string-matching-based methods, a sentence is viewed as a sequence of words and no grammar or structure analysis is performed. The only information used is the surface layer information of sentences, for example, word sequence, part of speech (POS) information. However, structure information is a kind of important content contained in sentences. It should be utilized in a sentence similarity measure for an improved performance.

(3) Adaptability and Recombination

Having retrieved a set of examples, with accompanied translations, the next step is to extract from the translations the appropriate fragments and to combine these so as to produce a grammatical target output. This process is referred to recombination or generation. This is arguably the most difficult step in the EBMT process which has not been investigated widely. The problem is twofold: (a) identifying which portions of the associated translations correspond to the matched portions of the source text, and (b) recombining these portions in an appropriate manner. Compared with other issues in EBMT, adaptability and recombination have received considerably less attention.

2.4 Conclusion

In this chapter, we briefly review the main issues and challenges of research in natural language processing, Chinese information processing, and example-based machine translation (EBMT). We particularly point out the necessity of conducting research in Chinese chunk segmentation and the similarity measure of Chinese sentences. We give a general review on the similarity measure of Chinese sentences, together with the other two main tasks in EBMT, bilingual sentences corpora construction and adaptability and recombination. Detailed reviews on Chinese chunk segmentation and the similarity measure of Chinese sentences are given in Chapters 3 and 4, respectively.

Chapter 3

Chunk Segmentation of Chinese Sentences Using a Combined Statistical and Rule-based Approach (CSRA)

3.1 Introduction

Computerized Chinese language understanding is a long-term objective of Chinese information processing. A fundamental problem to approach this objective is to perform sentence analysis. Unfortunately, the current techniques on deep parsing of sentences are still far from a mature technology. A promising alternative approach is to perform shallow analysis on sentences, which is helpful for many natural language processing (NLP) tasks. The main goal of a shallow parser is to segment a sentence into non-overlapping segments of certain syntactic units without producing too detailed information such as that from a deep parser.

Over the last few years, several methods have been proposed and applied to shallow parsing of sentences in different languages. These methods can be divided into two main categories: statistical methods and rule-based methods. Either method might not produce satisfactory performance individually, but a combined approach could be more attractive. Generally, a statistical method plays an important role in shallow parsing of languages. Liu et al proposed a statistical algorithm to recognize definite levels of Chinese chunks (Liu, et al., 2000). In the measure, Level-Tag is applied which is too complex to be practical. In (Xi & Sun, 2002), Xi and Sun proposed a method for automatic determination of Chinese phrase boundaries using a neural network. The method can obtain a high precision. However, it is very difficult to determine the number of nodes in every layer and the times of training. In (Li, et al., 2004(a)), Li et al proposed a support vector machine (SVM) based method for Chinese text chunking. For SVM-based method, it is normally used to the classification of bi-value problems. So many classifiers should be constructed for chunk identification which is too complex and time-consuming. In (Li, et al., 2002(b)), Li et al presented a combined rule-based and statistics-based method for Chinese chunk parsing. In their definition of chunks, they do not emphasize whether a chunk has a relatively independent meaning (RIM) while a chunk with an RIM is more suitable to be a translation unit. In this method, they used the learned rules to identify some special Chinese chunks. Li et al also proposed a method based on the maximum entropy principle to perform shallow parsing on Chinese sentences (Li, et al., 2003(b)). In this method, it is difficult to select the feature set. Furthermore, too many feature functions are used. In (Zhou, et al., 1999), Zhou et al proposed a shallow parsing scheme for analyzing Chinese sentences. Zhou also presented a statistical method to recognize Chinese chunks (Zhou, 1996). Li et al proposed a transductive HMM model for Chinese chunk segmentation (Li, et al., 2004(b)). In the scheme, the word boundary stem and constituent group were defined. The usefulness and efficiency are shown by experiments on the automatic segmentation and acquisition of Chinese grammar knowledge. A new E-chunk based multi-engine machine translation model is proposed by Li (Li, et al., 2002(a)). There are also other methods proposed for text chunking (Collins, 1996; Daelemans, et al., 1999; Halteren, 2000; Kudoh & Matsumoto, 2000; Kudoh & Matsumoto, 2001; Skut & Brants, 1998; Tjong Kim Sang, 2000; Tjong Kim Sang, 2002; Voutilainen, 1993; Zhou, et al., 2000). The error-driven learning algorithm was normally adopted in the domain (Wong, et al., 2001).

In this chapter, we propose a chunk definition of Chinese and a combined statistical and rule-based approach (CSRA) to segment chunks of Chinese sentences. In our scheme, the essence of chunk definition is the relatively independent meaning (RIM) which is different from the definitions given by others. For example, the definition given in (Li, et al., 2002(b)) was very similar to the definition of English chunks. The RIM requirement is also different from a phrase. For example, the sentence "我/r 很/d 不/d 喜欢/v 看/v 电视/n" (I do not like to watch TV at all.) can be segmented into three chunks with RIM, "我" (I), "很不喜欢看" (do not like to watch at all) and "电视" (TV). Different segmentation results would be produced if different definitions are adopted. We also propose an error-driven learning algorithm for obtaining the rules to correct wrongly segmented chunks. There are two main elementary sub-tasks in the chunk parsing of Chinese sentences, the segmentation and tagging of chunks. In our study, we focus on the first task, that is, the chunk segmentation. Our proposed CSRA includes two phases: the training and test phases.

The main steps for the training phase of the CSRA are:

(1) Conduct manual chunk segmentation of Chinese sentences in the training corpus.

(2) Obtain the statistical information from the training corpus for the statistical model.

(3) Use the statistical model to segment chunks of Chinese sentences in the training corpus.

(4) The wrongly segmented instances by the statistical model are used to build a decision tree using an error-driven rule learning mechanism.

(5) The manually segmented instances from the training corpus are used to set the parameters in the decision tree.

The block diagram of the training phase of the CSRA is shown in Fig. 3.1. Figure 3.2 shows the steps of using the CSRA to segment a Chinese sentence.

The remainder of this chapter is organized as follows. Section 3.2 describes our definition of chunks with RIM in Chinese sentences. Section 3.3 discusses a statistical model for pre-segmentation of Chinese sentences into chunks. This is followed by Section 3.4 on decision rule generation for dealing with exceptions. Experimental results are reported and discussed in Section 3.5. Finally, a conclusion is drawn in Section 3.6.

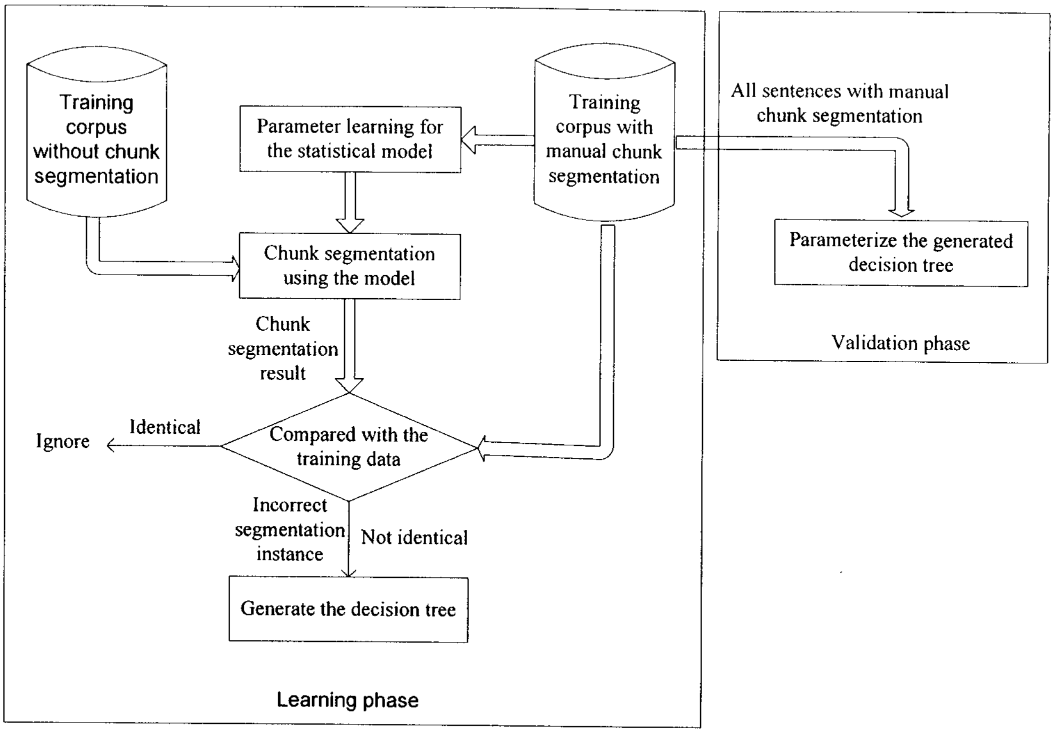

Training corpus without chunk segmentation
Parameter learning for the statistical model
Training corpus with manual chunk segmentation
All sentences with manual chunk segmentation
Parameterize the generated decision tree
Chunk segmentation using the model
Validation phase
Chunk segmentation result
Ignore
Identical
Compared with the training data
Incorrect segmentation instance
Not identical
Generate the decision tree
Learning phase

Figure 3.1: The training phase of the CSRA.

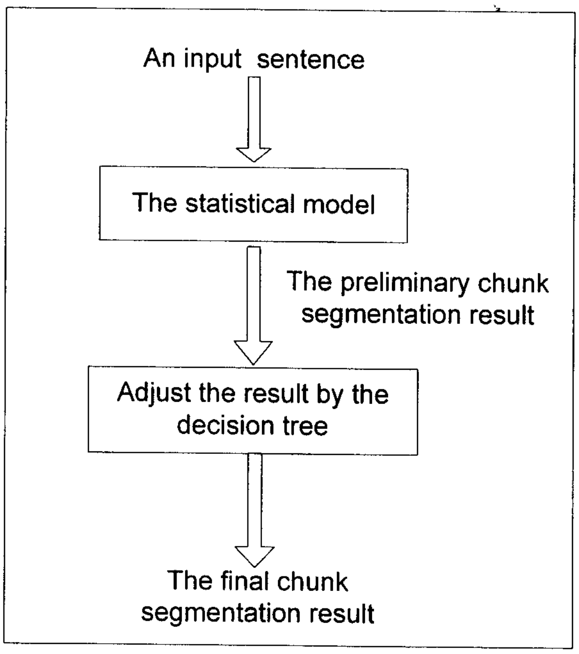

An input sentence
The statistical model
The preliminary chunk segmentation result
Adjust the result by the decision tree
The final chunk segmentation result

Figure 3.2: The test phase of the CSRA.

3.2 Chunks in Chinese Sentences

Shallow parsing is a popular research topic in natural language processing (NLP). Shallow parsing makes an important contribution to EBMT to avoid deep sentence analysis that cannot be satisfactorily carried out otherwise. However, different chunk definitions and segmentation methods will affect the subsequent processes in EBMT, especially for the similarity measure of Chinese sentences. Similar to Chinese parts of speech (POS's), there has been no unique definition on Chinese chunks due to the inherent complexity of the Chinese language. Different definitions of Chinese chunks have been proposed for different applications. In general, a chunk is a language unit that is more complex than a word and simpler than a sentence or a short sentence.

For English, Abney (Abney, 1991) presented an integrated chunk description scheme which is considered as the most authoritative definition. In the scheme, a *chunk* is the non-recursive core of an intra-clausal constituent, extending from the beginning of the preceding constituent to its head word, but not including post-head dependents. The author was interested only in the category and start/end points of a chunk. The author also gave a list of categories, including noun chunk (NX), verb chunk (VX), infinitive chunk (INF), gerund chunk (VGX), past participle chunk (VNX), adjective chunk (AX), and adverb chunk (RX). In addition, some special issues were considered in the chunk scheme, e.g., Wh-Phrases, punctuation and coordination.

Many researchers have worked on the shallow parsing of Chinese sentences. The E-chunk is an extended chunk proposed by Li et al (Li, et al., 2002(a)). In (Li, et al., 2003(b)), Li et al defined a Chinese chunk as a non-recursive structure which accords with a certain syntactical function. Every chunk has a head constituent which is the core of other constituents within the chunk. Some categories of Chinese chunks are also given syntactically but not semantically, functionally or lexically. In (Li, et al., 2004(b)), Li et al defined a Chinese chunk in the same way as in English. In (Liu, et al., 2000), Liu et al defined a Chinese chunk as a structure that contains a one- or two-level phrase according to a certain syntactical function and semantics.

For improving rationality and suitability of EBMT, in our definition of Chinese chunks we make some changes to the existing definitions and emphasize the following points:

(1) Relatively independent meaning (RIM). This means that a chunk has an RIM expressed by a core constituent and its adjunctive constituents. There is a close relationship between the core constituent and other adjunctive constituents. The adjunctive constituents can be at the beginning or the end of a chunk, which is different from Abney’s definition in which there is no post-head constituent. For example, the sentence “我/r 很/d 不/d 喜欢/v 看/v 电视/n” (I do not like to watch TV at all.) can be segmented into three chunks with RIM, “我” (I), “很不喜欢看” (do not like to watch at all) and “电视” (TV). The core constituents of the chunks are “我”(I), “喜欢看”(like to watch) and “电视”(TV). There is a close relationship between the core constituent “喜欢看”(like to watch) and its adjunctive constituents “很不”(do not…at all). In general, this relationship can be reflected by a high frequency at which the core constituent is followed or preceded by other

constituents.

(2) Non-nesting. This means that under no circumstance will two chunks contain each other. Non-nesting will ensure that all chunks in a sentence will not overlap with each other. For example, if a preceding adjective or several adjectives modify a noun, we do not segment the adjective(s) as a separate adjective chunk but consider the sequence including the adjective(s) and the noun as a single noun chunk. This is similar to the case of preposition chunk segmentation in which we do not consider the noun, e.g., the “桌子” in “在/p 桌子/n 上/f” (Refer to the Table 3.1 for POS tags) as an independent noun chunk but instead the whole sequence including both the preposition and the noun as a preposition chunk. Table 3.2 summaries the categories of Chinese chunks in our definition. In defining Chinese chunks, we take into consideration translating Chinese sentences to English using the EMBT approach in which chunks are more reliable units for measuring sentence similarity and for translation.

Table 3.1: Tags of Parts of Speech (POS’s)

Tag	Meaning	Tag	Meaning
V	Verb	T	Time
A	Adjective	F	Orientation
D	Adverb	R	Pronoun
an	Nominalized Adjective	U	Auxiliary
vn	Nominalized noun	M	Numeral
Q	Quantity	W	Punctuation

Table 3.2: Categories of Chinese chunks in our definition

Types of Chunks	*Description*	*Types of Chunks*	*Description*
NC	Noun chunk	NQC	Numeral quantity chunk
VC	Verb chunk	LC	Location chunk
PC	Preposition chunk	TC	Time chunk
ADJC	Adjective chunk	ADVC	Adverb chunk
NOTC	Not chunk	CC	Conjunction chunk
OC	Punctuation chunk	VPC	Verb Preposition Chunk

The following provides a more detailed account of these defined chunks. For each category, we only summarize the main cases since it is impossible to cover all the possible cases due to the complexity of Chinese language.

NC (Noun Chunk)

A NC extends a head noun from its preceding and sequential constituents that modify the head noun. Typical examples are described below. We firstly describe the basic noun chunk. then we describe possible preceding and sequential constituents. The basic noun chunks include:

- Nouns. There are about 3570 commonly used nouns most of which can act as a noun chunk individually.
- Pronouns. There are about one hundred pronouns which can be used as noun chunks, e.g., [我/r] (I, me), [你/r] (you), [这/r] (this).

- Proper nouns, including names of persons, and proper names of locations and organizations. For example, [香港特别行政区](Hong Kong Special Administrative Region), [长江三角洲/n] (Yangtze River Delta).

- All nominalized POS's. These include verbs and adjectives which act as nouns in a sentence, e.g., "工作" (working) in [一年/t 辛苦/v 的/u 工作/vn] (year-long hard working), [读书/vn] (reading).

The modifying constituents that precede a noun include:

- Possessive pronouns, e.g., [我/r 的/u] (my), [你/r 的/u] (your).

- Adjectives, e.g., [新/a] (new), [大/a] (big).

- Verb phrases. A verb phrase can include a verb followed by a noun and an auxiliary word, e.g., [操作/v 电视机/n 的/u 困难/an], (the difficulty in operating a TV set).

- Numeral and quantity phrases. A numeral and quantity phrase consists of a numeral word tagged as "m" and a quantity word tagged as "q", e.g., [第一/m 颗/q 人造卫星/n] (the first man-made satellite).

- Noun phrases. A noun phrase consists of a noun or several nouns.

VC (Verb Chunk)

Since a Chinese sentence normally consists of a verb chunk(s) and a noun chunk(s), VC is also an important chunk type. The basic verb chunks include:

- Verbs. For example, [获得/v] (obtain), [发射/v] (launch).

- The other POS's which act as a verb. Similar to nominalized POS's, some POS's can act the same role as a verb from the semantic point of view.

The constituents to modify a basic verb chunk include:

- Verbs. That means a core verb is modified by another verb to form a verb chunk, e.g., [开始/v 工作/v] (begin to work).
- Adverbs. Normally, an adverb is used to modify a verb, e.g., [多么/d 盼望/v] (desire very much).

PC (Preposition Chunk)

Actually, a PC chunk begins from a preposition. PC also plays an important role in Chinese sentences. We observed that most of preposition chunks contain a preposition except for the case that a preposition is omitted. Actually, a preposition corresponds to a POS which denotes a location or orientation, e.g., [在/p 公园/n 里/f] (in the park). Most PCs begin with a preposition and end with a noun or a POS denoting a location or orientation.

ADJC (Adjective Chunk)

The basic adjective chunk is an adjective. The modifying constituents preceding or following an adjective include:

- Adverbs. An adjective usually is modified by an adverb, e.g., [非常/d 多/a] (a good many).

- An adjective together with an auxiliary word, e.g., [累/a 得/u 满头大汗/a] (perspiring from exhausting).

Note that if an adjective is followed by a noun or several nouns, then the adjective will not be considered as an adjective chunk while they are combined with the noun(s) to form a noun chunk. Only if the adjective is not followed by any noun or a nominalized POS, it is segmented as an adjective chunk.

NOTC (Not Chunk)

We consider a sequence of words as NOTC which does not belong to any other chunk categories.

OC (Punctuation Chunk)

We consider punctuation in a sentence as OC which is a special constitution in a sentence. That means all punctuations in a sentence are recognized as *OC*s.

NQC (Numeral and Quantifier Chunk)

We consider a sequence of numeral and quantifier as NQC which does not modify any sequential constituents. The quantifier is omitted sometimes.

LC (Location Chunk)

Actually, LC denotes a location which does not modify any sequential constituents. For example, [走廊/n 上/f](on the hallway), [山坡/n 下/f](under the hillside).

TC (Time Chunk)

The basic TC is a POS denoting time, including date, week, etc. For example, [明天/t] (tomorrow).

The constituents which can modify a basic time chunk include:

- Time. It denotes a POS which represents a time concept, e.g., [明天/t 早上/t](tomorrow morning)
- Verb phrases, e.g., [上课/v 的/u 时候/t] (when having a class).
- Numeral and quantity which represents a time concept, e.g., [一/m 天/q 早上/t] (one morning).

ADVC (Adverb Chunk)

Actually, an ADVC is an adverb which does not modify any other constituents. For example, [原来/d](originally), [忽然/d](suddenly).

CC (Conjunction Chunk)

Actually, the POS's which can be classified into a CC include:

- POS's conjunction tagged as "cs" which joint two sub-sentences into a sentence, e.g., [另一方面/cs] (on the other hand). Note that a conjunction tagged as "cw" joints mainly two nouns or two verbs as an integrated part.
- Adverbs. Some adverbs in sentences can also be classified into a CC, e.g., [只要/d] (if only). Actually, in some cases, two adverbs classified as CC can be used in pair to joint two sub-sentences. For example, "只要/d 到/v 了/u 冬天/t ，/w 就/d 会/v 下雪/v 。/w " (It will snow if only winter comes).

VPC (Verb Preposition Chunk)

This is a special chunk which has not been defined elsewhere. VPC begins with a verb which is followed by a preposition. Many Chinese sentences have VPC chunks. That is, a preceding verb has a very compact modifying relationship with the sequential preposition phrase. It is not justified to categorise them into VC or PC chunks. So, we define a new chunk type VPC here. For example, [坐/v 在/p 桌子/n 旁边/f] (sit at the table), [停/v 在/p 河边/s] (park close to the river), [混/v 在/p 垃圾堆/n 里/f](mixed in dustheap).

3.3 A Statistical Model for Chunk Segmentation

3.3.1 Statistical Model

Suppose that $S = < W, T >$ is an input Chinese sentence after word segmentation and tagging. $W = w_1, w_2, ..., w_n$ is the word sequence of the sentence. $T = t_1, t_2, ..., t_n$ is the POS tag sequence corresponding to the word sequence. When we perform chunk segmentation, we should first obtain the tag sequences corresponding to the processed sentence. The chunk segmentation problem can be defined as finding a segmentation point sequence C' from the tag sequence which makes:

$$C' = \arg\max_{C' \in \{C\}} \prod_{i=1}^{N_{seg}-1} P(t_{c_{i-1}+1} ... t_{c_i}) \qquad (3.1)$$

where N_{seg} is the number of the segmentation points in a sentence (the number of chunks plus one). $C = c_0, c_1, c_2, ..., c_{N_{seg}-1}$ denotes the segmentation boundaries of a sentence, c_0 is the beginning of a sentence, which means there is always a boundary at the beginning of a sentence ($C_0 = 0$). Similarly, $c_{N_{seg}-1}$ is the end of the sentence which means there is always a boundary at the end of a sentence ($c_{N_{seg}-1} = n$ where n is the number of words in the sentence). $t_{c_{i-1}+1}...t_{c_i}$ corresponding to $w_{c_{i-1}+1}...w_{c_i}$ is the *i*-th segmented chunk. We consider the POS tag sequence of a sentence and determine the most probable chunk segmentation using Eq. 3.1. For example, if we have a sentence $w_1w_2w_3w_4w_5w_6w_7w_8w_9$ $(n = 9)$, that is, the sentence has 9 words, the corresponding POS tag sequence is $t_1t_2t_3t_4t_5t_6t_7t_8t_9$. One of possible segmentation results is $|_{c_0}\ t_1t_2t_3\ |_{c_1}\ t_4t_5t_6t_7\ |_{c_2}\ t_8t_9\ |_{c_3}$. In this case, $N_{seg} = 4$ and the number of segmented chunks is 3, being $t_1t_2t_3$, $t_4t_5t_6t_7$ and t_8t_9. Therefore, the joint probability of such segmentation is $p(t_1t_2t_3) \times p(t_4t_5t_6t_7) \times p(t_8t_9)$ each term of which can be obtained using the training database (so-called statistical information). Using Eq. (3.1) for Chinese chunk segmentation is a statistical approach. We can know that the number of possible chunk segmentation point sequences of a tag sequence with length i is 2^{i-1}. For the above example, it is $2^{9-1} = 256$.

3.3.2 Statistical Information Extraction and Parameter Estimation

From the manually processed corpus, we obtained the statistical information for chunk segmentation. For a tag sequence, the frequency of containing chunk segmentation points or not was obtained. For example, if we have the following tag sequence of POS's corresponding to the Chinese sentence "尽管/cs 他/r 力气/n 小/a ，/w"(even if he is weak,):

| *cs* | *r n* | *a* | *w* |

All possible sub-sequences of the above tag sequence are summarized in Table 3.3. Note that the punctuation tag *w* is ignored because it is always considered as a chunk.

Table 3.3: Statistical information extraction of a sample Chinese sentence

Sub tag sequence	Count of the tag sequence without any segmentation point in it	Count of the tag sequence with at least one segmentation point in it
cs r	0	1
cs r n	0	1
cs r n a	0	1
r n	1	0
r n a	0	1
n a	0	1

The punctuation following a sub-sentence or a sentence is always considered as a chunk boundary, meaning that any sub-sequence will not contain a punctuation symbol. The partial statistical information in the database is given in Table 3.4.

For convenience of computation, we list all three statistical counts although any two of them are sufficient to compute all the frequencies.

Table 3.4: Partial statistical information in the database (TagSeq: tag sequence; CntSeg: the count of the tag sequence with at least one segmentation point in it; CntNoSeg: the count of the tag sequence without any segmentation point in it; CntTotal: the total number of the tag sequence in the sentence corpus. The definition of tag symbols is given in Table 3.1.)

