密　级：　　学校代码：10075
分类号：　　学　　号：gxjs08123

文学硕士学位论文

机器翻译中英译汉长句分析研究

学位申请人：　李宏毅
指 导 教 师 ：　叶慧君　副教授
学 位 类 别 ：　文学硕士
学 科 专 业 ：　英语语言文学
授 予 单 位 ：　河北大学
答 辩 日 期 ：　二〇一〇年六月

Classified Index: CODE: 10075

U. D. C: NO: gxjs08123

A Dissertation for the Degree of M. Arts

A Study of Long Sentence Parsing in English-Chinese Machine Translation

Candidate: Li Hongyi

Supervisor: Associate Prof. Ye Huijun

Academic Degree Applied for: Master of Arts

Specialty: English Language literature

University: Hebei University

Date of Oral Examination: June, 2010

河北大学

学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得河北大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了致谢。

作者签名：[signature] 日期：2010年6月1日

学位论文使用授权声明

本人完全了解河北大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

本学位论文属于

1、保密 □，在_____年_____月_____日解密后适用本授权声明。

2、不保密 √。

（请在以上相应方格内打“√”）

保护知识产权声明

本人为申请河北大学学位所提交的题目为（A Study of Long Sentence Parsing in English-Chinese Machine Translation）的学位论文，是我个人在导师（叶慧君副教授）指导并与导师合作下取得的研究成果，研究工作及取得的研究成果是在河北大学所提供的研究经费及导师的研究经费资助下完成的。本人完全了解并严格遵守中华人民共和国为保护知识产权所制定的各项法律、行政法规以及河北大学的相关规定。

本人声明如下：本论文的成果归河北大学所有，未经征得指导教师和河北大学的书面同意和授权，本人保证不以任何形式公开和传播科研成果和科研工作内容。如果违反本声明，本人愿意承担相应法律责任。

声明人：李富融　　日期：2010年6月1日

作者签名：李富融　　日期：2010年6月1日

导师签名：叶慧君　　日期：2010年6月1日

Abstract

After decades of development and improvements, as the information and computer techniques are being perfected day by day, the research on Machine Translation (MT) has become a comprehensive discipline from the very beginning of computational linguistics, which is now related to many other subjects, such as semantics, mathematics, corpus, computer science, Artificial Intelligence(AI) technology and biology, etc. However, the translation quality of MT still can't meet requirements, especially lacking capacities of analyzing long sentences. Although compared to the past decades, there are qualitative leaps in the computer technology and many related technologies, the problem of parsing long sentences is still a formidable obstacle in this field.

The methods of parsing long sentences are quite different from finding word meanings from dictionaries. The lexical translation only needs lemmatization and tokenization to check the original forms of words in the data base. The lexical analysis is only the initial procedure of the parsing process, in the following, it also needs to identify the context-sensitive ambiguous, the far-distance related words. And then during the procedure of parsing the sentence structures, the problems, such as phrases included in the sub-clause, phrases introduce clauses, and the relations between them also need to be dealt properly. Thus, the subject of whether these complex sentence structures can be identified and transformed into Target Language (TL) with correct word order has become the restricting factor of the MT development. Complex long sentences can be seen everywhere in English, many of these long sentences can even be the whole paragraph. The result of long sentence analysis can impact the qualification and the readability of the MT. On the other hand, the demand for translation service is facing the severe insufficient circumstance, and this kind of insufficiency really may affect the capabilities of capturing useful information. Thus, the first feasible solution to this problem is using MT to fill the vacancies. Spontaneously, the problem of how to analyze long sentences has become the key problem of fulfilling the Informationization Strategy in China.

In this thesis, the history of MT and some basic translation methods were introduced at the very beginning, and then introduced the details of present MT systems' sentence parsing methods. Based on all the above, in the last part of the thesis long sentence parsing system

structures and several significant difficulties existing in current MT studies are discussed in detail. Especially in the Chapter 3, in which long sentence parsing is fully discussed, and thesis emphasizes on the problems of the basic principles and sentence parsing methods, of defining and identifying the long sentence and of difficulties in analyzing long sentence. Then MT dictionary and its structure are discussed in the following part, and the thesis also proposes some improving suggestions to the designs of MT dictionaries. At last a sample sentence was used to demonstrate how MT system works with an improved dictionary. In the final chapter, some questions that are encountered during the studying of this subject are discussed.

Key words machine translation long sentence sentence parsing FAHQMT Right-to-Left algorithm

摘 要

机器翻译在经历了长达几十年的发展和演变之后，随着信息技术和计算机技术的不断提高，对机器翻译的研究已经由当初简单的语言学与计算科学相结合的学科慢慢转变为融合了语义学、数学、语料库、计算科学、人工智能和生物科学的一个综合性研究领域。然而时至今日，机器翻译的译文质量依然不能达到人们所期望的水平。尤其是在长句处理问题上，尽管计算机等相关科学已经较十几年前有了质的飞跃，然而长句处理这一问题却依然是机器翻译研究领域中一个无法逾越的障碍。