TagSeq	CntSeg	CntNoSeg	CntTotal
d/v	8	543	551
v/v	13	402	415
m/q	1	187	188
p/n	1	168	169
n/f	1	133	134
r/v	168	120	288
d/v/v	5	119	124
v/r	80	92	172
r/n	3	91	94
n/u/n	0	78	78
a/u/n	1	74	75
v/u/n	38	69	107

3.4 Error-Correction Based Learning of Decision Rules

3.4.1 Main Idea

A statistical model cannot accommodate all the cases in Chinese chunk segmentation. There are always some exceptions. On the other hand, using decision rules only cannot solve the problem either. Therefore combining these two methods would be an attractive approach, making use of the advantages of both methods to produce better performance. The main steps in the training and test phases of the CSRA are illustrated in Fig. 3.1 and Fig. 3.2, respectively.

3.4.2 Decision Tree Generation

The chunk segmentation using our proposed statistical model is followed by segmentation refinement using a decision tree for improved performance. It is necessary to consider the context of the segmentation boundary of a wrong segmentation from the statistical model. The decision rules generated are organized into a decision tree. The syntax of a decision rule is:

<Wrong Segmentation> :: <Correct Segmentation>.

For example,

<d v | v n> :: <d v v | n>

For a tag sequence of POS’s “d v v n”, it is a wrong segmentation to break it into two segments “d v” and “v n”. It should be adjusted to become two segments “d v v” and “n”. Other examples are

(1) <n v y> :: <n | v y>

(2) <v v | m q> :: <v v m q>

There are mainly three cases of wrong segmentation from the statistical model: 1-chunk, 2-chunk and multi-chunk. Correspondingly there are three ways to process them, which will be discussed later. Three types of nodes are used in the decision tree: root node, internal node and leaf node. The root node attached with the total number of learned rules is the ancestor of all other nodes. An internal node contains properties “cntC”, “cntE”, “rate” and “TagPair”. The “cntC” and “cntE” are the numbers of correctly and wrongly segmented tag sequences going through the node, respectively. The “rate” is the error rate which is defined as the ratio of “cntE” over the summation of “cntC” and “cntE”. The “rate” denotes the proportion of wrongly segmented tag sequences going through the node. The “TagPair” is the tag sequence generated by the *node generation process* discussed later. A leaf node labeled as the substitution chunk sequences contains the property “cnt” denoting the count of the substitution chunk sequence used to revise the segmentation of the chunk sequence in its branch. In general, a wrong chunk segmentation resulting from the statistical model will pass all the tests on a branch from the root node to a leaf node and therefore the wrong segmentation is substituted by the correct segmentation of the chunk sequence attached to the leaf node. Although a correct chunk segmentation resulting from the statistical model will also be fed to the decision tree, it will normally fail a test on a branch before it reaches a leaf node, meaning no correction is required.

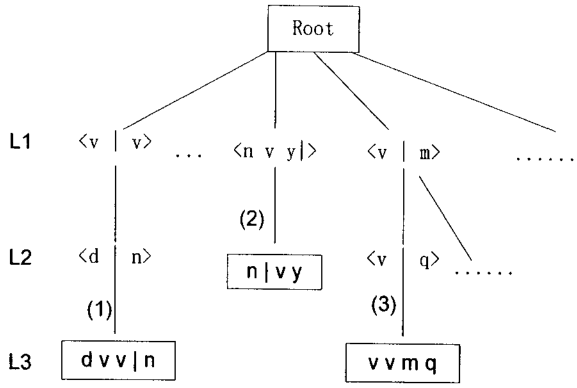


Figure 3.3: A decision tree for segmentation refinement of Chinese chunks.

All learnt rules are organized into the tree structure as shown in Fig. 3.3. All internal nodes are generated by the node generation process, for example, nodes <v|v> and <v|n> shown in the figure. L1 is the first level of the tree in which all nodes are the children of the root node. Similarly, all nodes in the second level L2 are the children of the nodes in L1, and so on. Based on the constructed decision tree, we can achieve search efficiency. Suppose that the maximal number of levels is k and the maximal number of branches of an internal node is n, then the searching complexity is $o(k)$.

The *node generation process* is described as follows:

(1) If there is no segmentation tag ("|") within a tag sequence, that is, the 1-chunk case, e.g., <a b c>, then the generated node is <a b c|>.

(2) If there is only one segmentation tag within a tag sequence, that is, the

2-chunk case, e.g., <a b c | d e>, several nodes are generated by forming the tags pair from both sides of the segmentation tag “|” one by one. If empty is met at left or right, then the remaining tag or tag sequence in the other side together with the segmentation tag “|” is considered as a node. A node generated early is the ancestor of the nodes to be generated next. For example, <a b c | d e> generates nodes <c | d>, <b | e> and <a |>. The node <c | d> is the father node of node <b | e> which is the father node of node <a |>.

(3) If there are more than one segmentation tag in it, that is, a multi-chunk sequence, then the sequence denotes a node, e.g., <a b c | d e | f g h>. This suggests that if the number of chunks in a wrongly segmented chunk sequence is larger than 2, then it will be considered as a node and added to the decision tree directly.

After we obtain the wrongly segmented instances from the statistical model, the decision tree is generated using the following steps:

(1) Obtain all nodes using the incorrect instances from the statistical model by the *node generation process*.

(2) Input the nodes generated sequentially to find the deepest matched node in current decision tree. If the test on all input nodes is passed, then the counter “cntE” of each node in the path is increased by one. If the test of a node is failed, the node and all its offspring nodes are added into the tree hierarchically and set all counters “cntE” to the default value 1.

3.4.3 Training Phase of the CSRA

The training process can be divided into two phases: the learning phase and the validation phase.

Learning phase

In the learning phase, the rules used to correct wrongly segmented chunks from the statistical model are constructed and stored in an XML file corresponding to the tree structure. The procedures are described as follows:

(1) Segment chunks using the statistical model. Wrongly segmented instances are recorded.

(2) Construct the decision tree. The internal nodes and leaf nodes are generated by the node generation process. The node addition operations are carried out to the decision tree.

Validating Phase of the CSRA

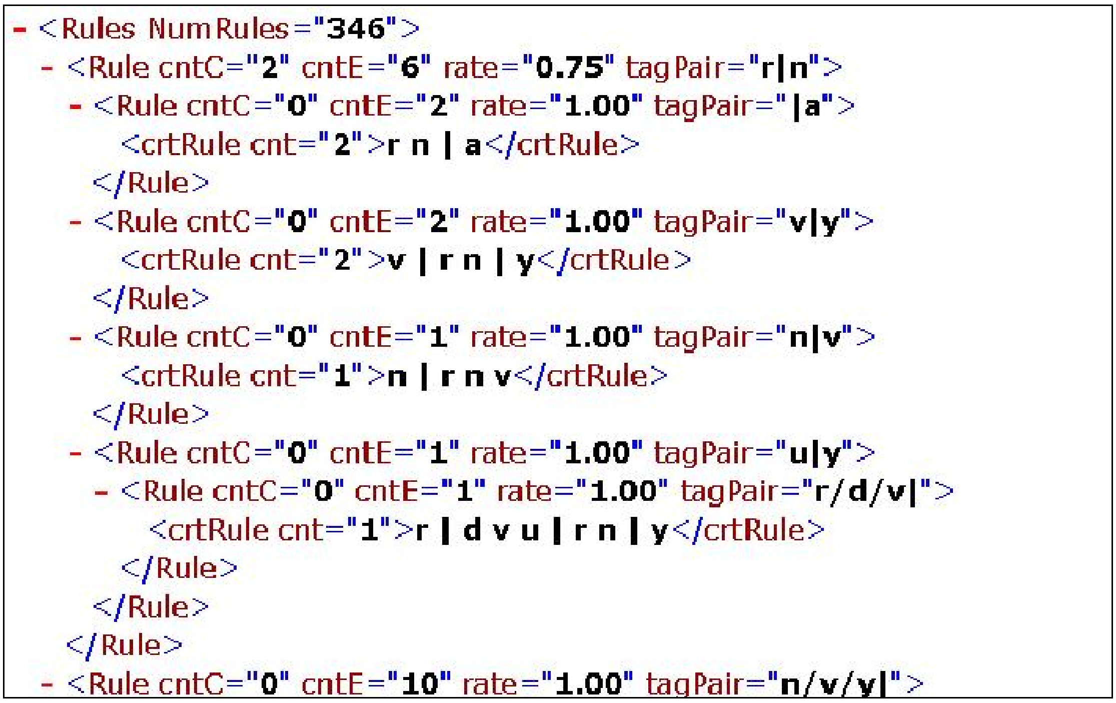
(1) Obtain all correct instances of chunk segmentation from the training corpus in which chunk segmentation has been carried out manually.

(2) Complete the setting of properties “cntC” and “Error Rate”. Corresponding to every individual chunk and every chunk sequence with different lengths, nodes are generated by the node generation process and based on the generated nodes the longest path is found in the decision tree. Then all “cntC” counters of the nodes recording the number of correct instances in the

path are updated. The “cntC” is increased by one when one instance passes it.

When every chunk or chunk sequence of all sentences in the training corpus with manually segmented chunks is fed to the decision tree, the error rate of each internal node is updated. The error rate will be used to test whether the context on the path will be replaced by the substitution chunk sequence or not. If the rate is larger than a preset threshold, we will replace the context on the path with its substitution chunk sequence which is attached to its leaf node. In our experiment, we use 0.5 as the threshold.

Some rule examples are shown in Fig. 3.4:



```
- <Rules NumRules="346">
  - <Rule cntC="2" cntE="6" rate="0.75" tagPair="r|n">
    - <Rule cntC="0" cntE="2" rate="1.00" tagPair="|a">
        <crtRule cnt="2">r n | a</crtRule>
      </Rule>
    - <Rule cntC="0" cntE="2" rate="1.00" tagPair="v|y">
        <crtRule cnt="2">v | r n | y</crtRule>
      </Rule>
    - <Rule cntC="0" cntE="1" rate="1.00" tagPair="n|v">
        <crtRule cnt="1">n | r n v</crtRule>
      </Rule>
    - <Rule cntC="0" cntE="1" rate="1.00" tagPair="u|y">
      - <Rule cntC="0" cntE="1" rate="1.00" tagPair="r/d/v|">
          <crtRule cnt="1">r | d v u | r n | y</crtRule>
        </Rule>
      </Rule>
    </Rule>
  - <Rule cntC="0" cntE="10" rate="1.00" tagPair="n/v/y|">
```

Figure 3.4: Rule examples in XML format.

In Fig. 3.4, the node “Rules” is the root node. Other nodes tagged as “Rule” are internal nodes which are organized into a hierarchical structure. The node tagged “crtRule” is a leaf node of the decision tree which contains the substitution chunk sequence for the context on the path and is used to replace the context. In these nodes, the meanings of elements “cntC”, “cntE” and “cnt” are explained in section 3.4.2.

3.4.4 Test Phase of the CSRA

A chunk sequence using the statistical model is obtained and used as an input chunk sequence to perform the matching process. The matching process is to compare the chunk sequence and the content in the property “tagPair” of the compared node. If they match, then we can say the chunk sequence is found in the tree. The complete steps of using the CSRA for Chinese chunk segmentation are described below (see Figure 3.2):

(1) If the length (the number of chunks) of the compared chunk sequence is larger than 2, for example, the chunk sequence |ABC|DE|FGH|IJ| contains chunks “ABC”, “DE”, “FGH” and “IJ” (‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’, ‘H’, ‘I’ and ‘J’ are POS tags), then the matching process is performed and go to step (2). Otherwise (the length is not longer than 2, for example, the chunks |ABC|DE| or |ABC|), go to step (4).

(2) If the compared chunk sequence can be found at the L1 level of the decision tree and the property “rate” of the found node is larger than the preset threshold (0.5 in our experiment), then the substitution chunk sequence of its leaf node with the maximal property “cnt” is used to replace the compared

chunk sequence, go to step (6); Otherwise go to step (3).

(3) If the compared chunk sequence can not be found at the L1 level of the decision tree, or the property “rate” of the found node is not larger than the preset threshold, then the new chunk sequence is generated by trimming its last chunk and used as a new input compared sequence. For example, the chunk sequence |ABC|DE|FGH| is used as a new input compared sequence after trimming |IJ| from |ABC|DE|FGH|IJ|. Go to step (1).

(4) All the tag pairs are generated from the compared chunk sequence with length 2 by the generation process and used to perform the matching process in the decision tree. For example, the generated tag pairs from |ABC|DE| are “C|D”, “B|E” and “A|”. If the following three conditions are satisfied, then the replacement operation is performed by using the correction chunk sequence in the corresponding leaf node with the maximal property “cnt”. Otherwise, go to step (5). After the replacement operation is performed, go to step (6).

A) All generated tag pairs can be found in the decision tree by the matching process;

B) The last reached node has at least one leaf node;

C) The property “rate” of the last reached node is larger than the threshold 0.5.

(5) If one of the above conditions is not satisfied, then the first chunk (|ABC|) by trimming the second one (|DE|) from the left two chunks (|ABC|DE|) is

considered as an input chunk sequence. If it can be found in the decision tree and the property “rate” of the found node is larger than the threshold, then the substitution operation is performed with the correct one contained in its leaf node with the maximal value of property “cnt”. Otherwise no correction operation is carried out to the first chunk, meaning it has been segmented correctly.

(6) Trim the chunks that have been validated as correct or replaced by the correct ones from the original segmented chunk sequence, then use the remained ones as a new input sequence and go through the step (1)-(6) until all pre-segmented chunks are validated. For example, if the chunk sequence |ABC|DE|FGH|IJ| and |ABC|DE|FGH| cannot be found in the decision tree while the chunk sequence |ABC|DE| can be matched and substituted by the correct one, then the chunk sequence |ABC|DE| is trimmed from the original chunk sequence |ABC|DE|FGH|IJ|, the remained chunks |FGH|IJ| are used as the new input chunk sequence.

3.5 Experimental Results and Discussion

3.5.1 Training Corpus

We use the sentences database provided by Tsinghua University, China, for our experiments. There are 2,030 sentences in the database. Some example sentences from the database are shown in Table 3.5.

Since size of the database is rather small, we perform the experiments based on cross validation. We adopted a 10-fold cross validation approach in which 2,030 sentences in the database were partitioned into ten subsets with each subset containing 203 sentences. The k-th subset (k = 1, 2, ..., 10) contains the (10*n + k)-th sentences in the database (n = 0, 1, 2, …, 202). Each of these subsets was used as the test set once. The remaining subsets (containing 1,827 sentences) after removing the test set were used as the training set. Therefore, ten experiments were conducted. The experimental results on the maximal, average and minimum values of each performance measure on the ten experiments are reported in this thesis.

Table 3.5: Examples in the sentence database

Index	*Sentences with word segmentation and tagging*
1	会议/n 决定/v 接受/v 他/r 的/u 请求/v
2	他/r 的/u 钱/n 被/p 偷/v 了/u
3	我/r 的/u 朋友/n 是/v 一/m 名/q 工程师/n 。/w
4	这里/r 讲/v 的/u 赵概/ngp ，/w 便/d 是/v 这样/r 的/u 一个/m 读书人/n 。/w
5	把/p 你/r 的/u 书/n 放/v 在/v 桌子/n 上/f
6	我们/r 的/u 实验室/n 去年/t 建成/v
7	这里/r 的/u 交通/n 很/d 不/d 发达/a
8	我/r 的/u 帽子/n 飞/v 到/v 树/n 上/f 去/v 了/y 。/w
9	伤兵/n 们/k 都/d 很/d 感激/v 照顾/v 他们/r 的/u 南丁格尔/ngp 。/w

The sentence database contains relative short sentences of diversified syntaxes and different chunk combinations. We performed manual chunk segmentation in the training set and made sure the training data is processed according to our definition of chunks based on RIM.

3.5.2 Experimental Results and Error Analysis

We used precision (*P*), recall rate (*RR*) and incorrect adjustment rate (*IAR*) defined below to evaluate the performance of Chinese chunk segmentation.

Precision (P)

This is defined as the ratio of the number of chunks segmented correctly over the total number of chunks segmented:

$$P = \frac{\text{the number of chunks segmented correctly}}{\text{the total number of chunks segmented}} \qquad (5.1)$$

Recall Rate (RR)

This is defined as the ratio of the number of chunks segmented correctly over the total number of chunks in the database:

$$R = \frac{\text{the number of chunks segmented correctly}}{\text{the total number of chunks in the database}} \qquad (5.2)$$

Incorrect Adjustment Rate (IAR)

This is defined as the ratio of the number of chunks wrongly adjusted using the correction rules over the total number of chunks adjusted using the correction rules:

$$IAR = \frac{\text{the number of chunks wrongly adjusted using the correction rules}}{\text{the total number of chunks adjusted using the correction rules}}$$

(5.3)

Based on the above definitions, we conducted the experiments based on cross-validation. Table 3.6 shows the experimental results on the maximal, average and minimum values for the statistical model (SM) and our proposed combined statistical and rule-based approach (CSRA) when the training sets containing 1,827 sentences were used to generate decision rules. We set the threshold of the error rate of all rules in the decision tree to 0.5.

Table 3.6: Experimental results on the whole training set and the test set

		Close test (in the whole training set)		Open test (in the test set)	
		SM	CSRA	SM	CSRA
Precision (%)	*Max.*	88.54	93.19	71.23	86.88
	Ave.	87.80	92.75	67.05	79.03
	Min.	87.14	92.20	63.98	69.28
Recall rate (%)	*Max.*	82.25	93.13	74.28	93.44
	Ave.	81.67	92.56	67.63	86.47
	Min.	80.49	92.13	61.65	81.38
IAR (%)	*Max.*	NA	4.20	NA	6.45
	Ave.	NA	3.37	NA	3.28
	Min.	NA	2.77	NA	0.00

From Table 3.6, we can see that the precision and recall rate in the close test are rather high especially with refinement by the decision tree. Although the precision and recall rate are relative low in the open test by using SM, the performance was improved significantly after using the decision tree, suggesting that the CSRA can produce a promising performance in Chinese chunk segmentation in terms of the precision and recall rate.

Error Analysis

We found that the wrong segmented chunks mainly contain the following cases:

(1) No correction has been carried out to an incorrectly segmented chunk sequence from the statistical model because the error rate of the nodes corresponding to the wrong segmentation did not pass the threshold (0.5 for our experiment). This accounts for about 56% of incorrect chunk segmentation cases in our experiment.

(2) Wrong correction rules were used to adjust the chunk segmentations. This accounts for about 22% of incorrect chunk segmentation cases in our experiment.

(3) N-gram conflict. In the training and test phases, we consider a tag sequence of POS segmented with n-chunk as a n-gram chunk sequence. Firstly, we test whether the n-gram chunk sequence as a node can be found in the decision tree. If it cannot be found, then the (n-1) chunks after deleting the last one are tested. This process is iterated until part of the sequence can be found, or the final unigram is reached. Therefore, a wrong segmentation will occur if an *i*-gram chunk sequence as a node is found in the decision tree while, in fact, the (*i*-1)-gram chunk sequence should be corrected. This is because the *i*-gram chunk sequence appears prior to the (*i*-1)-gram chunk so the *i*-gram

chunk sequence will be corrected firstly, which is unexpected. Such conflict will occur in spite of the search directions, either from n-gram to unigram or from unigram to n-gram. Such errors account for about 22% of cases.

3.5.3 Detailed Analysis on Decision Rule Learning

To examine the rule learning process and evaluate its performance, we conducted the 10 experiments based on cross validation. Experimental results on the maximal, average and minimum values of each performance measure were tabularized in Table 3.7 and illustrated in Figures 3.5 to 3.10 with analysis.

For each experiment, we construct 9 close test sets from each training set that contains 1,827 sentences. The first close test set consists of 203 tag sequences of POS's (one subset). The second close test set consists of 406 tag sequences of POS's which was formed by adding another subset of 203 tag sequences to the first set. Each new test set was formed by adding another subset of 203 tag sequences to its previous set. The ninth set contains 1,827 tag sequences of POS's, that is, the whole training set. The increment of the test data is to check the relationship between the performance (precision, recall rate and incorrect adjustment rate) and the size of the data set used for decision rule learning. In addition, we have an open test set consisting of 203 tag sequences of POS's. By using of the whole training set, we can obtain the statistical information used to perform chunk segmentation on the nine close test sets. For each close test set, decision rule learning was carried out. The learned rules are used to refine chunk segmentation results. Experimental results on rule leaning are shown in Table 3.7 where CT stands for the close test and OT the open test. CT-P is the precision achieved on the close test and CT-RR is the recall rate on the close test. Similarly, OT-P stands for the precision achieved on the open test and OT-RR is the recall rate on the open test. SM stands for the statistical model. CSRA stands for the combined statistical and rule-based approach. The first row shows the numbers of sentences in every close test set. The numbers of learned rules are given in the second row of the table.

Table 3.7: Experimental results on the maximal, average and minimal values of the ten experiments carried out

# of sentences in the close test set		203	406	609	812	1015	1218	1421	1624	1827
# of learned rules	Max.	85	146	183	230	265	291	321	354	386
	Ave.	82	139	176	218	254	280	308	340	380
	Min.	79	122	161	201	242	268	292	323	366
SM CT-P	Max.	86.39	87.17	88.37	88.74	89.01	88.89	89.02	89.40	88.54
	Ave.	85.12	85.15	86.40	86.88	87.21	87.39	88.19	88.69	87.80
	Min.	82.27	84.04	85.38	85.84	86.36	86.65	87.05	87.98	87.14
SM CT-RR	Max.	80.04	80.67	81.71	82.34	82.81	79.63	82.75	83.37	82.25

	Ave.	78.42	78.52	79.84	80.65	80.74	80.75	81.90	82.62	81.67
	Min.	75.66	77.14	78.53	79.24	79.12	82.08	80.15	81.52	80.49
CSRA CT-P	Max.	97.53	97.02	96.84	96.20	95.72	95.11	94.61	93.82	93.19
	Ave.	97.04	95.27	95.31	95.34	95.03	94.67	94.03	93.36	92.75
	Min.	95.09	94.51	94.67	94.78	94.43	94.16	93.49	92.84	92.20
CSRA CT-RR	Max.	96.44	95.90	95.97	94.93	95.26	94.83	94.19	93.61	93.13
	Ave.	95.91	94.62	94.22	94.45	94.36	94.14	93.62	93.13	92.56
	Min.	94.00	94.12	93.64	93.81	93.91	93.82	93.22	92.84	92.13
CSRA OT-P	Max.	81.79	81.61	83.12	83.99	84.79	87.10	86.88	86.88	86.88
	Ave.	74.21	75.80	76.99	77.49	78.47	78.68	78.78	78.80	79.03
	Min.	63.98	65.72	66.58	67.31	68.37	68.47	69.18	69.28	69.28
CSRA OT-RR	Max.	87.70	87.94	89.62	90.57	91.77	93.68	93.44	93.44	93.44
	Ave.	80.17	82.58	83.71	84.44	85.97	85.71	86.25	86.15	86.47
	Min.	73.85	76.01	77.54	78.57	80.61	80.61	81.38	81.38	81.38
CT-IAR	Max.	0.00	1.76	1.20	1.74	1.83	2.34	3.07	4.01	4.20
	Ave.	0.00	0.60	0.66	0.82	1.23	1.54	2.23	3.11	3.37
	Min.	0.00	0.37	0.28	0.43	0.53	1.08	1.63	2.23	2.77
OT-IAR	Max.	10.53	10.00	10.00	8.40	7.50	6.89	8.70	8.89	6.45

	Ave.	5.32	7.08	5.86	5.08	4.27	3.69	3.80	3.50	3.28
	Min.	0.00	1.33	1.30	1.28	1.25	1.25	0.00	0.00	0.00

We use the sixth column (1,218 sentences) in Table 3.7 as an example to explain the experimental result. On average, 280 rules were generated from the set. If only the statistical model is used only, the average precision and recall rate on the set are 87.39% and 80.75%, respectively. The average precision and recall rate can be improved significantly to 94.67% and 94.14%, respectively, if CSRA is used.

To show the experimental results from different aspects and make them easier to compare, we also show the experimental results in Figures 3.5 to 3.10.

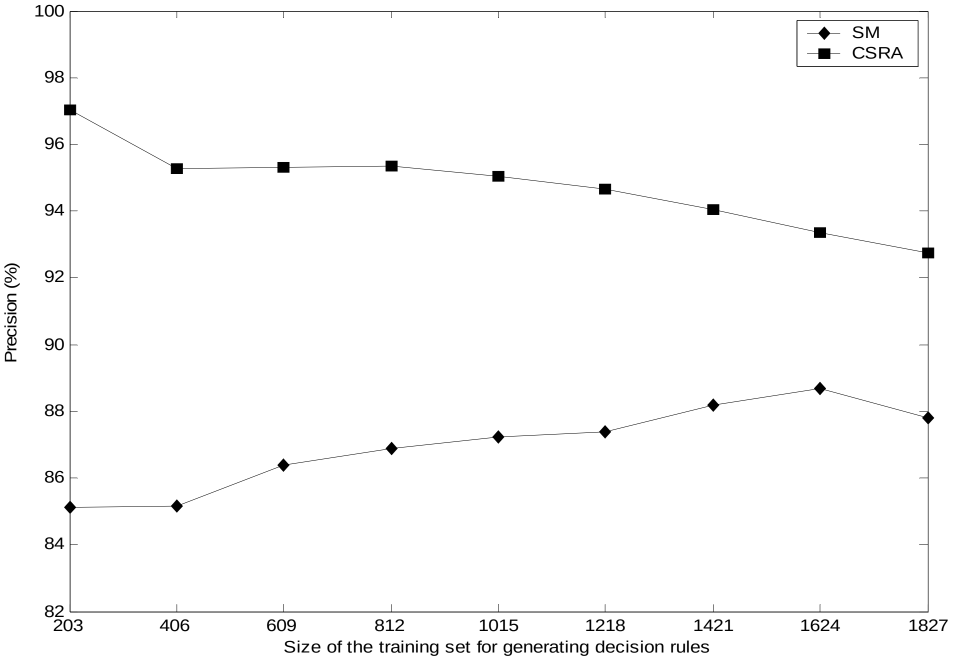


Figure 3.5: The average precision of the SM and CSRA approaches on the close test.

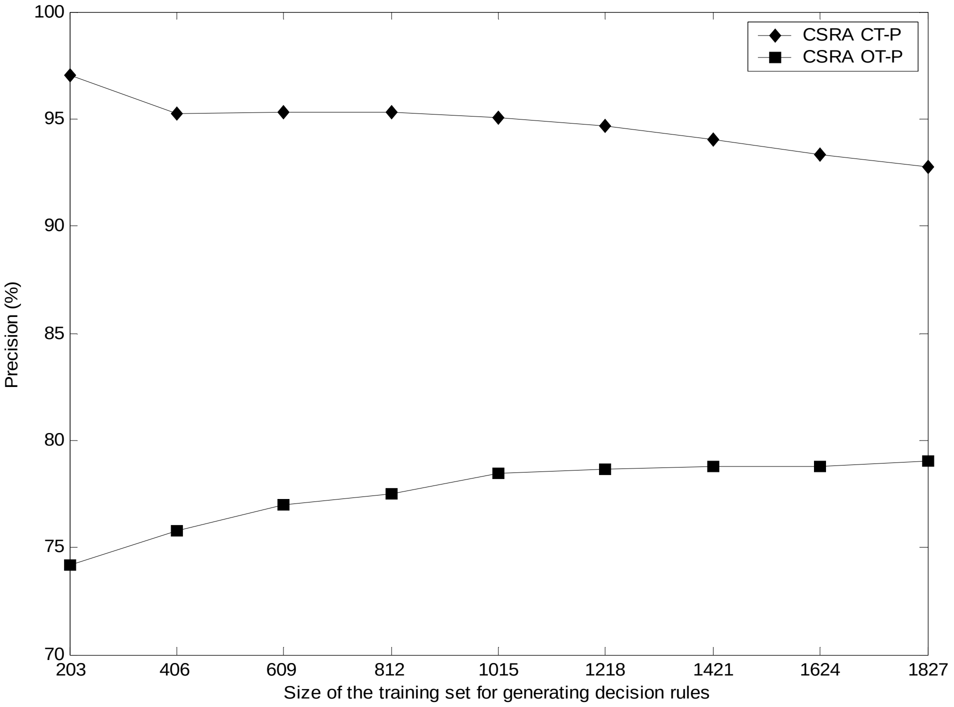


Figure 3.6: The average precision on the close test and the open test using the CSRA.

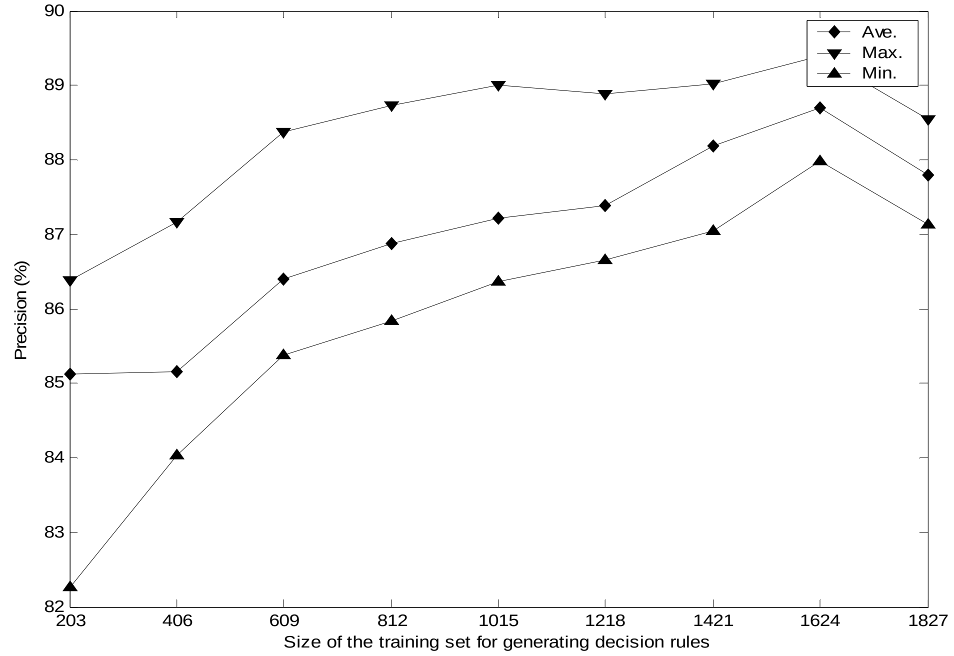


Figure 3.7: The maximal, average and minimal precisions of the SM approach on the close test.

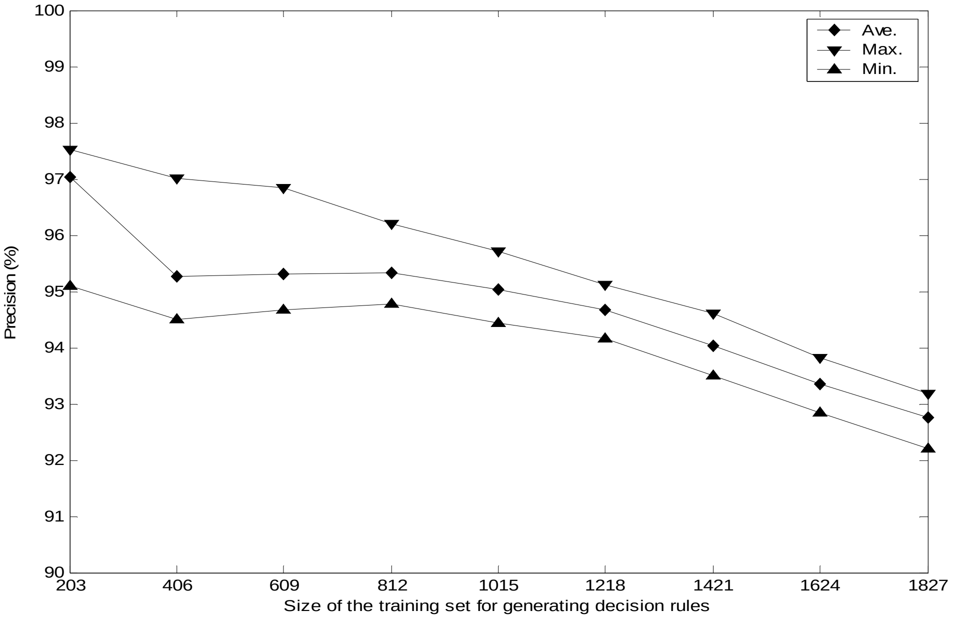


Figure 3.8: The maximal, average and minimal precisions of the CSRA approach on the close test.

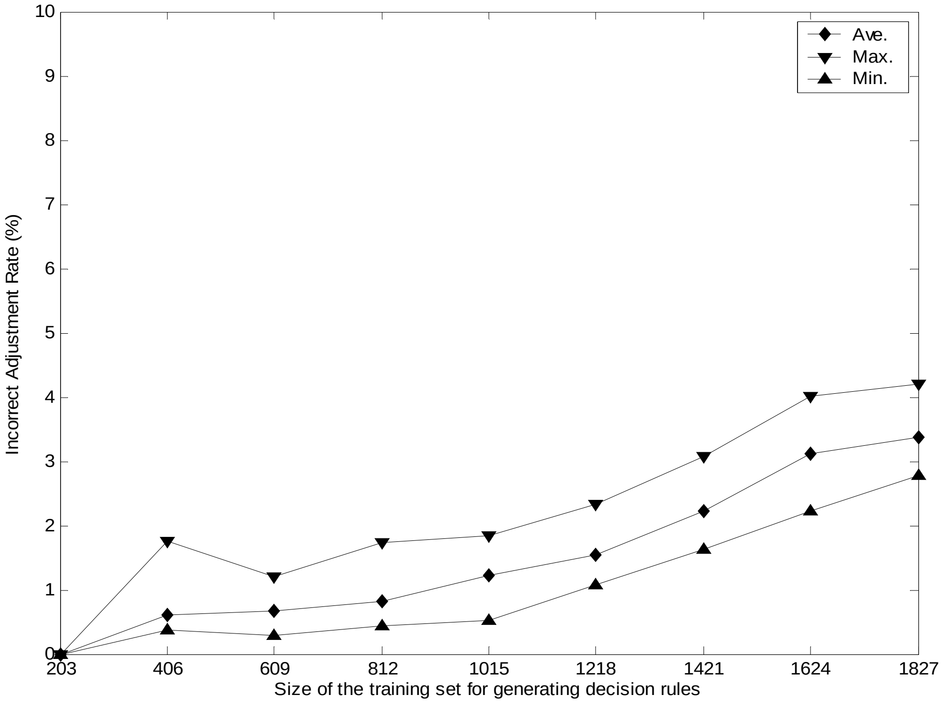


Figure 3.9: The maximal, average and minimal IAR on the close test.

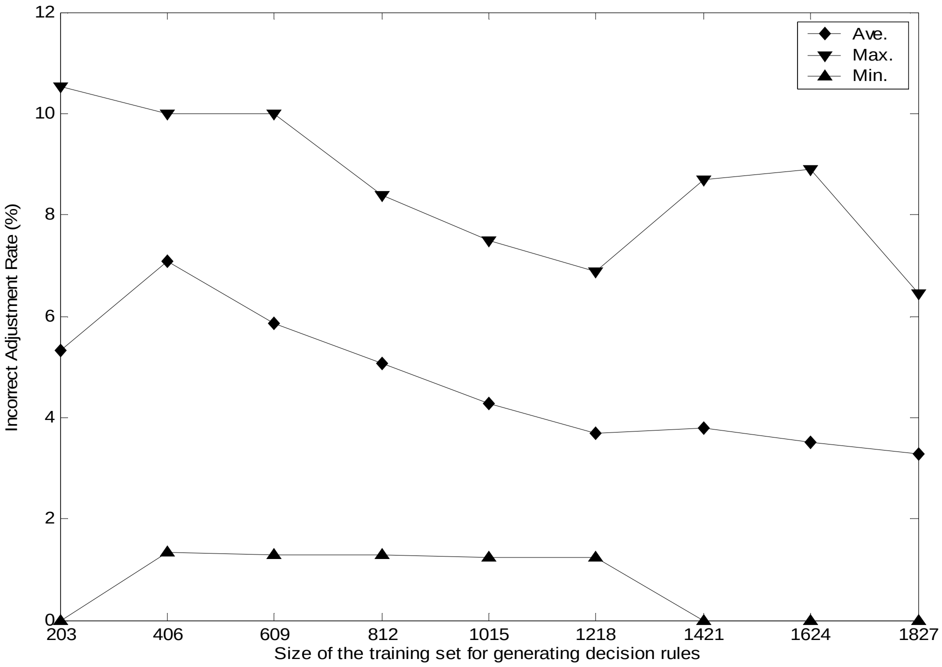


Figure 3.10: The maximal, average and minimal IAR on the open test.

Figure 3.5 shows the average precisions of the SM and CSRA approaches on the close test. We can see that the average precision of the statistical model increases slowly as size of the close test set increases. This is because the statistical parameters were obtained from the whole training set. The average precision of the CSRA decreases slowly as the size of the close test set increases, suggesting that the decision rules will play a less important role when size of test set increases, which accords with our intuition.

Figure 3.6 presents the average precision on the close test and the open test using the CSRA. We can see that the average precision of the open test increases while that of the close test decreases, suggesting again that the CSRA works well for chunk segmentation on unseen sentences, especially when the training set is sufficiently large.

Figures 3.7 and 3.8 show the maximal precision, average precision and minimal precision of the SM approach and CSRA approach on the close test, respectively. From the figures, we can see that when size of the test set increases gradually, the difference between the maximal precision and the average precision, and the difference between the minimal precision and the average precision are descending, which is consistent with our expectation.

Figures 3.9 and 3.10 show the maximal, average and minimal IAR on the close tests and open tests, respectively, versus size of the training set for generating decision rules. We can see that the average IAR on the close test increases slowly but that on the open test decreases slowly, suggesting better generalization ability is achieved when a larger training set is utilized.

3.6 Conclusion

In this chapter, we present a Chinese chunk segmentation method, CSRA, which combines a statistical model and a decision tree approach. In the statistical model, the best chunk segmentation based on the probability of the segmentation point set is obtained. The probability of every possible segmentation boundary is obtained from the training sentence corpus. A decision tree consisting of a number of decision rules is produced by using the incorrect segmentation instances from the statistical model, and is used to refine chunk segmentation. The parameters used in the decision tree are obtained using the training sentence set. Experiment results on a dataset of 2,030 Chinese sentences show that promising results have been produced by the CSRA. The merits of our approach include:

(1) The combined approach achieves both universality and particularity. The statistical information used in the statistical model represents the general distribution of chunk segmentation of a sentence database. On the other hand, the decision rules generated from incorrect segmentation from the statistical model accommodate special cases in chunk segmentation.

(2) Adaptive sentence length. In the rule learning, incorrectly segmented chunks are considered as the context information regardless of their length, meaning that we do not have to set a fixed window size which is commonly adopted in other methods. There are two main drawbacks in setting a fixed window size: (a) only part of the context knowledge used and (b) increased computing complexity for using a large window.

Chapter 4

Similarity Measure of Chinese Sentences

4.1 Introduction

The basis of Example-Based Machine Translation (EBMT) is a large volume of translated bilingual text. EBMT retrieves similar examples of the input sentence from a bilingual corpus (pairs of source/target phrases, sentences, or paragraphs), adapting the translations of the examples to translate the input one. If one of the retrieved examples is the same as the input one, the translation of the example will be directly used as the target translation of the input sentence. Otherwise, the structure of the example's translation is used as a template for translating the input sentence. The differences between the input sentence and the example are identified, and some modifications and substitutions are carried out to the template according to the differences to obtain the translation of the input sentence. Therefore translation template extraction is also very important which has been intensively investigated (Carl, 1999; Cicekli & Güvenir, 1996; Güvenir & Cicekli, 1998; Kaji, et al., 1992; Katoh & Aizawa, 1994; McTait & Trujillo, 1999; Watanabe & Takeda, 1998). The research on sentence templates definitely contributes to the EBMT.

Similarity is an important and fundamental concept which is used widely in a number of different domains. So far, several similarity measures have been proposed, such as information content, mutual information, Dice coefficient, cosine coefficient, distance-based measurements, and feature contrast model. McGill et al surveyed and compared 67 similarity measures used in information retrieval (Lin, 1998).

Similarity measure is a key problem in the EBMT system. For EBMT from Chinese to English, the similarity measure of Chinese sentences is used to determine how similar or dissimilar two Chinese sentences are. Its performance affects directly the final translation of an input sentence. The similarity measure of Chinese sentences can be categorized into two main groups according to the degree of analysis to sentences: string-matching-based and syntactic/semantic-based. For a string-matching-based method, a sentence is viewed as a sequence of words and no or little grammar or structure analysis is performed. The only information used is the surface layer information of sentences, e.g., word sequence and part of speech (POS) information. One simple and traditional string-matching-based measure is based on Dice Coefficient measure. That is the ratio of the number of identical words of two sentences and the average number of words in the two sentences. The following Table 4.1 shows the main idea of Dice Coefficient measure.

Table 4.1: Dice coefficient measure

	# of words in s_2	# of words not in s_2
# of words in s_1	c_{11}	c_{12}
# of words not in s_1	c_{21}	0

The similarity between the two sentence s_1 and s_2 is defined as

$$\frac{2 * c_{11}}{(c_{11} + c_{12}) + (c_{11} + c_{21})} \qquad (4.1)$$

Where c_{11} is the number of words both in the two compared sentence s_1 and s_2, c_{12} is the number of words in s_1 but not in s_2, c_{21} is the number of words in s_2 but not in s_1. Simplicity is the main advantage of this measure. It has been applied to both Chinese and English sentence similarity measure. There are always two sides to everything. The main drawback of the measure is that neither syntactic nor semantic information of two sentences is considered. So it cannot identify two sentences which are similar in structure. However, structure information is a kind of important content contained in sentences. It should be expressed in a good sentence similarity measure.

Another string-matching-based similarity measure for English sentences was introduced in (Mandreoli, et al., 2002). The similarity between two sentences is measured by a distance function defined as the minimum number of editing operations (i.e., insertions, deletions, and substitutions) of single terms (words) needed to transform the first terms (words) sequence into the second one. The measure is more suitable for English and other Latin-based languages than oriental languages such as Chinese, although an improved method using editing distance was proposed for the similarity measure of Chinese sentences (SMCS) (Che, et al., 2003).

For SMCS, some researchers proposed that the overall similarity of two sentences can be obtained by computing and summing up the similarity value of every word pair in two compared sentences (Li, 2002). In this method, some other language resources such as a semantic dictionary should be used for obtaining the similarity value of two words.

The syntactic/semantic-based sentence similarity measure makes use of the syntactic/semantic structures of sentences. For SMCS, Sui presented a syntactic/semantics-based similarity measure based on the skeletal dependency analysis (Sui & Yu, 1998). However, performing the skeletal analysis is not only time-consuming but also error-prone.

Currently, a trend for sentence similarity measure is to combine several methods so as to make use of information of several aspects (Chen, et al., 2001; Li, 2002; Li, et al., 2003(a)). For example, the method combining semantic and syntactic information was proposed in (Li, et al., 2003(a)), which makes use of other language resources such as HowNet. However, the accuracy of dependency analysis is not satisfactory, which affects significantly the final similarity measure performance. In (Chen, et al., 2001), a new multi-level feature-based approach was presented to extend the metrics of similarity measure of sentences from the lexical level to the syntactic and semantic levels. However, the algorithm is quite complex and the weight assignment is too subjective.

In this chapter, firstly we introduce a word/chunk-sequence-matching-based similarity measure and a similarity measure of Chinese sentences' structures for SMCS. The chunk segmentation results are generated using the chunk segmentation approach discussed in Chapter 3. For structural similarity measure, the POS tag sequences of the two sentences are compared and the relationship between the segments of the POS tag sequence is explored. Based on the relationship identified, we can measure the similarity of two sentences' structures. In both methods, we weigh the importance of different POS's in sentences by assigning different weights to them. Secondly, we combine these two methods with weight factors. For obtaining more reasonable weight assignment, we introduce a human-computer interaction approach to the current SMCS based on relevance feedback. Unlike the computer centric approach, where the weights are fixed, the proposed interactive approach allows the computer to fine tune the weights via users' relevance feedback. This part of work will be discussed in Chapter 5.

4.2 Perceptions and Assumptions

Since our objective is to provide a more reasonable definition of the intuitive concept of the similarity of Chinese sentences, we first describe our perceptions about similarity.

Perception 1: The similarity between sentence A and sentence B is related to their commonality. The more commonality they share, the more similar they are.

Perception 2: The similarity between sentence A and sentence B is related to the differences between them. The more different they are, the less similar they are.

Perception 3: The maximum similarity between sentence A and sentence B is reached when A and B are identical, no matter how much commonality they share (Lin, 1998).

Our goal is to arrive at a definition of similarity of Chinese sentences that captures the above perceptions.

Perception 3 indicates that the similarity measure reaches a maximum when the two Chinese sentences are identical. We assume the maximum is 1.

Assumption 1: The similarity between two identical sentences is 1.

When there is no commonality between sentences A and B, we assume their similarity is 0, no matter how different they are.

Assumption 2: The similarity between a pair of completely different sentences is 0.

4.3 Word-Sequence-Matching Based Similarity Measure

As we mentioned in the previous section, the Similarity Measure of Chinese Sentences (SMCS) plays an extremely important role in example-based machine translation from Chinese to English. We propose two methods for measuring the similarity of Chinese sentences by making use of word sequence information and POS tag sequence information, respectively.

In the word-sequence-matching-based (WSMB) method, we take three factors into consideration. They are the number of identical word sequences, the length of identical word sequences, and the average weighting (AW) of identical word sequences in two compared sentences. We will assign a weighting to every Chinese POS tag according to the importance of the POS in a Chinese sentence. Two sentences are more similar if the similarity measure between them is greater. That is, the more identical word sequences, the longer these word sequences, and the greater the average weighting of these word sequences, the more similar these two sentences.

Based on this idea, we propose the following sentence similarity measure:

$$Sim_1(s_1, s_2) = \frac{2n\sum_{i=1}^{n} i^2 (\sum_{j=1}^{C_i} AW_j^i)}{(m+n)n\sum_{l=1}^{n} w_l} \quad (m \geq n) \qquad (4.2)$$

The average weighting (AW) of the *j*-th identical word sequence with length *i* is defined as

$$AW_j^i = \frac{1}{i}\sum_{k=1}^{i} w_k^j$$

Where w_k^j is the weighting factor of POS tag corresponding to the *k*-th word in the *j*-th identical word sequence with length *i*.

The Eq. (4.2) can be further expressed as

$$Sim_1(u_1,u_2,...u_p) = \frac{2\sum_{i=1}^{n} i*(\sum_{j=0}^{C_i}\sum_{k=1}^{i} w_k^j(u_1,u_2,...u_p))}{(m+n)*(\sum_{l=1}^{n} w_l(u_1,u_2,...u_p))} \qquad (4.3)$$

where *m* and *n* are the numbers of words of the longer and shorter one of two sentences to be compared, *i* is the length of identical word sequences and C_i is the number of word sequences with length *i*. $w_k^j(u_1,u_2,...,u_p)$ is the weighting of the POS tag of the *k*-th word in the *j*-th identical word sequence with length *i*, e.g., if the POS tag of the word is tag *v*, then $w_k^j(u_1,u_2,...,u_p) = u_v$; $w_l(u_1,u_2,...,u_p)$ is the weighting assigned to the POS tag of the *l*-th word in the short sentence, $u_1, \ldots u_p$ are parameters used to weigh the importance of POS's when the POS tag is tag 1, tag 2, …, tag *p*, respectively. It is easy to show that the above equation takes values ranging from 0 to 1.

If no identical word sequence has a length of greater than 1, Eq. (4.3) is similar to Eq. (4.1) except that the weightings of POS's tags are considered in the former. It is obvious that Eq. (4.3) is superior to Eq. (4.1) because it also makes use of the lengths and the weightings of identical word sequences in two compared sentences. The proposed measure can distinguish among two special cases. The first case is that two sentences are identical, meaning that they have the same words in the same sequences. The second case is that two sentences have the same words, but these words are inversed in order. For these two cases, the similarity measure by Eq. (4.1) is the same, a value of 1. However, the similarity value of the first case is equal to 1 while that of the second case is much smaller than 1 if Eq. (4.3) is adopted.

An example of similarity measure using the word-sequence-matching-based method is given in Table 4.2. An assumption is that all sentences to be compared have been performed word segmentation and tagging. In the current measure, we assign value 3 to nouns, value 5 to verbs and 1 to all other POS's.

Table 4.2: An example of Chinese sentence and their similarity measure

Compared sentences	我/r 喜欢/v 看/v 电视/n 。/w (source sentence) (I like to watch TV.)	
	Our Method	The measure defined by Eq. (4.1)
我/r 不/d 喜欢/v 电视/n 。/w (I do not like to watch TV.)	0.16	0.75
我/r 喜欢/v 看/v 电影/n 。/w (I like to watch movie.)	0.59	0.75

From Table 4.2, we can see that our proposed method can distinguish among the two compared sentences while the measure defined in Eq. (4.1) cannot distinguish one from the other. Part of the experimental result is shown in the Table 4.3.

Table 4.3: Retrieved sentences with the word-sequence-matching based similarity measure (N1: the source sentence number; N2: the rank of retrieved sentences)

N1	N2	Retrieved Sentences
1	1	带动上周五晚道指收市升 112.38 点 (Motivated the Dow up 112.38 pts last Friday night,)
	2	上周五晚道指收市升 38.93 点，(The Dow up 38.93 pts last Friday night,)
	3	收市升 45.36 点，(Closed up 45.36 pts,)
	4	上周五晚道指收市跌 70.20 点，(The Dow down 70.20 pts last Friday night,)
	5	收市仍升 23.16。(Closed still up 23.16.)
2	1	回补大部份 13312/13360 下跌裂口后反覆回软，(Repeatedly eased after covering most of downward gap at 13312/13360,)
	2	回补大部份 13312/13360 下跌裂口后回调至 12884，(Corrected to 12884 after covering most of downward gap at 13312/13360,)
	3	回补大部份上升裂口，(Covering most of the upward gap,)
	4	回补大部份 13155/13323 上升裂口，(Covering most of the upward gap at 13155/13323,)
	5	回补大部份 13155/13323 上升裂口后反弹，(Rebounded after covering most of the upward gap at 13155/13323,)
3	1	港股今早表现反覆，(The HK stocks performed repeatedly this morning,)
	2	表现反覆，(Performed repeatedly,)

	3	港股今早表现造好，(The HK stocks performed well this morning,)
	4	带动港股今早表现造好，(The HK stocks were stimulated to perform well this morning,)
	5	港股今早反覆偏软，(The HK stocks performed repeatedly bearish this morning,)
4	1	其他原料股及能源股则窄幅波动。(Other raw material and energy stocks fluctuated within a narrow range.)
	2	其他原料股及航运股同告下挫，(Other raw material and shipping stocks slid together,)
	3	原料股个别发展，(Raw material stocks developed individually,)
	4	原料股普遍下跌，(Raw material stocks slid universally,)
	5	原料股普遍反弹，(Raw material stocks rebounded universally,)
5	1	嘉华建材(0027-HK)跌 4.87，(K. Wah Cons (0027-HK) slid 4.87,)
	2	嘉华建材 (0027-HK)跌 3、澳门实德 (0487-HK)跌 1.4，(K. Wah Cons (0027-HK) slid 3, Macau Success (0487-HK) slid 1.4,)
	3	嘉华建材 (0027-HK)升 3.68，(K. Wah Cons (0027-HK) jumped 3.68,)
	4	嘉华建材 (0027-HK)狂升 33.14，(K. Wah Cons (0027-HK) jumped greatly 33.14,)
	5	嘉华建材 (0027-HK)无起跌，(K. Wah Cons (0027-HK) flat,)

Table 4.4: Source sentence used in Table 4.3

No.	Source Sentences
1	带动上周五晚道指收市升 112.38 点，(Motivated the Dow up 112.38 pts last Friday night,)
2	回补大部份 13312/13360 下跌裂口后反覆回软，(Other raw material and energy stocks fluctuated within a narrow range,)

3	港股今早表现反覆，(The HK stocks performed this morning,)
4	其他原料股及能源股则窄幅波动。(Other raw material and energy stocks fluctuated within a narrow range.)
5	嘉华建材(0027-HK)跌 4.87，(K. Wah Cons (0027-HK) slid 4.87,)

From Table 4.3, we can see that the most similar sentence retrieved from the sentence database is the source sentence itself. However, in practical MT system, it is not always possible to find identical sentences. We can use the top similar sentences to generate translation of an input sentence. From the table, we can also see that the retrieved sentences for every source sentence, which are listed in the order of decreasing similarity measure, are quite consistent with human judgement. For example, for the first source sentence (带动上周五晚道指收市升 112.38 点), from our judgment, we know the second retrieved sentence (上周五晚道指收市升 38.93 点，) is more similar to the source sentence than the third retrieved one (收市升 45.36 点，). The same conclusion can be drawn for the third and the fourth retrieved sentences. It appears that our proposed method is quite promising for the similarity measure of Chinese sentences. It is also observed that some retrieved sentences are quite different from their source ones due to a small sentences database used. This problem can be overcome if the sentences database is significantly expanded. How to utilize the top retrieved sentences is another important problem in EBMT. Even if the retrieved sentences are not very similar to the source one due to a lack of similar sentences in the database, we can still use them as a reference for obtaining the translation of the input sentence.

4.4 Similarity Measure of Chinese Sentences with Chunks

Although the word-sequence-matching-based method for the similarity measure of Chinese sentences (SMCS) can obtain a better performance than conventional techniques, some further improvements can be made by overcoming some of its drawbacks. It is recognized that the structures of two Chinese sentences may be different even if they have completely identical POS tag sequences, meaning the two compared sentences may have different chunk components. As we know, sentence structure information is very important for sentence translation. So, a sophisticated sentence similarity measure should be able to differentiate the differences in the chunk components of two sentences. By extending the word-based approach to chunk-based approach, we can obtain a similarity measure based on chunk information.

4.4.1 Main Idea

The main idea of the chunk-based method can be described as follows.

(1) Firstly, we perform chunk segmentation on all sentences in the database by using the method introduced in Chapter 3.

(2) The similarity measure of two compared word sequences defined in Eq. (4.3) can be enhanced by including the chunk information. The

improved measure is expressed in Eq. (4.4).

(3) If two compared word sequences are identical and they have the same chunk segmentation, then these two word sequences have larger similarity value. That is, the coefficient is equal to 1. If two compared identical word sequences have different chunk segmentations, then their similarity value will be reduced by multiplying a scale which is less than 1.

The improved similarity measure is defined as

$$Sim_1^{'}(s_1, s_2) = \frac{2\sum_{i=1}^{n} i * (\sum_{j=1}^{C_i} (\frac{1+2*D_j}{1+A_{j1}+B_{j2}} \sum_{k=1}^{i} w_k^j))}{(m+n)*(\sum_{l=1}^{n} w_l)} \quad (m \geq n) \qquad (4.4)$$

where *m* and *n* are the numbers of words of the longer and shorter one of two sentences to be compared. D_j is the number of segmentation tags in the same position in two identical word sequences, A_{j1} and B_{j2} are the numbers of segmentation tags included in two identical word sequences respectively. Note that the chunk segmentation tag "|" is not considered when two word sequences are compared. For example, for two word sequences $AB|CD|EF|$ and $A|BC|D|EF|$ (*A*, *B*, *C*, *D*, *E* and *F* are Chinese words), we consider they are identical in word sequences. D_j is equal to 2, A_{j1} and B_{j2} are equal to 3 and 4 respectively. The other parameters, like *i*, w_k^j, C_i, w_l, are the same as Eq. (4.3). Similar to Eq. (4.3), it is easy to show that the Eq. (4.4) takes values ranging from 0 to 1.

4.4.2 Experiment Results and Discussion

In the experiment, we use the same sentences database as the one used in Chapter 3. There are 2,030 Chinese sentences in the database. Chunk segmentation was carried out first to all the sentences in the database. The main objective of our research is to develop methods for Chinese sentence similarity measure but due to the difficulty and time-consumed in collecting the data, it is only possible at this point to make use of a small database for evaluating our proposed methods. Some sentence examples in the database are shown in Table 4.5.

In our experiment, 30 Chinese sentences from the database were used as input sentences to retrieve top five most similar sentences from the database by using the word-sequence-matching-based method and the improved measure with chunk segmentation, respectively. Experimental results for the source sentences shown in Table 4.6 are shown in Table 4.7.

Table 4.5: Some examples from the sentence database

Index	Sentences with word segmentation and tagging, and chunk segmentation
1	安装/v 在/p 桌子/n 上/f 的/u 灯/n |nc 亮/a 了/y |adjc
2	他/r |nc 总是/d 第一/m 个/q 来/v |vc 。/w |oc
3	动力/n 问题/n |nc 必须/d 引起/v |vc 足够/a 的/u 重视/vn |nc
4	工厂/n |nc 决定/v 建立/v |vc 一/m 所/q 学校/n |nc
5	计算机/n |nc 可以/v 做/v |vc 很多/m 事情/n |nc
6	这儿/r |nc 本来/d 是/v |vc 海/n |nc 。/w |oc

7	我/r |nc 特别/d 喜欢/v |vc 农村/n 的/u 生活/n |nc
8	我/r |nc 有/v |vc 两/m 本/q 书/n |nc

Table 4.6: Sources sentence used in Table 4.7

No.	Source Sentences
1	村子里的人们都喜欢看戏 (All people in the village like to enjoy a drama)
2	这是一个最好的老师 (This is a greatest teacher)
3	这个书包是小亮的。(This schoolbag is Xiaoliang’s.)
4	妈妈买了八条手巾。(Mother bought eight hand towels.)
5	我们在学华语。(We are learning the Chinese.)
6	他们在教室里上课。(They have a class in the classroom)
7	小花请他们吃水果。(Xiaohua invites them for fruits.)
8	妈妈给小白一个小黑板。(Mother gives Xiaobai a small blackboard.)
9	我们坐汽车去动物园。(We go to the zoo by car.)
10	工厂决定建立一所学校 (The factory decides to establish a school)

Table 4.7: Retrieved sentences with the original and improved Chinese similarity measures based on word/chunk sequences (N1: the source sentence number; N2: the rank of retrieved sentences; I: the improved measure; O: the original measure)

N1	N2	Retrieved Sentences
1	1	I: 村子里的人们都喜欢看戏(All people in the village like to go to theatre.)
		O: 村子里的人们都喜欢看戏(All people in the village like to go to theatre.)
	2	I: 她很喜欢看电影 (She like to watch movie very much.)
		O: 村子里农民不多 (There are a few farmers in the village.)
	3	I: 你喜欢看电视，(You like to watch TV,)
		O: 你看，(Look,)
	4	I: 还是喜欢看电影？(Or like to watch movie?)
		O: 我喜欢这些书 (I like these books.)
	5	I: 我们一家人都喜欢参加活动。(All our family members like to participate in these activities.)
		O:我们一家人都喜欢参加活动。(All our family members like to participate in these activities.)
2	1	I: 这是一个最好的老师 (This is a greatest teacher.)
		O: 这是一个最好的老师 (This is a greatest teacher.)
	2	I: 这是一本书。(This is a book.)

		O: 这是舌头。(This is the tongue.)
	3	I: 这是舌头。(This is the tongue.)
		O: 他是一个学生。(He is a student.)
	4	I: 他是一个学生。(He is a student.)
		O: 王子又问这是什么野兽 (The prince asked what is this wild animal.)
	5	I: 埃及是一个文明古国，(Egypt is an ancient civilized nation.)
		O: 你是谁？(Who are you?)
3	1	I: 这个书包是小亮的。(This schoolbag is Xiaoliang’s.)
		O: 这个书包是小亮的。(This schoolbag is Xiaoliang’s.)
	2	I: 你是谁？(Who are you?)
		O: 你是谁？(Who are you?)
	3	I: 起初是这样，(In the beginning, it is like this,)
		O: 起初是这样，(In the beginning, it is like this,)
	4	I: 这是为什么呢？(Why is this?)
		O: 明天是植树节，(Tomorrow is an arbor day,)
	5	I: 明天是植树节，(Tomorrow is an arbor day,)
		O: 很高兴。(Very happy.)
4	1	I: 妈妈买了八条手巾。(Mother bought eight hand towels.)
		O: 妈妈买了八条手巾。(Mother bought eight hand towels.)
	2	I: 妈妈买了一些菜，(Mother bought some vegetables.)
		O: 妈妈买了一些菜，(Mother bought some vegetables.)

	3	I: 他不要手巾。(He does not want a hand towel.)
		O: 他不要手巾。(He does not want a hand towel.)
	4	I: 买文具的钱，(The money for buying stationery,)
		O: 很高兴。(Very happy.)
	5	I: 妈妈很担心，(Mother is very anxious.)
		O: 好不悲惨。(Very pathetic)
5	1	I: 我们在学华语。(We are learning Chinese.)
		O: 我们在学华语。(We are learning Chinese.)
	2	I: 学华语，(Learn Chinese.)
		O: 学华语，(Learn Chinese.)
	3	I: 我还要跟妈妈学，(I still need to learn from mother.)
		O: 何必学呢？(Why to learn?)
	4	I: 巴比星很认真地学，(Babixin studies very seriously.)
		O: 很高兴。(Very happy.)
	5	I: 何必学呢？(Why to learn?)
		O: 好不悲惨。(Very pathetic.)
6	1	I: 他们在教室里上课。(They are having a class in the classroom)
		O: 他们在教室里上课。(They are having a class in the classroom)
	2	I: 丁松在明亮的教室里上课。(Dingsong attends class in the bright classroom.)
		O: 丁松在明亮的教室里上课。(Dingsong attends class in the bright classroom.)

	3	I: 在教室外面有几只小鸡 (There are some chickens outside the classroom)
		O: 很高兴。(Very happy.)
	4	I: 他们在这儿种植农作物，(They plant crops here.)
		O: 好不悲惨。(Very pathetic.)
	5	I: 顺吉哥哥不是到夜校去上课，(The elder brother Sunjie is not having a class in night school.)
		O:上课的时候，(When we have a class,)
7	1	I: 小花请他们吃水果。(Xiaohua invites them to eat fruits.)
		O: 小花请他们吃水果。(Xiaohua invites them to eat fruits.)
	2	I: 姐姐去买水果。(The elder sister went to buy fruits.)
		O: 很高兴。(Very happy.)
	3	I: 他没有东西吃 (He has nothing to eat.)
		O: 好不悲惨。(Very pathetic.)
	4	I: 请听我说！(Please listen to me!)
		O: 他没有东西吃 (He has nothing to eat.)
	5	I: 经理通常和大家一起吃午餐 (The Manager usually has lunch with all colleagues.)
		O: 心中思潮起伏。(Have disquieting and surging thoughts in heart.)
8	1	I: 妈妈给小白一个小黑板。(Mother gave Xiaobai a small blackboard.)
		O: 妈妈给小白一个小黑板。(Mother gave Xiaobai a small blackboard.)
	2	I: 妈妈很担心，(The mother is very anxious.)
		O: 很高兴。(Very happy.)

	3	I: 我爱妈妈。(I love my mother.)
		O: 好不悲惨。(Very pathetic.)
	4	I: 哈山事先给了我们地址，(Hashan gave us the address in advance,)
		O: 心中思潮起伏。(Have disquieting and surging thoughts in heart.)
	5	I: 也仍然晶莹夺目。(Still crystal-clear and brilliant)
		O: 我们在山上一个小屋里度假 (We spent holidays in a cabin on the mountain.)
9	1	I: 我们坐汽车去动物园。(We went to the zoo by car.)
		O: 我们坐汽车去动物园。(We went to the zoo by car.)
	2	I: 伯伯坐汽车去火车站。(The uncle went to the station by bar.)
		O: 他坐在那里，(He sat there.)
	3	I: 他坐在那里，(He sat there.)
		O: 很高兴。(Very happy.)
	4	I: 你要去礼堂，(You will go to the hall.)
		O: 好不悲惨。(Very pathetic.)
	5	I: 自己骑上马去。(I Myself go by riding on a horse.)
		O: 心中思潮起伏。(Have disquieting and surging thoughts in heart.)
10	1	I: 工厂决定建立一所学校 (The factory decides to establish a school.)
		O: 工厂决定建立一所学校 (The factory decides to establish a school.)

	2	I: 工厂决定生产这种仪器 (The factory decides to produce this kind of instrument.)
		O: 工厂决定生产这种仪器 (The factory decides to produce this kind of instrument.)
	3	I: 学校决定录用他 (The school decides to employ him.)
		O: 十年前我在学校读书 (I studied at the school ten years ago.)
	4	I: 学校建在工厂旁边 (The school is built at the side of the factory.)
		O: 学校决定录用他 (The school decides to employ him.)
	5	I: 叭的一声，(“Ba”)
		O: 学校建在工厂旁边 (The school is built at the side of the factory.)

From Table 4.7, we can see that, for either method used, the most similar sentence retrieved from the sentence database is the source sentence itself. However, in practical MT system, it is not always possible to find identical sentences. We can use the top similar sentences to generate translation of an input sentence. From the second top retrieved sentence onwards, the retrieved sentences using the improved method with chunk segmentation appear to be better than the method without chunk segmentation, although such judgment is of subjective nature. For example, for the first source sentence (村子里的人们都喜欢看戏), from our subjective judgment, we know the second retrieved sentence (她很喜欢看电影) using the improved method is more similar to the source sentence than that (村子里农民不多) of using the original method. The same conclusion can be drawn for the third and the fourth retrieved sentences.

From other examples we also see that, in general, the retrieved sentences using the improved method are more similar to their source sentences than those retrieved by using the original method. It is also observed that some retrieved sentences are quite different from their source ones due to a small sentences database used. This problem can be overcome if the sentences database is significantly expanded. How to utilise the top retrieved sentences is another important problem in EBMT. Even if the retrieved sentences are not very similar to the source one due to a lack of similar sentences in the database for some input sentences, we can still use them as a reference for obtaining the translation of the input sentence. For example, if we want to translate the tenth source sentence (工厂决定建立一所学校), we can refer to the translation of the third retrieved sentence by the improved method (学校决定录用他) on the translation of “学校” and “决定”.

4.5 POS-Tag-Sequence-Matching Based Method

For a structure-matching-based method, its objective is to measure the similarity of Chinese sentences’ structures. If the constituents in two Chinese sentences are similar, then we can say these two Chinese sentences are similar in structure. The main idea of this similarity measure is that we perform matching between POS’s of two Chinese sentences. The POS weighting is also utilized in this process.

We used the directed graph to model the POS tag sequence of a sentence where the tag of POS is represented using a node and a directed weighted link is used to connect two sequent nodes. The start node is labeled with “*”. The end node is labeled with “#” with one directed link of weight $W_0\ (=0)$ (Fig. 4.1). Now suppose that we have the following two POS tag sequences to represent two sentences S_1 and S_2 (n≥m):

S_1 : $s_{1,1}s_{1,2}s_{1,3},\cdots,s_{1,n-2},s_{1,n-1},s_{1,n}$

S_2 : $s_{2,1}s_{2,2}s_{2,3},\cdots,s_{2,m-2},s_{2,m-1},s_{2,m}$

In above sequences, $s_{1,i}$ (1≤i≤n) denotes the *i*-th POS tag of S_1. And $s_{2,j}$ (1≤j≤m) denotes the *j*-th POS tag of S_2. To describe the algorithm more clearly, the above POS tag sequences can be represented in the directed graphs as shown in Fig. 4.1.

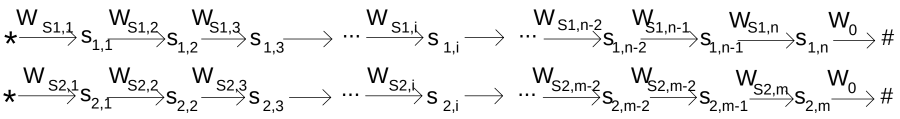


Figure 4.1: Directed graphs of sentences S_1 and S_2.

From Fig. 4.1, we can compute the structural similarity of two POS tag sequences. It can be carried out in two steps. The first step is to perform the preliminary forward matching from the start node (*) to the end node (#). The second step is to carry out the backward matching refinement from the end node to the start node. Figure 4.2 shows the flowchart of the two-step matching process.

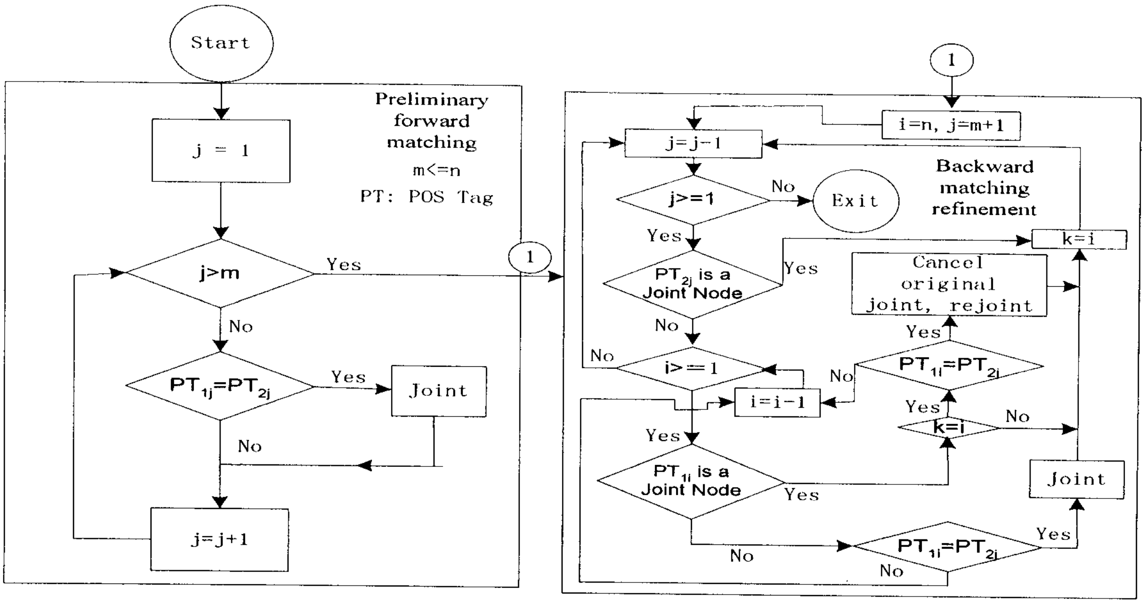


Figure 4.2: Flowchart of the two-step matching process.

(1) Preliminary Forward Matching

Firstly, we perform a forward matching from left to right between the tags (nodes) of POS's at the same positions of the two sequences. If two nodes of the same subscript are of the same tag, then they are joined as a Joint Node (JN). This process is repeated until all nodes in the shorter sequence are compared because the shorter sequence will reach its end first. After this process, it is possible that one or several closed loops are formed such as one shown in Fig. 4.3 where one closed loop was formed between $S_{1,2}$ ($S_{2,2}$) and $S_{1,\text{n-2}}$ ($S_{2,\text{n-2}}$).

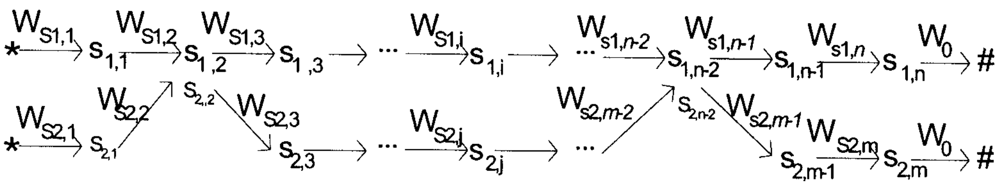


Figure 4.3: The preliminary forward matching process and a possible outcome.

(2) Backward Matching Refinement

The backward matching refinement is an important process for sentence structural similarity measure. The aim of the backward matching refinement process is to adjust the joint points to generate smaller closed loops. Smaller closed loops are considered more reasonable, which can be explained by two special cases. Firstly, if there is no closed loop, then the two compared structures are completely different. Secondly, if every closed is minimum (the joint node only), that is, every node is a joint node, then the two structures are identical. Moreover, in preliminary forward matching, two identical nodes in the same subscript are joined as a JN. So, if two nodes preceeding the current JN are identical, then they can be jointed to be another JN. That means the previous larger closed loop is divided into two smaller closed loops. The examples are showed in Fig. 4.4 and Fig 4.6. The backward scan is done in the shorter sequence and the longer sequence synchronously. A certain closed loop generated from the preliminary forward matching process is scanned backward to see if there is any node which is closer to node * and has the same tag as the JN (the one closer to node #). There are three cases for JN re-generation:

(1) Neither of two nodes in a closed loop is a joint node.

In a closed loop, if two identical nodes from the two sequences are not JN’s and they have a different tag from that of the closest JN’s on the right, then they are jointed as a JN. Therefore, the previous closed loop will be replaced by two smaller closed loops. This is to solve the case in which two identical nodes have not been jointed in the preliminary forward matching. An example is showed in

Figure 4.4.

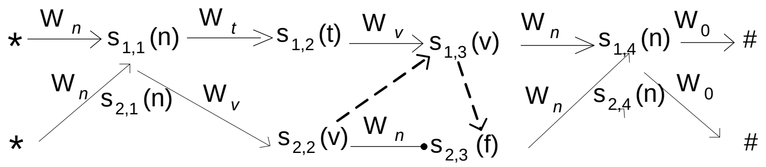


Figure 4.4: An example diagram after performing the backward matching refinement.

In Fig. 4.4, $s_{1,3}$ and $s_{2,2}$ are two identical nodes in the closed loop which are different from the right closest node $s_{1,4}$($s_{2,4}$), after performing the backward matching refinement, they are jointed and the directed link between node $s_{2,2}$ and node $s_{2,3}$ is removed. The dashed links are new directed links and the previous closed loop is divided into two smaller closed loops. Such adjustment will improve the similarity measure of sentence structures.

(2) In a closed loop, the node in the shorter sequence to be re-jointed is not a JN and the node in the longer sequence to be re-jointed is a JN.

In the backward matching refinement process, the nodes in the shorter sequence and the longer sequence are revisited. If there are two identical nodes from two sequences, where the one from the shorter sequence is not a JN and the one from the longer sequence is a JN, then they are re-jointed to improve the similarity measure of sentence structures. An example is showed in Fig.4.5 and Fig. 4.6. In Fig. 4.5, nodes $s_{1,4}$ and $s_{2,2}$ are identical and in the same closed loop, after the backward matching refinement, they are jointed with the new directed dashed links shown in Fig. 4.6. The original directed link between node $s_{2,2}$ and node $s_{2,3}$ is removed.

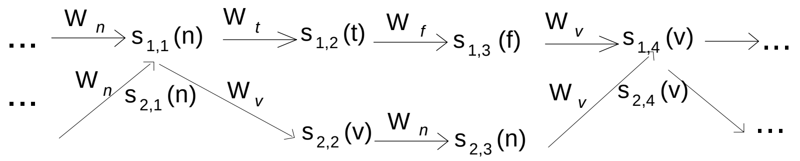


Figure 4.5: An example diagram after the preliminary forward matching process.

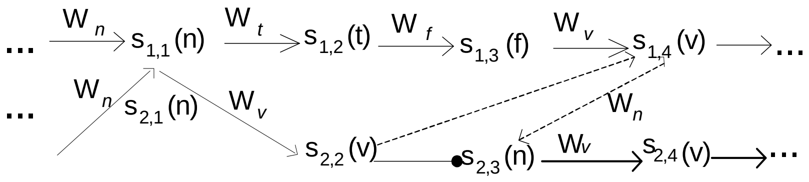


Figure 4.6: The example diagram of Fig.4.5 after the backward matching refinement.

(3) Two identical nodes out of any closed loop

If two nodes out of any closed loop are identical, then they are re-jointed to form a new JN. Figure 4.7 shows an example, where the dashed link is added to joint node $s_{1,5}$ and node $s_{2,4}$ which are identical.

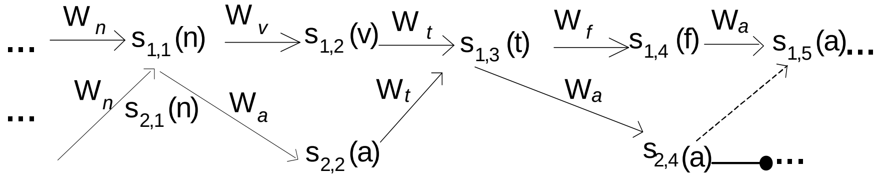


Figure 4.7: An example diagram after the backward matching refinement.

Based on our proposed joint directed diagram, we can define a structural similarity measure as

$$Sim_2(v_1,v_2,...v_q)=\frac{2\sum_{\mathrm{i}\in\mathrm{C}}\frac{1}{1+\sum_{\mathrm{j}\in\mathrm{D_c}}w_j(v_1,v_2,...v_q)}w_i(v_1,v_2,...v_q)}{\sum_{\mathrm{k}\in\mathrm{E}}w_k(v_1,v_2,...v_q)} \tag{4.4}$$

Where *C* is the set containing all joint nodes excluding nodes “*” and “#” in two sequences, D_c is the set containing all the nodes in a closed loop excluding the JN’s**.** From Eq. (4.4), we can see that if size of D_c is smaller, the similarity value is larger. When we compute the D_c term, we consider the node “*” as the start node of the first closed loop. *E* is the set containing all the nodes in two sequences. w_i is the weighting factor for node *i*. $v_1 \dots v_p$ are parameters which are similar to those defined in Eq. (4.3) to weigh POS’s tags of two compared sentences. It is easy to show that the similarity measure take values within the range from 0 to 1.

An example is given below:

Source sentence 1: 鲁迅浙江绍兴人。(Lu3Xun4 Zhe4 Jiang1 Shao4 Xing1 Ren2.)

After word segmentation: 鲁迅/n 浙江/ns 绍兴/ns 人/n 。/w

Source sentence 2: 鲁迅浙江人。(Lu3 Xun4 Zhe4 Jiang1 Ren2.)

After word segmentation: 鲁迅/n 浙江/ns 人/n 。/w

The corresponding POS tag sequences are as follows:

S_1: n ns ns n w.

S_2: n ns n w.

The directed graphs are shown in Fig. 4.8.

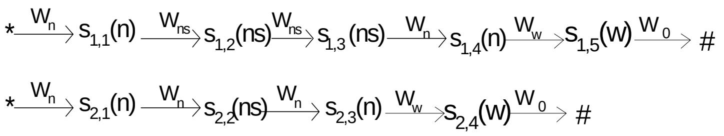


Figure 4.8: Directed graphs for sentences S_1 and S_2 .

The result after the preliminary forward matching process is shown in Fig. 4.9.

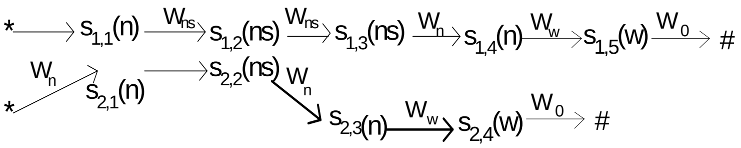


Figure 4.9: The result after the preliminary forward matching process.

After the backward matching refinement process, the directed diagram is shown in Fig. 4.10.

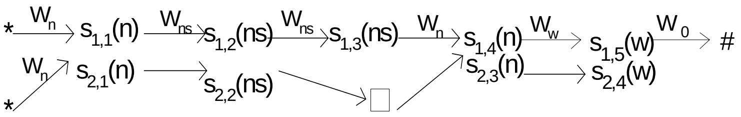


Figure 4.10: The directed diagram after the backward matching refinement process. ("□" denotes a null node).

We know that C = {$s_{1,1}$, $s_{1,2}$, $s_{1,4}$, $s_{1,5}$}, $D_1 = D_2 = D_4 = \Phi$ (null), $D_3 = s_{1,3}$, E contains all the nodes in two sequences. Let us assign a weight factor of 3 to a noun POS, 5 to a verb POS and 1 to all other POS's. By using to the similarity measure defined in Eq. (4.4), we can obtain the similarity value of 0.929 for the two sentences' structures.

4.6 Combination of Two Sentence Similarity Measures

While both the word-sequence-matching-based method and the POS-tag-sequence-matching-based method show their efficiency and effectiveness, they also have some shortcomings. For the word-sequence-matching-based method, the main drawback is that no structural information is considered as it cannot measure the structural similarity of two sentences. For the POS-tag-sequence-matching-based (PTSMB) method, no word sequence information is considered. To have a more sophisticated similarity measure we can combine these two measures by using the weighted average.

Suppose two sentences s_1 and s_2 can be viewed from two independent perspectives, the word sequence information and the POS tag sequence information. The sentences' similarity can be computed separately from each perspective, just as we have introduced in the previous sections. For example, the similarity between two documents can be calculated by comparing the sets of words in the documents or by comparing their stylistic parameter values, such as average word length, average sentence length, average number of verbs per sentence, etc. We assume that the overall similarity of the two documents is a weighted average of their similarities computed from different perspectives (Lin, 1998). This is similar to the similarity measure of Chinese sentences. The overall similarity of two Chinese sentences is a weighted average of their similarities computed by the word-sequence-matching-based measure and the POS-tag-sequence-matching-based measure.

In this chapter we measure the similarity of two Chinese sentences (s_1 and s_2) by making use of both word sequence information and POS tag sequence information. The weighted average of two measures is defined as

$$Sim(s_1,s_2) = t_1 Sim_1(s_1,s_2) + t_2 Sim_2(s_1,s_2) \quad (4.5)$$

$$\text{Subject to: } t_1 + t_2 = 1 \quad (4.6)$$

In Eq. (4.5), t_1 is the weight for the similarity measure using the word-sequence-matching-based method (Sim_1), and t_2 is the weight for the similarity measure using the POS-tag-matching-based method (Sim_2). The summation of weights t_1 and t_2 should be equal to 1. The values of t_1 and t_2 can be determined by experiments. To use Eq. (4.5) to measure the similarity of two Chinese sentences, the weight assignment of two similarity measures and other parameters used in each measure will be a key problem. In Chapter 5, we will present a relevance feedback scheme and a neural network model to optimize the model parameters to accommodate users' preferences and intentions.

We perform the comparisons between the proposed combination method and the Dice Coefficient measure. The partial experiment results are listed in the Table 4.8.

Table 4.8: Retrieved sentences with Dice Coefficient method and proposed WSMB+PTSMB method

		Retrieved Sentences	
N1	N2	WSMB+PTSMB	Dice Coefficient
1	1	油价回落带动美股上升，(As the easing oil prices led the US stocks up,)	油价回落带动美股上升，(As the easing oil prices led the US stocks up,)
	2	油价回吐令股价下跌， (Stock price fell on retreated oil prices,)	油价回落，(Oil prices retreated,)
	3	油价回落，(Oil prices retreated,)	油价大幅回落，(Oil prices retreated greatly,)
	4	油价续升拖累上周五晚美股大跌， (The surging oil prices dragged the US stocks to slump last Friday night,)	但油价回落，(But oil prices retreated,)
	5	油价持续创新高， (Continually renewed highs in oil prices,)	加上美股昨晚显著上升，(And the significantly rising US stocks last night,)
2	1	蓝筹股表现偏软， (Blue Chips performed sluggish.)	蓝筹股表现偏软， (Blue Chips performed sluggish.)
	2	蓝筹股表现牛皮， (Blue Chips performed sluggish.)	蓝筹股普遍偏软， (Blue Chips were sluggish broadly,)
	3	蓝筹股普遍偏软， (Blue Chips were sluggish broadly,)	股价偏软，(Stock prices were sluggish,)
	4	股价造好， (Stock prices performed well,)	蓝筹股表现牛皮， (Blue Chips performed sluggish.)
	5	股价下跌，(Stock prices slump,)	中资股亦表现偏软，(H-shares were sluggish,)
3	1	上海复地(2337-HK)跌 3.125%，(FORTE (2337-HK) down 3.125%)	上海复地(2337-HK)跌 3.125%，(FORTE (2337-HK) down 3.125%)
	2	上海石化(0338-HK)跌 1.71%。(Shanghai Pechem (0338-HK) sunk 1.71 per cent.)	只上海复地 (2337-HK)无起跌。 (Except that FORTE (2337-HK) remained unchanged.)
	3	泰山石化(1192-HK)升 8.3%； (Titan Petrochem (1192-HK) up 8.3 per cent;)	新地(0016-HK)跌 0.33%， (SHK PPT (0016-HK)was down 0.33 per cent,)
	4	泰山石化 (1192-HK)跌 7.69%， (Titan Petrochem (1192-HK) down 7.69 per cent,)	恒地(0012-HK)跌 0.79%， (Henderson Land (0012-HK) down 0.79 per cent,)
	5	中国人寿 (2628-HK)升 3%， (China Life (2628-HK) up 3 per cent,)	新地 (0016-HK)无起跌， (SHK PPT (0016-HK) remained unchanged,)
4	1	在 12811 至 12868 间窄幅争持， (Struggled within a narrow range between 12811 and 12868,)	在 12811 至 12868 间窄幅争持， (Struggled within a narrow range between 12811 and 12868,)

	2	在 13269 至 13335 间窄幅徘徊，(Struggled within a narrow range between 13269 and 13335,)	在 13269 至 13335 间窄幅徘徊，(Struggled within a narrow range between 13269 and 13335,)
	3	期指整天持续在 12819 至 12880 间窄幅徘徊， (The futures still struggled within a narrow range between 12819 and 12880,)	以 13027 低开后在 13025 至 13099 间窄幅争持，(Struggled within a narrow range between 13025 and 13099 after a lower opening at 13027,)
	4	以 13027 低开后在 13025 至 13099 间窄幅争持， (Struggled within a narrow range between 13025 and 13099 after a lower opening at 13027,)	然后窄幅争持，(Then struggled within a narrow range in the afternoon,)
	5	下午期指持续窄幅争持，(The futures continually struggled within a narrow range in the afternoon,)	下午窄幅争持，(Struggled within a narrow range in the afternoon,)
5	1	复牌后大升 113%；(Soared 113 per cent after resumption of trading;)	复牌后大升 113%；(Soared 113 per cent after resumption of trading;)
	2	大升 8.76%，(Up greatly 8.76%,)	大升 21%后停牌；(Trading was suspended following a 21 per cent surge;)
	3	复牌后跌 6.5%，(Dropped 6.5 per cent after resumption of trading,)	上午大升 124%后突然停牌；(Trading was suspended suddenly following a 124 per cent surge;)
	4	复牌后显著挫 9%。(Dropped greatly 9 per cent after resumption of trading.)	复牌后股价仍大挫近 60%。(Dropped greatly 60 per cent after resumption of trading.)
	5	大升 274 点，(Up greatly 274 pts,)	但昨天大升后，(But after yesterday’s surge,)

Table 4.9: Source sentence used in Table 4.8

No.	Source Sentences
1	油价回落带动美股上升，(As the easing oil prices led the US stocks up,)
2	蓝筹股表现偏软，(Blue Chips were sluggish.)
3	上海复地(2337-HK)跌 3.125%，(FORTE (2337-HK) down 3.125%)
4	在 12811 至 12868 间窄幅争持， (Struggled within a narrow range between 12811 and 12868,)
5	复牌后大升 113%；(Soared 113 per cent after resumption of trading;)

In Table 4.8, the first column (*N1*) is the serial numbers of source sentences which are also listed in Table 4.9. The second column (*N2*) is the serial number of the retrieved top five sentences for each source sentence. The retrieved sentences with the combination measure of the word-sequence-matching-based method and the POS-tag-sequence-matching-based method (WSMB+PTSMB) are listed in the third column, and the ones with the Dice Coefficient are listed in the fourth column. In Table 4.8, the first retrieved sentence is the one which is identical to the source sentence. That indicates if there is an identical sentence in database, it can be retrieved. Let us take source sentence 1 as an example, from the two groups of retrieved top five sentences, we can see that the result with combined method is more reasonable according to our perception. Because the retrieved sentence 2 (油价回吐令股价下跌，) in third column is more similar to the source sentence (油价回落带动美股上升，) than the third one (油价回落，). But if the Dice coefficient measure is used, the second retrieved sentence (油价回落,) and the third one (油价大幅回落,) are less similar to the source sentence 1 than the second retrieved one in column 3 does. If we consider source sentence 4 to be input, from the table, we also can know the retrieved result using combined method is superior to the one using Dice Coefficient measure. Because the retrieved sentences 3 in column 3 is more similar to the source sentence 4 than the fourth retrieved sentence which is the third one if the Dice Coefficient measure is used. Similarly, for other source sentences, the retrieved results using the proposed combined method are generally more reasonable according to human perception. It shows that the combined method has better performance than the Dice Coefficient measure on the sentence database.

4.7 Conclusion

In this chapter, we present a word sequence based method and a POS tag sequence based method for The similarity measure of Chinese sentences (SMCS). Word sequence and corresponding POS tag sequence are important information of a Chinese sentence. A robust similarity measure of Chinese sentences should make use of these two kinds of information. For alignment in both measures, different word orders will produce different corresponding POS sequences which will generate different alignment results and similarity values. In both measures, we weigh the different POS's with different factors in the compared sentences to reflect their importance in sentence matching. A combination of two measures by a weighted average has also been proposed in this Chapter. The aim of this chapter is to bring out the theoretical treatment and analysis of two measures and their combination. Results of experiments on the combined measure and error analysis will be given in Chapter 5.

Chapter 5

Parameter Optimization of Similarity Measures with a Users' Relevance Feedback Scheme

5.1 Introduction

In engineering domains, for example in information and image retrieval domains, the relevance feedback scheme has been commonly adopted to learn users' preference and intention.

In this chapter, we present a human-computer interaction approach to optimize the combined similarity measure of Chinese sentences (SMCS) discussed in Chapter 4, based on a relevance feedback scheme.

In a computer centric approach the weight factors of parts of speech (POS's) are fixed, which cannot effectively model high-level semantic concepts and human's perception. Furthermore, the weight assignment imposes a huge burden as it requires comprehensive knowledge of the low-level feature representation of sentences.

To address the difficulties faced by a computer centric approach, we present a relevance feedback based approach to the similarity measure of Chinese sentences in which human and computer interact. Relevance feedback is a powerful technique used in many different domains. For example, in the image processing domain it has been applied to content-based image retrieval (Celentano & Sciasicio, 1998; Doulamis, et al., 2000; Picard, et al., 1996; Rui, et al., 1997; Rui, et al., 1998; Rui & Huang, 1999; Vasconcelos & Lippman, 2000; Wang, et al., 2003; Wu, et al., 2000; Yoon & Jayant, 2001). In optimizing similarity measure, the relevance feedback process is to obtain the user's preference and intention in ranking the sentence similarity and to optimize the weight assignment in the similarity measure by using the users' feedback information. Under an assumption that high-level concepts can be captured and mapped to the low-level features, the relevance feedback technique tries to establish the link between high-level concepts and low-level features. To our best knowledge, application of a user's relevance feedback scheme to the similarity measure of Chinese sentences is pioneered by us.

There are two distinct characteristics of Chinese sentences that are similar to images: (1) the gap between high-level semantic concepts and low-level features and (2) the subjectivity of human' understanding sentences and perceiving the similarity between two sentences.

5.2 Corpus Construction

We used a corpus of daily stock reports (a bilingual sentence database) to perform parameter optimization. For the current study, we only need to process Chinese sentences with word segmentation and tagging performed by making use of the tools we purchased from the Institute of Computational Linguistics of Peking University, China. The following are the reasons why we choose daily stock reports to construct the corpus for parameter optimization of the similarity measure of Chinese sentences:

(1) It is easy to collect these bilingual daily stock reports in Hong Kong.

(2) The article style and sentence structures in this domain are relatively stable and well structured.

(3) Database construction can be completed with a reasonable period of time and with modest effort.

(4) The reports are provided daily, so we can gradually expand the corpus.

We constructed a training set which consists of fifty source Chinese sentences chosen from the database. For every sentence in the set, we obtained its top 10 similar sentences from the database. In total, 550 sentences were used for parameter optimization. By using such a small training set we intend to make a preliminary study on the feasibility of our optimization method. To make a system useful in practice we need to increase the size of training sentences significantly. We understand that a huge amount of time needs to be spent by each user to provide the feedback data if the training set is large. Included in the source sentence set are simple Chinese sentences that cover almost all the POS's.

A complex or compound sentence can be decomposed into shorter sub-sentences. If we can solve the problems of simple sentences satisfactorily, complex sentences can also be processed without significant difficulties. This is because in Chinese, phase and simple sentence have identical grammar structures. Moreover, most complex or compound sentences consist of simple sentences that can be processed more easily. So processing simple sentences is the fundamental task of processing a complex sentence. There are 2,100 sentences in the corpus.

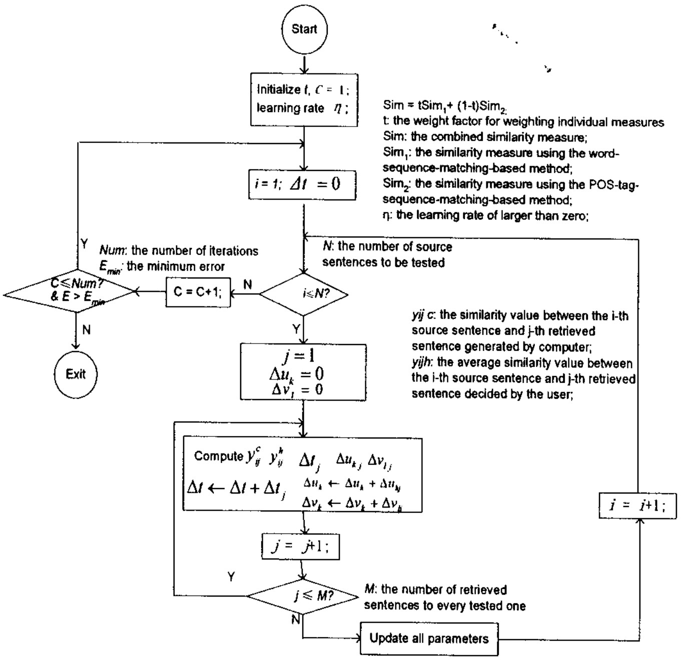
5.3 Design of the Relevance Feedback Network Scheme

5.3.1 Main Idea

For every group of sentences, there is one source sentence and ten compared sentences in a descending order of similarity values to the source sentence. A user is required to re-rank the sentences based on his/her judgment if necessary. This task was performed by web-based questionnaires which were designed for obtaining human feedback data. The new ranking of sentences is transferred into a set of numerals to reflect the similarity of the sentences to the source sentence. Twenty research students fluent in Chinese language were asked to complete the questionnaires. The data collected is used to train a neural network to optimize the parameters in the similarity measure of Chinese sentences.

In collecting feedback data, we cannot ignore one problem, that is, different persons or the same person under different circumstances may perceive the same pair of sentences differently. This is called *human perception subjectivity*. The subjectivity exists at various levels. For example, one person may be more interested in a sentence's structure while another may be more interested in the word sequence information. Even if two persons are both interested in the structure, the way in which they perceive the similarity of two sentences may be quite different (Rui, et al., 1998). The training process is to optimize the parameter set by considering the overall responses of different users.

The block diagram of the relevance feedback scheme on parameter optimization is showed in Fig.5.1.



Start
Initialize t, $C = 1$;
learning rate η;
Sim = tSim$_1$+ (1-t)Sim$_2$
t: the weight factor for weighting individual measures
Sim: the combined similarity measure;
Sim$_1$: the similarity measure using the word-sequence-matching-based method;
Sim$_2$: the similarity measure using the POS-tag-sequence-matching-based method;
η: the learning rate of larger than zero;
$i = 1$; $\Delta t = 0$
Y
Num: the number of iterations
E_{min}: the minimum error
N: the number of source sentences to be tested
$C \leq Num$? & $E > E_{min}$
C = C+1;
N
$i \leq N$?
N
Exit
Y
yij c: the similarity value between the i-th source sentence and j-th retrieved sentence generated by computer;
yijh: the average similarity value between the i-th source sentence and j-th retrieved sentence decided by the user;
$j = 1$
$\Delta u_k = 0$
$\Delta v_l = 0$
Compute y_{ij}^c y_{ij}^h Δt_j Δu_{kj} Δv_{lj}
$\Delta t \leftarrow \Delta t + \Delta t_j$
$\Delta u_k \leftarrow \Delta u_k + \Delta u_{kj}$
$\Delta v_k \leftarrow \Delta v_k + \Delta v_{k}$
$i = i+1$;
$j = j+1$;
Y
$j \leq M$?
M: the number of retrieved sentences to every tested one
N
Update all parameters

Figure 5.1: Block diagram of the relevance feedback scheme for parameter optimization.

The parameter optimization should address issues including weight initialization, training procedure, and stopping criteria.

(1) Weight initialization

We initialize the weights $W = [t, u, v]$ with t = 0.5, and the weight sets of *u* and *v* are set corresponding to the POS's set (see Table 5.1, the column of Original Parameter Value). We also set the learning rate η =0.5.

(2) Training

There are two training modes available, on-line and off-line modes. In the on-line mode, we update the parameters immediately after every group of sentences are ranked by respondents. Then we use new parameters to retrieve the top 10 similar sentences for the next source sentence. In the off-line mode, we update the parameters after all groups of sentences are ranked by respondents. The training is carried out after all users complete the questionnaires. The on-line training mode is very time consuming for a user to complete the questionnaire because the new group sentences should be re-computed by the updated parameters. In our study, we used the off-line training mode.

(3) Training Stopping Conditions

If the number of iterations reaches the maximum number or the total mean square error is not larger than the preset minimum error requirement, the training terminates.

Whether its weights can be updated continuously distinguishes the proposed relevance feedback approach from the computer centric approach. In the computer centric method, its weights are fixed which cannot effectively model high-level concepts and human perception subjectivities because of the fixed weights.

5.3.2 The Training Algorithm

As mentioned above, in our questionnaire system there are fifty source sentences. Top 10 similar sentences with a source sentence form a group. The top 10 retrieved sentences will be listed under its source sentence in the descending order of similarity values calculated by using Eq. (4.5) in Chapter 4. Then, respondents are required to give a new ranking of similarity by giving an integer number ranging between 1 (the most similar one) and 10 (the most dissimilar one) to every sentence in the list. After the normalization, the value 1.0 is for the most similar sentence, 0.9 to the second similar one, and 0.1 to the least similar sentence in the list. All assigned values corresponding to users' feedback are considered as the desired values. The difference between the similarity value y_{ij}^{c} calculated by a computer and the assigned value y_{ij}^{h} corresponding to a user's feedback can be computed. Based on the difference, the modification of parameters can be done using a gradient based algorithm discussed next.

The next task we should focus on is to train a neural network to obtain optimal parameters, which is the optimal assignment of weights of POS's tags and the important factors of the word-sequence-matching-based method, and the POS-tag-sequence-matching-based method. After all users' feedback information is used to train the network, a final parameter set which collectively accommodates the intentions and preferences of a number of respondents will be determined. Such a parameter set is expected to work as well as human beings in the similarity measure of Chinese sentences if a large training set is available. The network diagram to be used for parameter optimization is shown in Fig. 5.2.

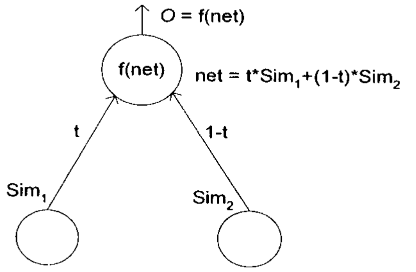


Figure 5.2: The architecture of the neural network for the parameter optimization of the combined similarity measure of Chinese sentences.

In the parameter optimization scheme, a sigmoid function defined below is applied to the output node to achieve a non-linear mapping:

$$o = f(net) = \frac{1}{1+e^{-net}} \qquad (5.1)$$

The net input the output node is defined as

$$net_{ij} = t * Sim_1^{ij} + (1-t) * Sim_2^{ij} \qquad (5.2)$$

In Eq. (5.2), i represents the i-th source sentence and j is the j-th compared sentence. Parameters t and (1-t) reflect the different emphasis of two measures in the overall similarity. Sim1 is the word-sequence-matching-based similarity measure (Eq. (4.3) and Sim2 is the POS-tag-sequence-matching-based (PTSMB) similarity measure (Eq.(4.4)).

The output of the network, y_{ij}^c, is defined as

$$y_{ij}^c = f(net_{ij}) = f[tSim_1(u_{ij}^1,...u_{ij}^p) + (1-t)Sim_2(v_{ij}^1,...v_{ij}^q)]$$

where $u_{ij}^{p}\ (i=1,2,\cdots p)$ are the parameters for the word-sequence-matching-based measure and $v_{ij}^{q}\ (i=1,2,\cdots q)$ are the parameters for the POS-tag-sequence based similarity measure.

The error function to be minimized is defined as

$$E=\frac{1}{2}(y_{ij}^{h}-y_{ij}^{c})^{2}$$

where y_{ij}^{h} is the similarity measure between the *i*-th source sentence and the *j*-th retrieved sentence converted from the ranking of a user. The modification of parameters, including t, u and v, can be performed by the gradient learning method given by

$$\Delta t_{ij}=-\eta\frac{\partial E}{\partial t_{ij}}=\eta(y_{ij}^{h}-y_{ij}^{c})f'(net_{ij})(Sim_{1}-Sim_{2})$$

$$\Delta u_{ij}^{k}=-\eta\frac{\partial E}{\partial u_{ij}^{k}}=\eta t(y_{ij}^{h}-y_{ij}^{c})f'(net_{ij})\frac{\partial Sim_{1}}{\partial u_{ij}^{k}}$$

$$\Delta v_{ij}^{l}=-\eta\frac{\partial E}{\partial v_{ij}^{l}}=\eta(1-t)(y_{ij}^{h}-y_{ij}^{c})f'(net_{ij})\frac{\partial Sim_{2}}{\partial v_{ij}^{l}}$$

Where η is a learning rate. The derivative of Eq. (5.1) is given by

$$f'(net)=f(1-f) \qquad (5.3)$$

To obtain the gradient, we make the following derivation.

$$\frac{\partial Sim_1}{\partial u_p} = \frac{(2\sum_{i=1}^{n}(i(1-\delta_1(C_i))*N_i^p))*[(m+n)*\sum_{k=1}^{n}w_k] - (2\sum_{i=1}^{n} i*(\sum_{j=1}^{c_i}\sum_{k=1}^{i} w_k^j))*(m+n)*N^p}{[(m+n)*\sum_{k=1}^{n}w_k]^2}$$

(5.4)

$$\delta_1(C_i) = \begin{cases} 1 & C_i = 0 \\ 0 & C_i \neq 0 \end{cases}$$

In Eq. (5.4), N^p is the number of POS tag *p* (with the weight u_p) appeared in the shorter sentence. N_i^p is the number of POS tag *p* in the identical word sequences with length *i*.

$$\frac{\partial Sim_2}{\partial v_q} = \frac{[2\sum_{i=1}^{C}\frac{\delta_2(POS(i),v_q)*(1+\sum_{j=1}^{D_i}w_j^c) - w_i * N_i^q}{(1+\sum_{j=1}^{D_i}w_j^c)^2}]*(\sum_{k=1}^{E}w_k^a) - [2\sum_{i=1}^{C}\frac{w_i}{1+\sum_{j=1}^{D_i}w_j^c}]*N^q}{(\sum_{k=1}^{E}w_k^a)^2}$$

(5.5)

$$\delta_2(POS(i),v_q) = \begin{cases} 1 & \text{if } POS(i) = v_q \\ 0 & \text{else} \end{cases}$$

In Eq. (5.5), N_i^q is the number of POS tag *q* (with the weight *v*$_q$) in the *i*-th closed loop (see the discussion in Section 4.5). N^q is the number of POS tag *q* in the two sentences. The identical POS tag in a sentence or a sequence is repeatedly counted if it appears more than once.

5.4 Experimental Results and Error Analysis

The ultimate goal of the relevance feedback technique is to help obtain the optimal parameter set that can be used in the similarity measure of Chinese sentences for achieving a human-like performance.

It is assumed that a user's intention is consistent when doing relevance feedback. That is, a user does not change his or her judgment during the feedback process. But for different users, it is obvious that they have different judgments.

There are mainly five factors that affect the behavior of the algorithm, i.e., the grammar information contained in sentences and POS distribution of the collected sentence set, the feedback data set, the number of iterations, and the initial values of parameters. The performance of the system is expected to improve gradually as more feedback iterations are carried out. Collecting feedback data is very time-consuming (one and a half hours for a user to complete the questionnaire). In our experiment, we managed to collect the feedback data from 20 research students who have very good knowledge of Chinese language.

In our experiments, we assign an initial weight value of 3 to nouns, 5 to verbs and 1 to all other POS's purely based on our common sense. However, it is undoubtedly true that in some sentences this setting might cause problem because a noun may play more important role than a verb. It also should be mentioned that the initial weight values are of a subjective nature. The weight values will tend to accord with the real situation after a certain number of training iterations. After parameter optimization is carried out, we can obtain new parameters shown in column 2 in Table 5.1.

Table 5.1: Parameters before and after the training process

Para. No.	New Para. Value	Original Para. Value	POS	Remarks
1	3.366560	3	n	Nouns. Increased importance.
2	3.023259	3	nr	Name noun. Little change.
3	3.087817	3	ns	Space noun. Little change.
4	3.000000	3	nt	Organization noun. No change since there is no such sample in the training set.
5	2.831603	3	nx	Not a Chinese character string (非汉字串). Decreased importance.
6	3.031262	3	nz	Other proper noun. Little change.
7	3.000000	3	an	Noun adjective (名形词). No change since there is no such sample in the training set.
8	3.004381	3	vn	Noun verb (名动词). Little change.
9	1.202004	1	t	Time. Increased importance.
10	1.053833	1	s	Space. Little change.
11	1.286397	1	f	Orientation (方位词). Increased importance.
12	1.110609	1	q	Quantity. Increased importance.
13	1.089220	1	b	Distinguish words (区别词). Little change.
14	1.763441	1	a	Adjective. Increasing importance significantly.
15	1.145355	1	ad	Adverb adjective (副形词). Increased importance.
16	1.000000	1	z	State word. No change since there is no such sample in the training set.

17	4.560758	5	v	Verb. Still very important but not as important as we expected.
18	1.000000	1	vd	Adverb verb (副动词). No change since there is no such sample in the training set.
19	1.155311	1	m	Numeral. Increased importance.
20	1.108447	1	r	Pronoun. Increased importance.
21	1.002257	1	y	Modal particle. Little change.
22	1.000000	1	o	Onomatopoeia. No change since there is no such sample in the training set.
23	1.000000	1	e	Exclamation. No change since there is no such sample in the training set.
24	1.127111	1	u	Auxiliary. Increased importance.
25	1.139976	1	j	Abbreviation. Increased importance.
26	1.000000	1	l	Idiom temporarily (临时性习语). No change since there is no such sample in the training set.
27	1.000000	1	h	Constituent followed by others(前接成分). No change since there is no such sample in the training set.
28	0.660136	1	w	Punctuation. Decreasing importance significantly.
29	1.006418	1	k	Constituent following others (后接成分). Little change.
30	1.000000	1	g	Morpheme. No change since there is no such sample in the training set.
31	1.119985	1	Dg	Adverb morpheme (副词性语素). Increased importance.
32	1.000000	1	Tg	Time morpheme (时间词性语素). No change since there is no such sample in the training set.
33	1.000000	1	x	Not morpheme (非语素字). No change since there is no such sample in the training set.
34	1.032604	1	p	Preposition. Little change.
35	1.503441	1	d	Adverb. Increasing importance significantly.

36	1.189667	1	c	Conjunction. Increased importance.
37	1.000000	1	i	Idiom. No change since there is no such sample in the training set.
38	1.148404	1	Ng	Noun morpheme (名词性语素). Increased importance.
39	1.014666	1	Ag	Adjective morpheme (形容词性语素). Little change.
40	1.040884	1	Vg	Verb morpheme (动词性语素). Little change.
41	0.512938	0.5	t’	Weighting for word-sequence-matching based measure, increased importance. (1-0.512938) = 0.487062 for tag-sequence-matching based measure, decreased importance.

From Table 5.1, we know that the modification of parameters accord with human’s perceptions and our expectation. In the table, the weight of nouns increased indicates that it plays an important role in SMCS. Similarly, the weight of punctuation decreased significantly shows that it is not that important in SMCS. It is also observed that some of patterns never appear, e.g., “an”. This is due to the problem of limited data. Unless we have a huge database, we may encounter a similar problem whereby some patterns are omitted. Due to the difficulties in constructing a large database, here we focus on the method itself. From the experimental result obtained, we observe that the approach can address all patterns appeared in the training set.

After the parameters are optimized, we retrieve top 10 similar sentences for every input source sentence. To certify how the optimized parameters work, a comparison between the top retrieved sentences for the same source sentence

(Table 5.2) by using the parameters before and after optimization is given in Table 5.3.

Table 5.2: Source sentences used for the experiment

No.	Source Sentences
1	买盘推动纽约 11 月期油再创新高， (The New York November oil futures was driven by buying to hit another fresh high.)
2	地产股及工商股全线向上， (Property and industry & commerce stocks all advanced.)
3	北京燕化(0325-HK)、镇海炼油(1128-HK)分别跌 2.31%及 1.27%， (Beijing Yanhua (0325-HK) and Zhenhai Refine (1128-HK) lost 2.31 per cent and 1.27 per cent, respectively.)
4	下午期指持续窄幅争持， (HSI future continually struggled within a narrow range in the afternoon.)
5	其他原料股及能源股则窄幅波动。 (Other raw material and energy stocks fluctuated within a narrow range.)

Table 5.3: Retrieved sentences with the weight sets before and after parameter optimization (N1: the source sentence number; N2: the rank of a retrieved sentence)

		Retrieved Sentences	
N1	N2	**The optimized weight set used**	**The original weight set used**
1	1	买盘推动纽约 11 月期油再创新高， (The New York November oil futures was driven by buying to hit another fresh high.)	买盘推动纽约 11 月期油再创新高， (The New York November oil futures was driven by buying to hit another fresh high.)

	2	刺激纽约 11 月期油再创新高， (Stimulating the New York November oil futures to rise to a fresh high.)	带动纳指大幅上升， (The Nasdaq was simulated.)
	3	带动纳指大幅上升， (The Nasdaq was simulated.)	比估计试 13069/13048 还差， (Worse than the predicted testing 13069/13048.)
	4	比估计试 13069/13048 还差， (Worse than the predicted testing 13069/13048.)	刺激纽约 11 月期油再创新高， (Stimulating the New York November oil futures to rise to a fresh high.)
	5	有报导指内地或调升电价， (On report that power rate might be raised in the Mainland.)	有报导指内地或调升电价， (On report that power rate might be raised in the Mainland)
2	1	地产股及工商股全线向上， (Property and industry & commerce stocks all advanced.)	地产股及工商股全线向上， (Property and industry & commerce stocks all advanced.)
	2	地产股及工商股则表现靠稳。 (Property and industry & commerce stocks held steady.)	地产股及工商股则表现靠稳。 (Property and industry & commerce stocks held steady.)
	3	原料股、航运股全线下跌； (Raw material and shipping stocks sunk across the board.)	原料股、航运股全线下跌； (Raw material and shipping stocks sunk across the board.)
	4	半导体类股份下跌， (Drops in semiconductor stocks.)	半导体类股份下跌， (Drops in semiconductor stocks.)
	5	地产股普遍下跌， (Property stocks declined broadly.)	化工股、保险股个别发展； (Chemical and insurance stocks developed individually.)
3	1	北京燕化(0325-HK)、镇海炼油(1128-HK)分别跌 2.31%及 1.27%， (Beijing Yanhua (0325-HK) and Zhenhai Refine (1128-HK) lost 2.31 per cent and 1.27 per cent respectively.)	北京燕化(0325-HK)、镇海炼油(1128-HK)分别跌 2.31%及 1.27%， (Beijing Yanhua (0325-HK) and Zhenhai Refine (1128-HK) lost 2.31 per cent and 1.27 per cent respectively.)

	2	上海石化(0338-HK)、镇海炼油(1128-HK)分别跌4.1%及3.09%， (Shanghai Pechem (0338-HK) and Zhenhai Refine (1128-HK) gave up 4.1 percent and 3.09 percent respectively.)	不过北京燕化(0325-HK)、镇海炼油(1128-HK)则先升后跌； (But Beijing Yanhua (0325-HK) and Zhenhai Refine (1128-HK) declined after rises.)
	3	不过北京燕化(0325-HK)、镇海炼油(1128-HK)则先升后跌； (But Beijing Yanhua (0325-HK) and Zhenhai Refine (1128-HK) declined after rises.)	上海石化(0338-HK)、镇海炼油(1128-HK)分别跌4.1%及3.09%， (Shanghai Pechem (0338-HK) and Zhenhai Refine (1128-HK) gave up 4.1 percent and 3.09 percent respectively.)
	4	恒地(0012-HK)、新地(0016-HK)分别微升0.28%及0.36%； (Henderson Land (0012-HK) and SHK PPT (0016-HK) edged up 0.28 per cent and 0.36 per cent respectively.)	恒地(0012-HK)、新地(0016-HK)分别微升0.28%及0.36%； (Henderson Land (0012-HK) and SHK PPT (0016-HK) edged up 0.28 per cent and 0.36 per cent respectively.)
	5	长实(0001-HK)、和黄(0013-HK)分别跌0.37%及0.82%； (Cheung Kong (0001-HK) and Hutchinson Whampoa (0013-HK) dipped 0.37 per cent and 0.82 per cent respectively.)	长实(0001-HK)、和黄(0013-HK)分别跌0.37%及0.82%； (Cheung Kong (0001-HK) and Hutchinson Whampoa (0013-HK) dipped 0.37 per cent and 0.82 per cent respectively.)
4	1	下午期指持续窄幅争持， (Continually struggled within a narrow range in the afternoon.)	下午期指持续窄幅争持， (Continually struggled within a narrow range in the afternoon.)
	2	下午期指持续窄幅上落， (The futures waved within a narrow range in the afternoon.)	下午期指持续窄幅上落， (The futures waved within a narrow range in the afternoon.)
	3	下午期指回软， (The futures retreated in the afternoon.)	下午期指回软， (The futures retreated in the afternoon.)
	4	目前期指下调低见13155， (The HSI lost ground at 13155,)	下午期指徘徊靠稳， (The futures hovered towards firmness.)

	5	下午期指徘徊靠稳， (The futures hovered towards firmness.)	今天期指结算， (The futures will be settled today.)
5	1	其他原料股及能源股则窄幅波动。 (Other raw material and energy stocks fluctuated within a narrow range.)	其他原料股及能源股则窄幅波动。 (Other raw material and energy stocks fluctuated within a narrow range.)
	2	其他原料股及航运股同告下挫， (Other raw material and shipping stocks also lost.)	其他原料股及航运股同告下挫， (Other raw material and shipping stocks also lost.)
	3	其他能源股及钢铁股则普遍下跌。 (While other energy and steel stocks declined broadly.)	其他能源股及钢铁股则普遍下跌。 (While other energy and steel stocks declined broadly.)
	4	其他公路股及钢铁股亦靠稳， (Other expressway and steel stocks also held stable.)	以原料股及航运跌幅较大， (H-shares sunk with raw material and shipping stocks more abruptly.)
	5	以原料股及航运跌幅较大， (H-shares sunk with raw material and shipping stocks more abruptly.)	其他公路股及钢铁股亦靠稳， (Other expressway and steel stocks also held stable.)

In Table 5.3, the first retrieved sentence is actually the source sentence, indicating that the identical sentence is always retrieved first if it is in the database. Let us take source sentence 1 as an example, from the two groups of top five retrieved sentences, we can see that the result with the optimized parameters is more consistent with human’s perception. As it is expected, the retrieved sentence 2 (刺激纽约 11 月期油再创新高，) in third column is more similar to the source sentence (买盘推动纽约 11 月期油再创新高，) than the third one (带动纳指大幅上升，) and the fourth one (比估计试 13069/13048 还

差，). But using the original weight set cannot produce the correct ranking. If we consider the source sentence 3 to be input, from the table we can see the retrieved sentences using the optimized parameter set is superior to the one using original weight set. It is also observed that some retrieved sentences are quite different from their source one. This problem can be solved if the sentences database is expanded. How to make use of the top retrieved sentences is another important problem in EBMT. Experimental results show that our relevance feedback scheme can accommodate human intension to a certain degree. The errors or incorrect retrievals occur mainly due to the following reasons:

(1) The feedback data set is not large enough. Collecting feedback data is very time-consuming (one and a half hours for a user to complete the questionnaires).

(2) The number of parameters used is not large enough. In the current experiment, the parameters mainly include the weights of POS tags and the weighting ratio between two measures.

(3) The incorrect word segmentation and tagging. The word segmentation and tagging tool does not have very high accuracy for the data set we used because some professional vocabulary has not been included in the dictionary of the tool.

Currently, we restrict our experiments to a small database to verify the efficiencies of the proposed methods due to the limited time frame. Our proposed methods for SMCS have obtained interesting and promising results by making use of the relevance feedback scheme. However, more improvements can be achieved if the following problems are addressed:

(1) A larger collection of the feedback data

(2) More parameters to be tuned by improving the modeling

(3) Improved word segmentation and tagging

There are some other difficulties in SMCS. As we know, some ambiguities are contained in both languages and further introduced in processing language resources, e.g., different definition of Chinese words from different points of view. Another difficulty is that sentence analysis at the semantic level is very difficult based on current theory and techniques. There is still no mature semantic analysis technique available. It is well understood that the semantic analysis of sentence is a long-term pursuit of natural language processing, which will greatly contribute to the implementation of computer based language understanding systems. Our proposed model based a relevance feedback scheme and neural network training is actually grammar-based but not semantic-based. The semantic information is represented partially by the grammar information, but no semantic information is considered specifically.

5.5 Conclusion

Similarity measure is an important and fundamental task in artificial intelligence and many other fields. The similarity measure of Chinese sentences (SMCS) as a very important component for Example-Based Machine Translation (EBMT) has emerged as one of the most active research focus for some time now.

In this chapter, we have presented a relevance feedback scheme and a neural network model to obtain the optimal parameter set to combine two similarity methods discussed in Chapter 4, namely the word-sequence-matching based method and the POS-tag-sequence-matching based method. Unlike a computer centric approach, where the weights are fixed, the proposed interactive approach allows a computer to update the weight set via a users' relevance feedback scheme. The effectiveness of the proposed approach has been validated by experiments, although the database used is relatively small due to difficulties in constructing a large database.

Some improvements can be expected from future research work. In fact, we cannot ignore the fact that the same Chinese word plays different roles in different sentences. And the same parts of speech in a sentence may also play different roles. This suggests that it is not ideal to assign the same weight to all words of the same POS. Weight factors for context information (group of words) may be considered to improve the situation. Another problem needs to be addressed in the future is to develop a relatively-objective approach to evaluate the performance of similarity measure.

Chapter 6

Conclusion

6.1 Summary of Achievements

In this thesis, advanced Chinese chunk segmentation and the similarity measure of Chinese sentences (SMCS) techniques are presented with potential application to example-based machine translation (EBMT).

For Chinese chunk segmentation, a statistical model combined with decision rules generated using an error-driven mechanism is presented. The parameters used in the statistical model are estimated from a corpus of sentences with manually segmented chunks. In this combined chunk segmentation method, firstly, preliminary chunk segmentation results are obtained by the statistical model. A set of decision rules is learned to refine the segmentation results. An error-driven mechanism is used in rule learning by comparing the chunk segmentation results from the statistical model and the manually segmented chunks. In Chapter 3, we show promising experimental results from this combined chunk segmentation method. The proposed Chinese chunk segmentation method can be applied to Chinese information processing, e.g., the similarity measure of Chinese sentences for EBMT.

The similarity measure of Chinese sentences plays a very important role in

Chinese information processing. To retrieve similar example(s) from a large sentence database, an efficient and effective Chinese sentence measure is required. In this thesis, similarity measures of Chinese sentences using word/chunk sequences and Part of Speech (POS) tag sequences are presented, which is discussed in Chapter 4. For Chinese sentences, there is no delimiter between Chinese words, which is different from English sentences. Therefore, Chinese word/chunk segmentation should be performed first before a similarity measure based on words/chunks can be carried out. In making use of POS tag sequences for similarity measure, we partially take into consideration structural information contained in the sentences by using directed graphs.

In the word-sequence-matching-based measure, we mainly consider the three factors which determine the similarity of two sentences. They are the number of identical word sequences, the length of identical word sequences, and the average POS weighting (AW) of the identical word sequences. We assign a weight to every Chinese POS according to the importance of the POS in Chinese sentences based on common sense. We have also made use of chunk segmentation results in the similarity measure. The ratio between the numbers of chunks with the same segmentation and different segmentation in every pair of identical word sequences is used in the measure for an improved performance.

In the POS-tag-sequence-based similarity (PTSMB) measure, the objective is to measure the similarity of Chinese sentences’ structures. If the constituents in two Chinese sentences are similar, then these two Chinese sentences are similar in structure. The main idea of this similarity measure is that we perform matching between POS's of two Chinese sentences to be compared by using directed graphs. Similar to word-sequence-matching-based method, the POS weighting is also utilized in this process. We used directed graph to model the POS sequence of a sentence where the tag of POS is represented using a node and a directed weighted link is used to connect two sequent nodes.

After the two similarity measures, one based on word/chunk sequences and the other based on POS tag sequences, are discussed we propose a novel Chinese sentence measure by combining these two measures. In the combined measure, the POS's weighting and the weighting of two measures are the parameters to be optimized. To obtain optimal parameters, we propose a relevance feedback scheme and a neural network model in Chapter 5. In our approach, the user's preferences and intensions are captured and used to optimize the model parameters for achieving human-like performance. We designed a web-based questionnaire to obtain human feedback data, which was used to train a neural network model to optimize the parameters. Experimental results have shown a visible improvement in measure accuracy. The proposed scheme on parameters optimization for the similarity measures of Chinese sentences has improved the measure accuracy.

A preliminary Chinese to English EBMT system was also implemented, which was included in Appendix A. In this trial EBMT system, after a Chinese sentence is input, word segmentation and tagging is performed with built-in software tools. The next step is to obtain the relevant examples by using index files. This is followed by the sentence similarity measure to retrieve top similar sentences and their translations. In implanting this trail EBMT system, issues such as sentence indexing, data structures, and matching efficiency are discussed. However, the last step in EBMT, the post-edit task has not been discussed because it is a very challenging task in EBMT which is not included in this PhD study. Development of a practical EBMT system is a long term pursuit.

6.2 Directions for Future Work

Shallow parsing can be done on non-restricted sentences in a more efficient and reliable way, and therefore it is a promising technique that can be integrated into an example-based machine translation (EBMT) system for an improved performance. We proposed an EBMT system with chunk parsing shown in Fig. 6.1, which can be described as follows:

(1) Chunk segmentation and tagging. The same as the word segmentation and tagging, chunk segmentation and tagging is also a fundamental task in Chinese information processing. Under current theories and techniques of NLP, shallow parsing is a compromised and reasonable choice.

(2) Sentence analysis based on segmented chunks. This process will contribute tremendously to the similarity measure of sentences and the

target translation generation. Base on identical Chinese chunks, the sentence analysis is performed to find relationship between chunks and to explore the sentence’s structure.

(3) Transfer and translation generation. This is also a very important task in EBMT. However, this research direction has not been explored sufficiently due to the complexity and difficulties involved.

(4) Syntactic check. The target of this task is to check the preliminary translation and refine it if necessary. Various language knowledge databases are required for this task.

Several issues need to be addressed in the future in order to develop a sophisticated EBMT system.

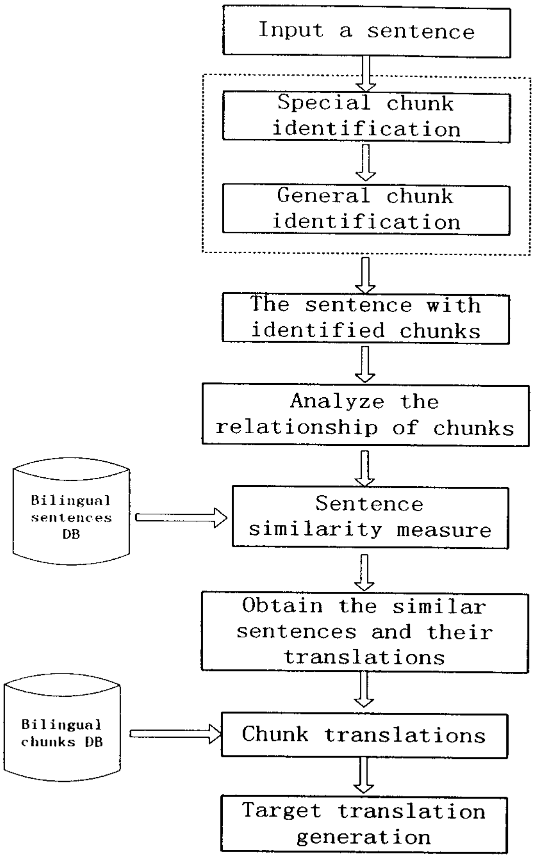

Input a sentence
Special chunk identification
General chunk identification
The sentence with identified chunks
Analyze the relationship of chunks
Bilingual sentences DB
Sentence similarity measure
Obtain the similar sentences and their translations
Bilingual chunks DB
Chunk translations
Target translation generation

Figure 6.1: The block diagram of a future EBMT system with chunk parsing.

Chinese Chunk Segmentation

Chinese chunk segmentation contains two components: special chunk identification and general chunk identification. Actually, Chinese chunk identification can be considered as a parallel process to Chinese word segmentation and tagging. The critical difference between them is the size of granularity. A chunk is normally considered to be a phrase instead of a word. At present, we perform Chinese chunk segmentation on sentences in which Chinese word segmentation and tagging has been performed.

(1) Special chunk identification

Many words or word sequences are considered as special chunks (also called name entities). Special chunks include entity names (persons, locations and organizations), times (dates and times), and various quantities such as monetary values and percentages. Due to the particularity of name entities, it is not possible to use a general method to recognize and tag them. Therefore, it is necessary to process them separately. The research in this direction is crucial. Machine learning techniques with statistical and contextual information could be a promising research direction.

(2) General chunk identification

For chunk segmentation, we have proposed a statistical model combined with a rule-based correction mechanism. Future work will focus on a direct approach for Chinese chunk segmentation with or without using machine learning techniques.

Construction of a Bilingual Chunk Corpus

Language resources play an extremely important role in natural language processing. Language resources include computer-readable dictionaries, uni-lingual, bilingual or multi-lingual aligned sentence databases with word segmentation and tagging, and so on.

Up to now, bilingual chunk (phrase) corpus construction has attracted little attention. According to the main concept of our proposed EBMT system with chunk parsing, a bilingual aligned chunk corpus will be essential in the similarity measure and sentence translation. Construction of a large Chinese chunk corpus is certainly an important task in the future. The work to be done for bilingual chunk collection includes:

(a) Collection of Chinese name entities and their translations

(b) Automatic retrieval of aligned chunks

(c) An efficient way to store and search for bilingual chunks

Sentence Similarity Measure

The performance of the sentence similarity measure will affect the quality of retrieved examples. An efficient sentence similarity measure is also desirable for dealing with a very large sentence database. In future work, we will investigate a high-performance similarity measure of Chinese sentences based on segmented chunks. Similar to the similarity measure based on word and POS tag sequences reported in this thesis, we will assign weights to various types of chunks and use a relevance feedback scheme to optimize the model parameters.

Appendix A

A Prototype for Example-Based Chinese to English Machine Translation

A.1 Introduction

Chinese is a language spoken by the largest population in the world, whereas English is the most common language used in the largest number of counties and regions. Therefore, machine translation between Chinese and English has become one of the most important research endeavors today. As an example of applying our proposed similarity measure of Chinese sentences (SMCS), in this appendix we will introduce a preliminary example-based machine translation (EBMT) system that has been implemented by us for translating Chinese to English. In this system, the post-edit task in EBMT has not been implemented due to limited time. Our target is not to implement a complete EBMT system but rather a small system to test our similarity measures of Chinese sentences. The system will provide the similar sentences to the input and their translations. The main implementation task of this prototype was carried out by a final year project student under my close supervision.

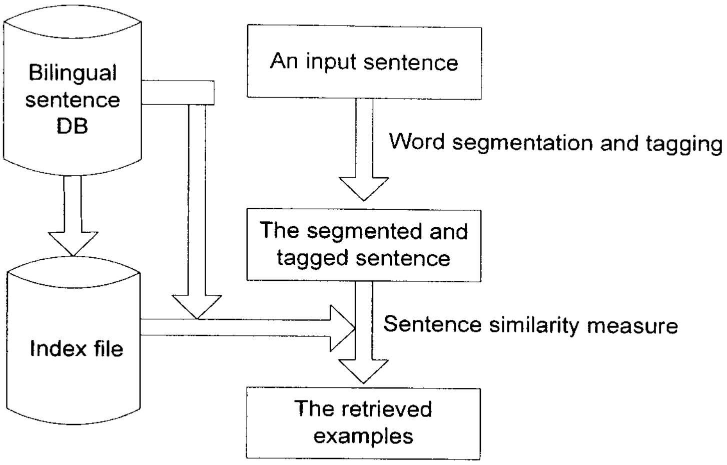

Bilingual sentence DB
An input sentence
Word segmentation and tagging
The segmented and tagged sentence
Index file
Sentence similarity measure
The retrieved examples

Figure A.1: Block diagram of a preliminary EBMT system.

The block diagram of our system is shown in Fig. A.1. The basic resources required in this preliminary system are a bilingual sentence corpus and corresponding sentence index files. In Fig. A.1, when a Chinese sentence is input, word segmentation and tagging is carried out first. This is followed by retrieving relevant examples by using the index files, and using the similarity measure to obtain the most similar Chinese sentences and their translations. The number of similar examples to be retrieved can be set by the user. All Chinese sentences in the database are indexed by the words appearing in the sentences. The translations of the retrieved sentences are used by the user as references for generating the translation of the input sentence.

The main implementation tasks of this system include: (1) building a user-friendly interface for the computer-aided EBMT system; (2) integrating the word segmentation and tagging program into the system to preprocess input sentences; (3) constructing a small corpus and corresponding index files for testing the translation system; and (4) integrating our proposed similarity measure of Chinese sentences into the system.

In this system, we made use of the advantages of XML to provide database support. We have implemented in the system the key functions that are helpful for users when they translate a whole paragraph, a sentence or a single word. It can also let users save their translation preferences. In the future, the system can be further enhanced to provide more functions to make it more powerful and more user-friendly.

In the following sections, we discuss the main issues in the implementation of this EBMT prototype.

A.2 Data Storage

All data is stored in XML files instead of a relational database management system (RDBMS). The translation engine will interact with the XML database using Microsoft's MSXML Software Development Kit (SDK), together with the XML Path Language (XPath). Besides these tools, all the programs for querying and updating the database are developed by us.

A.2.1 Database Structure

The database is illustrated as Fig. A.2.

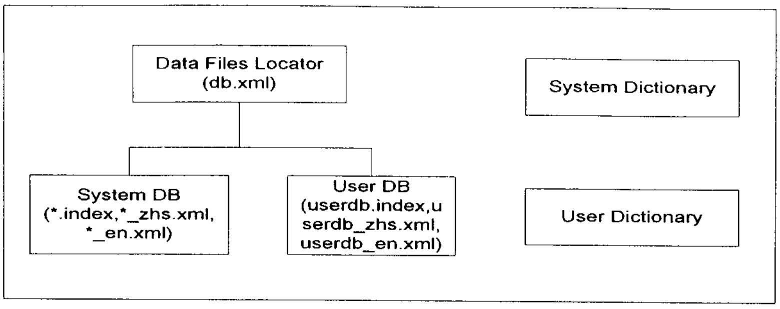


Figure A.2: The database structure of the preliminary EBMT system.

The database is mainly divided into two parts: the system database and the user database. The system database contains the data that are private to the system. Users cannot modify the data stored in the system database, except for adding new data to it. The user database has the same structure as the system database, but it is intended for storing the user-edited data. For example, when users edit a sentence and add it to the database, it will be stored in the user database instead of the system database. When the translation is performed, the translation engine will give a higher priority to the user database. That is, the user data will override the system data. When a Chinese sentence is found in both the user database and the system database, the resultant translation will be obtained from the user database rather than the system database.

A.2.2 Data Files

The database consists of four types of data files: index file, Chinese sentence file, English sentence file, and dictionary file. The index file and the dictionary file are in plain text format, while the sentences file is in XML format. Every text unit (e.g., an article) will be stored in a separate sentence file together with an index file. For example, if the article is named as "Test", then three files will be generated: “Test.index”, “Test_zhs.xml” (Chinese sentences) and “Test_en.xml” (English sentences). The article repository will be indexed by a file locator named “db.xml” that indicates which article will be used by the translation engine.

The “db.xml” looks like:

```
- <Database>
   <File>userdb</File>
   <File>fromInternet</File>
   </Database>
```

In the above example, the text unit named “userdb” and “fromInternet” will be used during the translation process. The dictionary files are also in plain text format, and they will be used to obtain the translations of Chinese words.

Index Files (*.index)

An index file, with the extension “.index”, contains all word entries of the sentences in the Chinese sentence file. Each entry indicates from which sentence an entry has come. By using an index file, it is not neccesary to search for the sentences containing a certain word in the whole database, and therefore the search efficiency is improved. This is similar to the index used in a relational database management system. The format of the index files looks like:

```
決策 :36/
員 :35/
得以 :32/
企業 :31/38/
發言 :20/
便 :19/21/22/24/
動議 :15/
致辭 :14/
```

The number(s) followed a Chinese word is the number(s) of the sentence(s) that contain the word.

Chinese Sentence Files (*_zhs.xml)

A Chinese sentence file, with the extension "_zhs.xml", stores the Chinese sentences in a text unit in XML format. Each entry will be formatted in the <s> element (the sentence number), with a parent <a> element (the paragraph number). The sentences stored in the file are accompanied by their word segmentation information which is necessary for the sentence similarity measure. Both the <a> element and the <s> element have an "id".

A Chinese sentence file looks like:

```
    - <TEXT_BODY>
    - <a id="1">
        <s id="1">因特網/經濟/如果/你/以/每/小時/90/英里/的/速度/開車
/，/你/就/會/更/快/到達/目的地/——/但是/即使/路上/的/一個/小/坑/也/
可能/導致/一/場/災難/。/</s>
        </a>
    - <a id="2">
        <s id="1">這/簡明/而/形象/地/概括/了/美國/經濟/目前/的/困境/。
/</s>
        </a>
        </TEXT_BODY>
```

English Sentence Files (*_en.xml)

An English sentence file, with the extension "_en.xml", is similar to a Chinese sentence file. The sentences are also stored in the <s> element, with a parent element <a>. Both the <s> and <a> elements have an id, and one <a> element only contains one <s> element, more <s> elements may be added for an enhancement version. However, in the English sentence file, no "/" tag is used to separate the words. The translation engine will use the ID to obtain the translation corresponding to the Chinese sentence from the file. An English sentence file looks like:

```
    <TEXT_BODY>
    - <a id="1">
      <s id="1">The Internet Economy When you drive a car at 90 mph, you get to your destination faster-but even a pothole in the road can turn into a disaster.</s>
      </a>
    - <a id="2">
      <s id="1">That, in a nutshell, sums up the dilemma of the U.S. economy these days.</s>
      </a>
      </TEXT_BODY>
```

Dictionary (*.dict)

A dictionary file is stored in plain text format. The translation engine will look up the dictionary file when it tries to translate a Chinese word. One Chinese word may have multiple English translations which are separated by “/”s.

A dictionary file looks like:

```
一個 /a/
也   /also/
一   /one/
時期 /a period in time or history/
目的地        /destination (location)/
以   /to use/
埃及 /Egypt/
的   /(possessive particle)/
經濟 /economy/
```

```
地  /(subor. part. adverbial)/
黟 /black and shining ebony/
```

The User Database

Besides the system database, there is also a user database for storing the user' translations to obtain customized results. Three files named "userdb.index", "userdb_zhs.xml" and "userdb_en.xml" store the sentence translations refined by the user. In addition, a file named "user.dict" stores word translations that provided by the user. All data stored in the user database will have a higher priority. The user can update the user database via the user interface.

A.3 Summary on System Implementation

The similarity measures discussed in previous section are implemented in this simple Example-Based Machine Translation (EBMT) system. Two external tools named "SegTag" and "Aligner" have been integrated into the system to perform word segmentation and tagging, and alignment of bilingual text units, respectively. The C-E translation pairs in the database are obtained by the tool "Aligner" which was purchased from the ICL of Peking University.

Since the efficiency is one of the main concerns in the implementation, C/C++ has been chosen as the programming language to implement the system. The system is aimed in providing a user-friendly interface to assist users to translate Chinese sentences into English ones. The system can also accommodate the user' intention and preference in translation. The performance of the system will be gradually improved as a user makes use of it.

The system consists of three main components: the user interface, the translation engine, and the database. The system architecture is illustrated in Fig. A.3. The inputted Chinese text from the interface is fed to the translation engine. Some necessary pre-processing to the input sentence, such as word segmentation and tagging is performed. The sentence similarity measure is then carried out to find the most similar sentences from the database. Finally, translations of the most similar sentences will be retrieved and displayed in the interface. The user can then modify them if necessary.

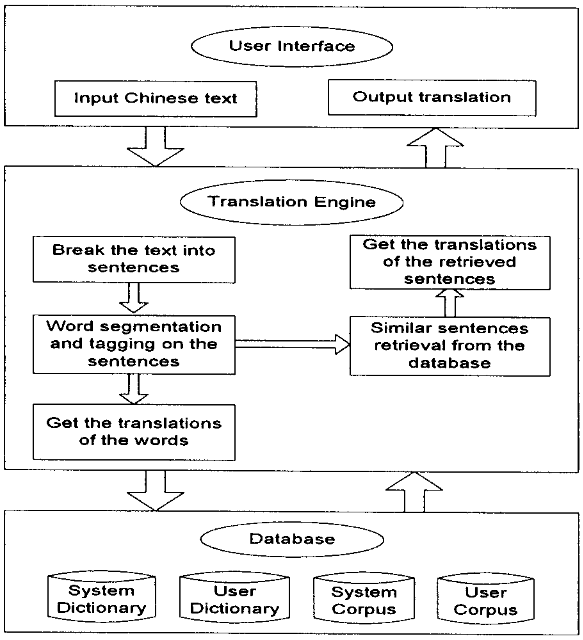

User Interface
Input Chinese text
Output translation
Translation Engine
Break the text into sentences
Get the translations of the retrieved sentences
Word segmentation and tagging on the sentences
Similar sentences retrieval from the database
Get the translations of the words
Database
System Dictionary
User Dictionary
System Corpus
User Corpus

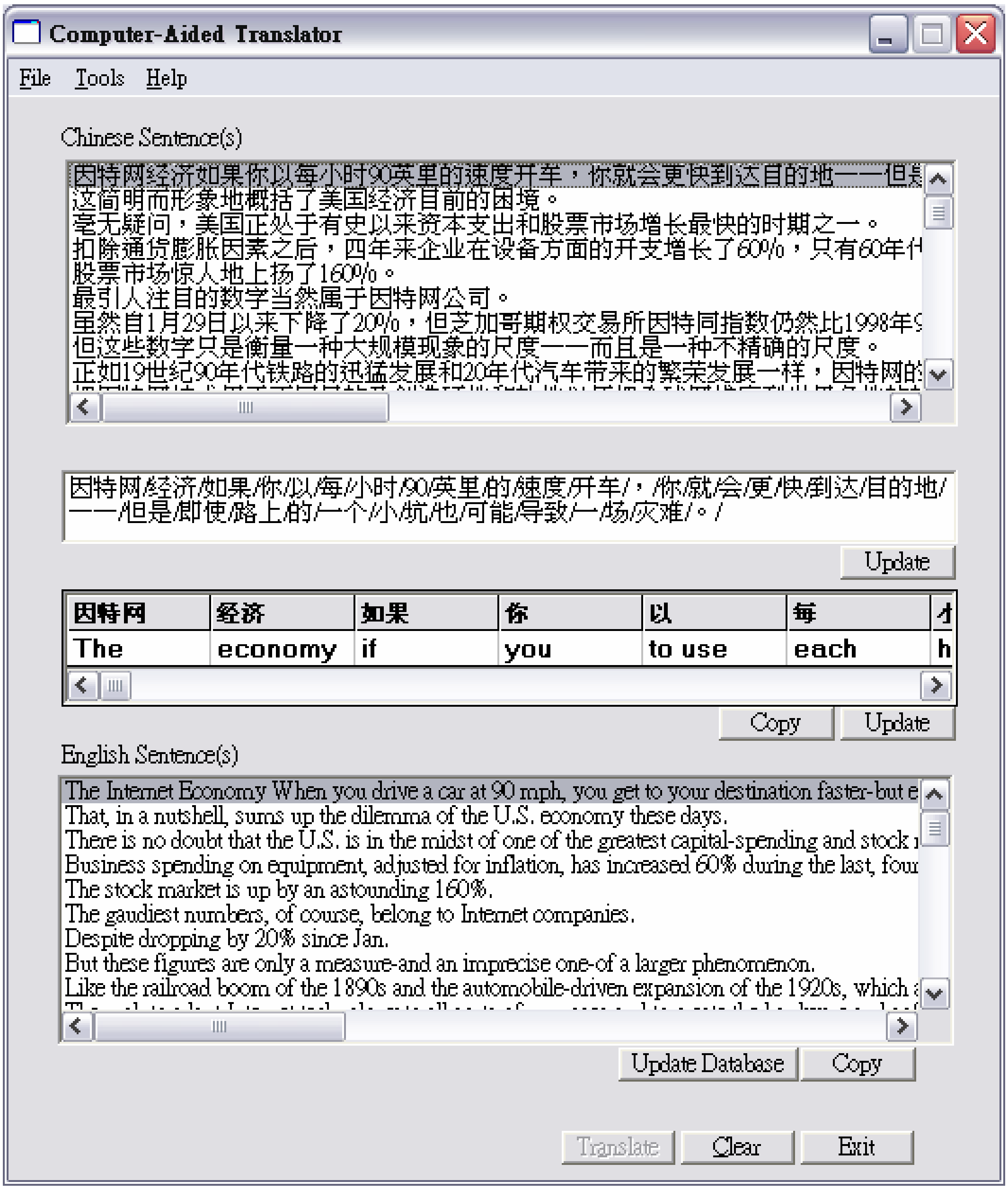
Figure A.3: The system architecture of the preliminary EBMT system.

A.4 User Interface

The user interface provides various functions for the user to input Chinese text consisting of one or more sentences, or maybe a single Chinese word to be translated. The user can also add their own Chinese corpora to the database, or update the database to meet their needs. Moreover, within the interface, the user can set various parameters for the system, e.g., the number of similar sentences to be retrieved.

A.5 Translation of Chinese Text

It is more reasonable to consider the system as a computer-aided translation system. This is because the final task in an EBMT system, which is the target translation generation, has not been implemented in the current system. To translate Chinese text into English, the user can input the text to be translated. He/she can highlight one or more sentences of the input text to obtain the translations by clicking the “Translate” button (Fig. A.4). The result will be displayed in the interface. The user can then modify the translation of the input sentence based on the retrieved examples.



因特网经济如果你以每小时90英里的速度开车，你就会更快到达目的地——但
这简明而形象地概括了美国经济目前的困境。
毫无疑问，美国正处于有史以来资本支出和股票市场增长最快的时期之一。
扣除通货膨胀因素之后，四年来企业在设备方面的开支增长了60%，只有60年代
股票市场惊人地上扬了160%。
最引人注目的数字当然属于因特网公司。
虽然自1月29日以来下降了20%，但芝加哥期权交易所因特网指数仍然比1998年9
但这些数字只是衡量一种大规模现象的尺度——而且是一种不精确的尺度。
正如19世纪90年代铁路的迅猛发展和20年代汽车带来的繁荣发展一样，因特网的

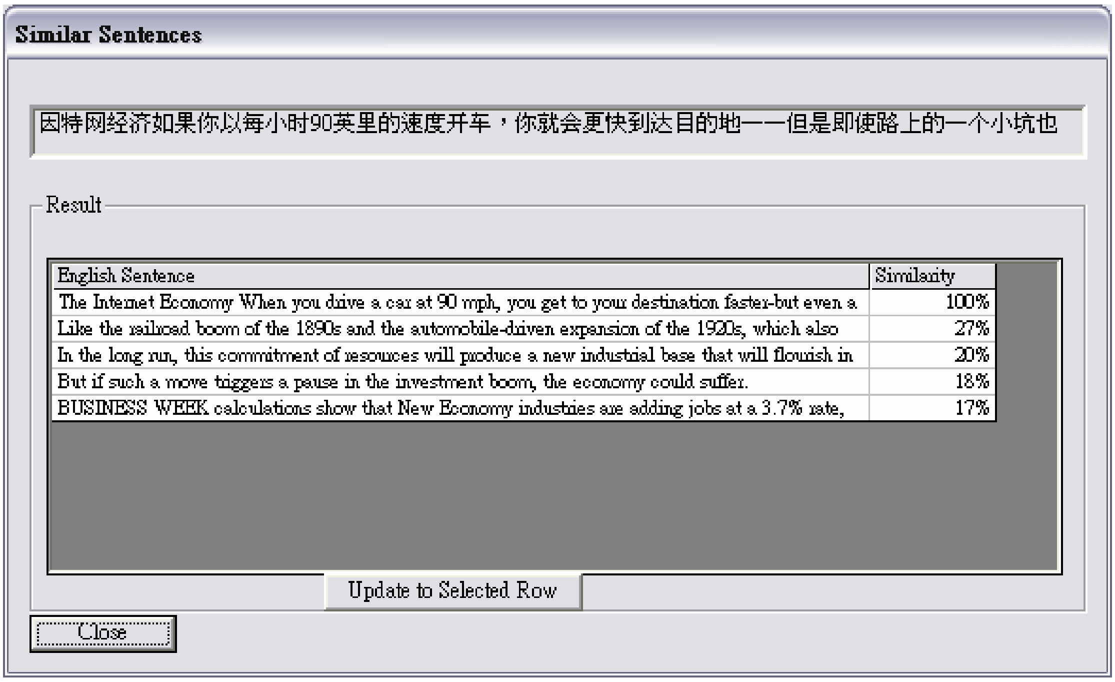
因特网/经济/如果/你/以/每/小时/90/英里/的/速度/开车/，/你/就/会/更/快/到达/目的地/
——/但是/即使/路/上/的/一个/小/坑/也/可能/导致/一/场/灾难/。/

因特网	经济	如果	你	以	每
The	economy	if	you	to use	each

The Internet Economy When you drive a car at 90 mph, you get to your destination faster-but e
That, in a nutshell, sums up the dilemma of the U.S. economy these days.
There is no doubt that the U.S. is in the midst of one of the greatest capital-spending and stock
Business spending on equipment, adjusted for inflation, has increased 60% during the last, four
The stock market is up by an astounding 160%.
The gaudiest numbers, of course, belong to Internet companies.
Despite dropping by 20% since Jan.
But these figures are only a measure-and an imprecise one-of a larger phenomenon.
Like the railroad boom of the 1890s and the automobile-driven expansion of the 1920s, which

Figure A.4: The interface for the translation of input sentences.

When the user clicks on a Chinese sentence in the upper list box, the corresponding segmentation result of the sentence will be shown in the second text box. At the same time, the translation of the Chinese words in the sentence will also be displayed in the grids below it. The user can double click any grid to edit the translation result, and can click the “Update” button to update the result to the user database. When the user double clicks on a row in the bottom English list box, a pop-up dialog will be shown. The dialog window (Fig. A.5) contains the English translation corresponding to the most similar Chinese sentences for the user to edit. These sentences are listed in a descending order of the similarity measure.



因特网经济如果你以每小时90英里的速度开车，你就会更快到达目的地——但是即使路上的一个小坑也

English Sentence	Similarity
The Internet Economy When you drive a car at 90 mph, you get to your destination faster-but even a	100%
Like the railroad boom of the 1890s and the automobile-driven expansion of the 1920s, which also	27%
In the long run, this commitment of resources will produce a new industrial base that will flourish in	20%
But if such a move triggers a pause in the investment boom, the economy could suffer.	18%
BUSINESS WEEK calculations show that New Economy industries are adding jobs at a 3.7% rate,	17%

Figure A.5: The interface for viewing and editing the most similar sentences.

After the user finishes their sentence editing operations, he/she can return to the main interface where he/she can click the “Update Database” button to update the edit result to the user database.

A.6 Other Functions

Translation of Chinese Words

Besides Chinese sentences, this application also provides a dictionary function for the user to translate a single Chinese word.

Adding a Text Unit to the Database

As we know, the more bilingual sentences are collected so does the performance of an EBMT system improve. So data (sentences) collection is a very important task for EBMT. This system allows the user to add bilingual text units to the database.

Up to now, ten bilingual text units are collected from the website of The Hong Kong Exchanges and Clearing Limited comprising mainly of news about the Hong Kong stock market. In addition, we have also collected eleven passages on labor laws from the Labor Department of The Hong Kong SAR. We have collected another thirteen passages on Hong Kong laws from the Department of Justice of The Hong Kong SAR.

Data Conversion

To facilitate the user to add text units to the database, a standalone data conversion tool has also been developed. This data conversion tool has a simple and easy-to-use interface. It contains two text boxes. One is for inputting the path of the Chinese sentence file and the other for inputting the path of the corresponding English sentence file.

A.7 Translation Engine

The translation engine is the most important part of this application. Its main task is to search for the most similar sentences to the input one from the database based on our proposed similarity measures of Chinese sentences.

After the preprocessing of the input Chinese sentence including word segmentation and tagging, the similarity measure can be performed. It involves the following steps:

(1) All the words of the current sentence to be translated are extracted. Using each word in the sentence as the entry, the translation engine retrieves all corresponding sentence IDs from the index file. By the IDs, all corresponding sentences will be exacted.

(2) The similarity measure between all retrieved sentences and the current one will be performed using the similarity measure discussed in Chapter 4. The top similar sentences are shown at a descending order of the similarity values.

A.8 Conclusion

Machine translation between Chinese and English has become one of the most important research endeavors today. In this appendix, we describe a preliminary example-based machine translation (EBMT) system for translating Chinese sentences to English ones. In this system, we implement our proposed similarity measure of Chinese sentences discussed in Chapter 4. The system is still in its infancy stage and should be further enhanced in the future, including the use of a larger-scale bilingual sentences database and exploration of more sophistical searching techniques for a huge sentence database and integration of post-edit task. A practical EBMT system is a long term pursuit.

References

Abney, S. (1991). “Parsing by Chunks,” in: Berwick A. & Tenny (eds), *Principle-based Parsing,* pp. 257-278, Kluwer Academic Publisher.

Belongie, S., Malik, J. & Puzicha, J. (2002). “Shape Matching and Object Recognition using Shape Contexts,” *IEEE Transaction on Pattern Analysis and Machine Intelligence, 24(4)*, pp. 509-522.

Brown, P. F., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R. & Roossin, P. (1988). “A Statistical Approach to French/English Translation,” in *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI-88)*, Pittsburgh, PA, USA, June 12-14. Panel 2: Paradigms for MT.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. & Roossin, P. S. (1990). “A Statistical Approach to Machine Translation,” *Computational Linguistics*, Vol. 16, pp. 79-85.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lafferty, J. D. & Mercer, R. L. (1992). “Analysis, Statistical Transfer, and Synthesis in Machine Translation,” in *Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation(TMI-92)*, pp. 83-100, Montreal. Canada, June 25-27.

Brown, P. F., Della Pietra S. A., Della Pietra, V. J. & Mercer, R. L. (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation,” *Computational Linguistics 19*, pp. 263-311.

Carl, M. (1999). “Inducing Translation Templates for Example-Based Machine Translation,” in *Proceedings of Machine Translation Summit VII*, pp. 250-258, Singapore, September 13-17.

Celentano, A., & Sciasicio, E. D. (1998). “Feature Integration and Relevance Feedback Analysis in Image Evaluation,” *Journal of Electronic Imaging*, 7(2), pp. 308-317.

Che, W. X., Liu, T., Qin, B. & Li, S. (2003). “Chinese Sentences Similarity Computational Oriented the Searching in Bilingual Sentence Pairs,” *Joint Symposium on Computational Linguistics 2003*, pp. 81-86, Harbin, China, August 8-11.

Chen, H. M., Varshney, P. K., and Arora, M. K. (2003). “Performance of Mutual Information Similarity Measure for Registration of Multi-temporal Remote Sensing Images,” *IEEE Transactions on Geoscience and Remote Sensing*, 41(11), pp. 2445-2454.

Chen, Z. X., Xia, Y. Q., Huang, H. Y. & Wang, J. D. (2001). “A Multi-level Feature-based Approach to Calculate Similarity of Two Sentences in IHSMTS system,” In *Proceeding of the 1st International Symposium on Natural Language Processing and Developing*, pp. 1-6, Shanghai, China.

Cicekli, I. & Güvenir, H. A. (1996). “Learning Translation Rules from a Bilingual Corpus,” in *Proceedings of the Second International Conference on New Methods in Language Processing*, pp. 90-97, Ankara, Turkey, September.

Collins, M. J. (1996). “A New Statistical Parser based on Bigram Lexical Dependencies,” in *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics*, pp. 184−191, California, USA, June 24-27.

Cranias, L., Papageorgiou, H. & Piperidis, S. (1994). “A Matching Technique in

Example-Based Machine Translation,” in *Proceedings of 15th International Conference on Computational Linguistics(Coling 1994)*, pp. 100-104, Kyoto, Japan, August 5-9.

Daelemans, W., Buchholz, S. & Veenstra, J. (1999). “Memory-based Shallow Parsing,” in *Proceedings of Computational Natural Language Learning (CoNLL-99)*, pp53-60, Bergen, Norway, June 12.

Doulamis, N. D., Doulamis, A. D. & Kollias, S. D. (2000). “Non-linear Relevance Feedback Improving the Performance of Content-based Retrieval System,” in *IEEE International Conference on Multimedia and Expo (ICME2000)*, Vol. 1, pp. 331-334, Hilton New York & Towers, New York City, NY, USA, July 30 - August 2.

Eidenberger, H. & Breiteneder, C. (2003). "Visual Similarity Measurement with the Feature Contrast Model," in *Proceedings SPIE Storage and Retrieval for Media Databases Conference*, Vol. 5021, pp. 64-76, SPIE, Santa Clara, California, USA, January 20-24.

El-Naqa, W., Wernick, M. N., Yand, Y. & Galatsanos, N. P. (2000). “Image Retrieval based on Similarity Learning,” in *Proceedings of International Conference on Image Processing (ICIP2000)*, Vol. 3, pp. 722-725, Vancouver, Canada, September 10-13.

Furuse, O. & Iida, H. (1994). “Constituent Boundary Parsing for Example-Based Machine Translation,” in *Proceedings of the 15th International Conference on Computational Linguistics (Coling1994)*, pp. 105-111, Kyoto, Japan, August 5-9.

Gale, W. A. & Church, K. W. (1993). “A Program for Aligning Sentences in Bilingual Corpora,” *Computational Linguistics 19*, pp. 75-102.

Gowda, K. C. & Diday, E. (1992). “Symbolic Clustering Using a New Similarity

Measure,” *IEEE Transactions on Systems, Man, and Cybernetic* 22(2), pp. 368-378.

Grishman, R. (1994). “Iterative Alignment of Syntactic Structures for a Bilingual Corpus,” in *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC2)*, pp. 57-68, Kyoto, Japan, August 4.

Gudivada, V. N. & Raghavan, V. V. (1995). “Design and Evaluation of Algorithms for Image Retrieval by Spatial Similarity,” *ACM Transactions on Information Systems*, 13(2), pp. 115-144.

Guo, F., Jin, J. & Feng, D. (1998) “Measuring Image Similarity using the Geometrical Distribution of Image Contents,” in *Proceedings of 4th International Conference on Signal Processing*, Vol. 2, pp. 1108-1112, Beijing, China, October.

Güvenir, H. A. & Cicekli, I. (1998). “Learning Translation Templates from Examples,” *Information Systems,* 23(6), pp. 353-363.

Halteren, H. V. (2000). “Chunking with WPDV Models,” in *Proceedings of CoNLL-2000 and LLL-2000,* pp. 154-156, Lisbon, Portugal, September 13-14.

Hutchins, W. J. (1995). “Machine Translation: a brief history,” *Concise History of the Language Sciences: from the Sumerians to the Cognitivists*, pp. 431-445, Oxford: Pergamon Press.

James, H., Miles, O., Susan, A. & Walter, D. (2002). “Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing,” *Journal of Machine Learning Research 2*, pp. 551-558.

Jones, D. (1992). “Non-hybrid Example-based Machine Translation Architectures,” in *Proceedings of 4th International Conference on Theoretical* *and Methodological Issues in Machine Translation(TMI-92)*, pp. 163-71, Montreal, Canada, June 25-27.

Jones, D. (1996). *Analogical Natural Language Processing*, London: UCL Press, UK.

Jones, D. & Somers, H. (Eds.). (1997). *New Methods in Language Processing*, London: UCL Press, UK.

Juola, P. (1994). "A Psycholinguistic Approach to Corpus-Based Machine Translation," in *Proceedings of the 3rd Conference on the Cognitive Science of Natural Language Processing (CSNLP 1994)*, pages not numbered, Dublin, Ireland, July 6-8.

Kaji, H., Kida, Y. & Morimoto, Y. (1992). "Learning Translation Templates from Bilingual Text," in *Proceedings of the 14th International Conference on Computational Linguistics(Coling1992)*, pp. 672-678, Nantes, France, August 23-28.

Katoh, N. & Aizawa, T. (1994). "Machine Translation of Sentences with Fixed Expressions," in *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 28-33, Stuttgart, Germany, October 13-15.

Kitano, H. (1994). "Speech-to-Speech Translation: A Massively Parallel Memory-Based Approach," *Computational Linguistics*, 21(4), pp. 590-592.

Kudoh, T. & Matsumoto, Y. (2000). "Use of Support Vector Learning for Chunk Identification," in *Proceedings of CoNLL-2000 and LLL-2000*, pp. 142-144. Lisbon, Portugal, September 13-14.

Kudoh, T. & Matsumoto, Y. (2001). "Chunking with Support Vector Machines," in *Proceedings of The Second Meeting of the North American Chapter of the*

Association for Computational Linguistics (NAACL2001), pp. 192-199, Pittsburgh, PA, USA, June 2-7.

Li, B., Liu, T., Qin, B. & Li, S. (2003a). “Chinese Sentence Similarity Computing based on Semantic Dependency Relationship Analysis,” *Application Research of Computer* 20(12), pp. 15-17.

Li, H., Zhu, J. B. & Yao, T.S. (2004a). “SVM Based Chinese Text Chunking,” *Journal of Chinese Information Processing*, 18(2), pp. 1-7.

Li, H., Tan, Y. M., Zhu, J. B. & Yao, T. S. (2004b). “Recognition of Chinese Chunk,” *Journal of Northeast University (Natural Science), China*, 25(2), pp.114-117.

Li, M., Lv, X.Q. & Yao, T. S. (2002a). “A Machine Translation Model Based on E-chunk,” *Journal of Software*, 13(4), pp. 669-676.

Li, S. J. (2002). “Research of Relevancy between Sentences based on Semantic Computation,” *Computer Engineering and Applications* 38(7), pp. 75-76.

Li, S. J., Liu, Q. & Bai, S. (2002b). “Chinese Chunking Parsing using Rule-based and Statistics-based Methods,” *Journal of Computer Research and Development*, 39(4), pp.385-391.

Li, S. J., Liu, Q. & Yang, Z. F. (2003b). “Chunk Parsing with Maximum Entropy Principle,” *Chinese Journal of Computers*, 26(12), pp. 1722-1727.

Lim, J. H., Wu, J. K., Singh, S. & Narasimhalu, D. (2001). “Learning Similarity Matching in Multimedia Content-based Retrieval,” *IEEE Transactions on Knowledge and Data Engineering*, 13(5), pp. 846-850, September.

Lin, D. K. (1998). “An Information-theoretic Definition of Similarity,” in *Proceedings of the 15th International Conference on Machine*

Learning(ICML-98), pp. 296-304, Wisconsin, Madison, USA, July 24-27.

Liu, F., Zhao, T. J., Yu, H., Yang, M. Y. & Fang, G. L. (2000). “Statistics-Based Chinese Chunk Parsing,” *Journal of Chinese Information Processing*, 14(6), pp.28-32, 39.

Lu, S., Li, X. L., Bai, S. & Wang, S. (2000). “An Improved Approach to Weighting Terms in Text,” *Journal of Chinese Information Processing* 14(6), pp. 8-13.

Mandreoli, F., Martoglia, R. & Tiberio, P. (2002). “Searching Similar (sub)Sentences for Example-based Machine Translation,” in *Proceedings of the 10th National Congress on Advanced Database Systems (SEBD 2002)*, pp.208-221, Isola d'Elba, Italy, June.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press, Massachusetts, USA.

Maruyama, H. & Watanabe, H. (1992). “Tree Cover Search Algorithm for Example-based Translation,” in *Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation(TMI-92)*, pp. 173-84, Montreal. Canada, June 25-27.

Matsumoto, Y., Ishimoto, H. & Utsuro, T. (1993). “Structural Matching of Parallel Texts,” in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 23-30, Columbus, Ohio, June 22-26.

McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*, Edinburgh: Edinburgh University Press..

McLean, I. (1992). “Example-based Machine Translation using Connectionist Matching,” in *Proceedings of 4th International Conference on Theoretical and*

Methodological Issues in Machine Translation(TMI-92), pp. 35-43, Montreal. Canada, June 25-27.

McTait, K. & Trujillo, A. (1999). “A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns,” in *Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pp. 98-108, Chester, UK, August 23-25.

Megyesi, B. (2002). “Shallow Parsing with POS Taggers and Linguistic Knowledge,” *Journal of Machine Learning Research 2*, pp. 639-668.

Meyers, A., Yangarber, R. & Grishman, R. (1996). “Alignment of Shared Forests for Bilingual Corpora,” in *Proceedings of the 16th International Conference on Computational Linguistics (Coling1996)*, pp. 459-465, Copenhagen, Denmark, August 5-9.

Molina, A. & Pla, F. (2002). “Shallow Parsing Using Specialized HMM,” *Journal of Machine Learning Research 2*, pp. 595-613.

Nagao, M. (1984). “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle,” in: A. Elithorn & R. Banerji (eds.), *Artificial and Human Intelligence*, NATO Publications, pp173-180.

Nirenburg, S., Domashnev, C. & Grannes, D. J. (1993). “Two Approaches to Matching in Example-Based Machine Translation,” in *Proceedings of 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1993)*, pp. 47-57, Kyoto, Japan, July 14-16.

Nirenburg, S., Shell, P., Cohen, A., Cousseau, P., Grannes, D. & McNeilly, C. (1992). “Multi-Purpose Development and Operation Environments for Natural Language Generation,” in *Proceedings of the Third Conference on Natural Language Applications*, pp. 255-256, Trento, Italy, March 31 - April 3.

Nirenburg, S., Beale, S. & Domashnev, C. (1994). “A Full-Text Experiment in Example-Based Machine Translation,” in *Proceedings of the International Conference on New Methods in Language Processing*, pp. 78-87, Manchester, England, September 14-16.

Nomiyama, H. (1992). “Machine Translation by Case Generalization,” in *Proceedings of the 14th International Conference on Computational Linguistics(Coling1992)*, pp. 714-720, Nantes, France, August 23-28.

Oi, K., Sumita, E., Furuse, O., Iida, H. & Higuchi, T. (1994). “Real-Time Spoken Language Translation Using Associative Processors,” in *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 101-106, Stuttgart, Germany, October 13-15.

Osborne, M. (2002). “Shallow Parsing Using Noisy and Non-stationary Training Material,” *Journal of Machine Learning Research 2*, pp. 695-719.

Patrice, B. & Konik, H. (2000). “Texture Similarity Queries and Relevance Feedback for Image Retrieval,” in *Proceedings of the 15th International Conference on Pattern Recognition*, Vol. 2, pp. 55-58, Barcelona, Spain, September 3-8.

Picard, R., Minka, T. & Szummer, M. (1996). “Modeling User Subjectivity in Image Libraries,” in *Proceedings of International Conference on Image Processing (ICIP1996)*, pp. 777-780, Lausanne, Switzerland, September 16-19.

Ramshaw, L. A. & Marcus, M. P. (1995). “Text Chunking Using Transformation-based Learning,” in *Proceedings of ACL’95 Workshop on Very Large Corpora*, pp. 82-94, Massachusetts, USA, Jun 1.

Ren, F. J. (1999). “Super-function Based Machine Translation,” *Communications of COLIPS*, 9 (1), pp. 83-100.

Rui, Y. & Huang, T. S. (1999). "A Novel Relevance Feedback Technique in Image Retrieval," in *Proceedings of the 7th ACM International Conference on Multimedia*, pp. 67-70, Orlando, USA, October 30 - November 5.

Rui, Y., Huang, T. S., Ortega, M. & Mehrotra, S. (1998). "Relevance Feedback: a Power Tool in Interactive Content-based Image Retrieval," *IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Segmentation, Description, and Retrieval of Video Content,* 8(5), pp. 644-655.

Rui, Y., Huang, T. S. & Mehrotra, S. (1997). "Content-based Image Retrieval with Relevance Feedback in MARS," In *Proceedings of International Conference on Image Processing (ICIP1997),* pp. 815-818, Washington, DC, USA, October 26-29.

Shyu, M. L., Chen, S. C., Chen, M., Rubin, S. H. (2004). "Affinity-Based Similarity Measure for Web Document Clustering," in *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2004),* pp. 247-252, Las Vegas, Nevada, USA, November 8-10.

Skut, W. & Brants, T. (1998). "Chunk Tagger: Statistical Recognition of Noun Phrases," in *Proceedings of ESSLLI-1998 Workshop on Automated Acquisition of Syntax and Parsing*, pp. 189-192, Saarbruucken, Germany, August 17-28.

Sobashima, Y., Furuse, O., Akamine, S., Kawai, J. & Iida, H. (1994). "A Bidirectional, Transfer-driven Machine Translation System for Spoken Dialogues," in *Proceedings of 15th International Conference on Computational Linguistics (Coling1994)*, pp. 64-68, Kyoto, Japan, August 5-9.

Somers, H. L. (1997). "The Current State of Machine Translation," *MT Summit VI: Machine Translation Past Present Future*, pp. 115-124, San Diego, California.

Somers, H. (1998). “Further Experiments in Bilingual Text Alignment,” *International Journal of Corpus Linguistics 3*, pp. 1-36.

Somers, H. (1999). “Review Article: Example-based Machine Translation,” *Machine Translation 14*, pp. 113-157.

Stricker, M. & Orengo, M. (1995). “Similarity of Color Images,” in *Proceedings of SPIE Storage and Retrieval for Image and Video Databases III*, Vol. 2420, pp. 381-392, San Diego/La Jolla, CA, USA, February 5-10.

Sui, Z. F. & Yu, S. W. (1998). “Skeleton-based Computational Model of Chinese Sentence Similarity,” *International Conference on Chinese Information Processing (ICCIP’98)*, pp. 458-465, Beijing, China, November 18-20.

Sumita, E. & Iida, H. (1991). “Experiments and Prospects of Example-based Machine Translation,” in *Proceedings of 29th ACL Meeting*, pp. 185-192, Berkeley, California, USA, June 18-21.

Sumita,E., Iida, H. & Kohyama, H. (1990). “Translating with Examples: A New Approach to Machine Translation,” in *Proceedings of The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pp. 203-212, Texas, USA, June 11-13.

Sun, H. L. & Jurafsky, D. (2004). “Shallow Semantic Parsing of Chinese,” in *Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (NAACL-HLT2004)*, pp. 249-256, Boston, USA, May 2-7.

Sun, H. L. & Yu, S. W. (2000). “Overview of Shallow Parsing,” *Contemporary Linguistics*, 2(2), pp.74-83.

Tjong Kim Sang, E. F. (2000). “Text Chunking by System Combination,” in

Proceedings of CoNLL-2000 and LLL-2000, pp. 151-153, Lisbon, Portugal, September 11-14.

Tjong Kim Sang, E. F. (2002). “Memory-based Shallow Parsing,” *Journal of Machine Learning Research 2*, pp. 559-594.

Vasconcelos, N. & Lippman, A. (2000). “Bayesian Relevance Feedback for Content-based Image Retrieval,” in *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 63-67, June 12.

Voutilainen, A. (1993). “NPTool, a Detector of English Noun Phrases,” in *Proceedings of the Workshop on Very Large Corpora*, pp. 48-57, Columbus, Ohio, USA, June.

Wang, R. B. & Chi, Z. R. (2005). “A Similarity Measure Method of Chinese Sentence Structures,” *Journal of Chinese Information Processing* 19(1), pp. 21-29.

Wang, R. B., Chi, Z. R., Chang, B. B. & Bai, X. J. (2005). “An Improved Similarity Measure of Chinese Sentences,” *Journal of Information* 8(1), pp. 139-145.

Wang, Z. Y., Chi, Z. R. & Feng, D. G. (2003). “Content-based Image Retrieval with Relevance Feedback using Adaptive Processing of Tree-structure Image Representation,” *International Journal of Image and Graphics* 3(1), pp. 119-143.

Watanabe, H. (1992). “A Similarity-driven Transfer System,” in *Proceedings of the 14th International Conference on Computational Linguistics(Coling1992)*, pp. 770-776, Nantes, France, August 23-28.

Watanabe, H. & Takeda, K. (1998). “A Pattern-Based Machine Translation System Extended by Example-Based Processing,” in *Proceedings of 36th Annual*

Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL1998), pp. 1369-1373, Montreal, Quebec, Canada, August 10-14.

Wong, K. F., Chan, K. C. & Cheng, C. H. (2001). "An Investigation on Transformation-based Error-driven Learning Algorithm for Chinese Noun Phrase extraction", *International Journal of Computer Processing of Oriental Languages*, 14(1), pp. 47-69, World Scientific, USA.

Wu, K., Shi, B., Lu, J. & Zhu, X. F. (2004). “Feature Selection and Weighting Scheme based on Text Set Density,” *Journal of Chinese Information Processing* 18(1), pp. 42-47.

Wu, Y., Tian, Q. & Huang, T. S. (2000). “Integrating Unlabeled Images for Image Retrieval Based on Relevance Feedback,” in *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'2000)*, Vol. 1, pp.21-24, Barcelona, Spain, September 3-8.

Xi, C. H. & Sun, M. S. (2002). “Automatic Prediction of Chinese Phrase Boundary Location with Neural Networks,” *Journal of Chinese Information Processing*, 16(2), pp. 20-26.

Xu, J. L., (2002). “The State-of-the-Art and the Related Strategic Considerations,” *Journal of Chinese Information Processing*, 15(2), pp. 1-8.

Yoon, Y. & Jayant, N. (2001). “Relevance Feedback for Semantics based Image Retrieval,” in *Proceedings of International Conference on Image Processing (ICIP2001)*, Vol. 1, pp. 42-45, Thessaloniki, Greece, October 7-10.

Zhang, K. & Shasha, D. (1997). “Tree Pattern Matching,” in A. Apostolico & Z. Galil (eds), *Pattern Matching Algorithms*, pp. 341-371, New York: Oxford University Press, USA.

Zhang, T., Damerau, F. & Johnson, D. (2002). “Text Chunking Based on a Generalization of Winnow," *Journal of Machine Learning Research 2*, pp. 615-637.

Zhou, G. D., Su, J. & Tey, T. G.. (2000). “Hybrid Text Chunking,” in *Proceedings of CoNLL-2000 and LLL-2000*, pp. 163-166, Lisbon, Portugal, September 13-14.

Zhou, Q. (1996). “A Model for Automatic Prediction of Chinese Phrase Boundary Location,” *Journal of Software*, Vol 7, Supplement, pp.315-322.

Zhou, Q., Sun, M. S. & Huang, C. N. (1999). “Chunk Parsing Scheme for Chinese Sentences,” *Chinese Journal of Computers*, 22(11), pp. 1158-1165.