长句的处理不同于词的查询。词的查询只需要对所查询词进行词例还原并将还原词例与数据库中的条目进行比对搜索就可以轻易找到对应的译文，而其中词例还原所需要涉及的转换规则对于现有的技术水平来说是十分容易的。相比之下，要想进行长句的分析处理，则不仅需要进行词一级的信息比对，而且还需要对句子中的各种结构进行区分和关系辨别，这其中涉及到了歧义处理，词的远距离相关，句相关等一系列的问题，诸如从句中包含短语，短语中含有从句，主、从句被插入语隔开等情况。而所有这些复杂的句结构关系能否被成功识别并通过规则成功转换为正常语序的目标语言正是制约机器翻译发展的一个关键性课题。因为在英语的各类文献中，长达数行的英语长句极为常见，如果长句处理不好则会严重影响机器译文的质量和可读性。 而随着我国经济和社会的日益发展，巨大的外文资料翻译需求面对的却是翻译力量的严重不足，这就严重影响了快速获取信息资源的效率。而要解决这一问题，机器翻译无疑成为了目前比较可行的方法！因此，机器翻译系统中如何正确处理英语长句的问题就成为了我国信息化战略的关键性问题。

本文主要研究内容就是从对机器翻译的发展历史和其基本方法的研究入手，进而对现阶段的机器翻译系统中句分析方法进行剖析，而后对长句分析的系统结构和相关关键问题进行探讨。本文的第三章从长句分析的基本原则和方法、长句的定义和识别、长句分析难点和基本处理方法以及现有机器翻译系统中长句处理方式入手，对长句翻译的机器字典设计和结构进行了详细的分析研究。本研究还针对机器翻译系统中最重要的资源性结构——机器字典以及长句分析算法进行了试探性的改进，并通过实例展示加以说明。论文的最后一部分则主要阐述了研究者在进行相关学习和研究过程中对机器翻译所

持有的一些观点和态度，同时对机器翻译的发展趋势及其可能产生的影响进行了尝试性的分析和预测。

关键词 机器翻译 长句 句分析 FAHQMT 自右向左寻首式分析算法

Contents

Chapter 1 Introduction

1.1 Preliminary Remarks

As we have stepped into Information Age, language, the information carrier, which has become the most significant means for human to communicate since we began to talk, have been considered as the barrier of communications between people from different cultures. The problem of converting a language into another quickly and efficiently has become a problem of common concern for humanity.

It is no exaggeration to say that almost everyone, directly or indirectly, needs the translation services. No translation, no communication! But the truth is, as the contacts between countries have been increasing rapidly for decades, the literature of science and technology is still growing now, the capacity of human translation has been far from meeting this enormous need for translation service. For example, in China, the translation market need, with the scale of 10 billion Yuan per year, only can be digested about 1/10.

And at present, the problem of how to obtain up-to-date information as quickly as possible has become the key to win the drastic international competition for our country. Under this kind of situation, Machine Translation can be the solution.

Machine translation, the abbreviation is MT, is a kind of computer-based automatic translation. It directly belongs to the field of Computational Linguistics, which investigates the methods of using computer software to translate one natural language, either text or speech, into another one. At the very basic level, MT can change words of one language into another. The best use of MT at the very beginning is to help people to find out the single words' meanings, the best sample for this kind of use that can be seen in our daily life is Electronic Dictionary.

On the surface, this kind of translation method seems very easy and simple, and of course, it can be efficient. In fact, the truth is far beyond our imaginations. For we all know that we can't translate an article only by substituting words into another language. A translator must interpret and analyze all of the elements in the text and know about how words influence the others. This requires extensive expertise in grammar of both TL and SL, in syntax (sentence structure) and semantics (meanings), etc. But this doesn't mean that human translator may be better than Machine Translation Systems. The challenges are equal! For example, with the

same text, to get identical translations from two human translators is impossible, and different translators may have different translation ideas, according to their own likes and dislikes, or even to meet some special purpose with the work. Then barriers for the people who come from different cultures have been created. But this is only a tiny problem, the greater challenge lies in how to make machine translation systems produce publishable quality translations, which has no need to be modified and can be published directly.

Actually, MT is a kind of organic system which combined various subjects and technologies in it. It is not only related to computer technology directly, but also is inextricably linked with mathematics, philosophy and literature, etc. Maybe in the future, this subject may also build up the relationship with biology and the Brain Science.

1.2 The History of Machine Translation and Research Status in China

1.2.1 The History of Machine Translation

The beginning: The exact concept of the MT may date back to the 17th century. R. Descartes and G. W. Leibniz, the French and German philosopher and mathematician, both of them had puts forward the imagination of using the mechanical dictionaries to overcome barriers in languages. But unfortunately, their ideas hadn't been realized but just stayed in principle design level.

About the mid 1930s, one proposal of using paper tapes and bilingual dictionaries to translate single words was put up by a French engineer, whose name is Georges Artsrouni. That machine was entitled Machine Brain. It uses two paper tapes to record information. One tape records words from Source Language (SL) and the Target Language (TL), the other one records code of the words, one word one code, and these codes were recorded by holes in the paper. People used key boards to punch holes, and of course use these boards to find the words. It will take 10 to 15 seconds to search one word. Unfortunately the World War II broke out, Artsrouni and his Machine Brain was ruined by the war.

From 1937 to 1946, anyone paid attention to the MT research, till to the year 1946, the first computer was created; the English engineer A. D. Booth proposed the idea of using computer to do automatic translation. The concept of MT, which is similar as today's had been gradually taking shape since then.

The Pioneering stage: In 1954, about 8 years after the first computer was invented, cooperated with IBM, the American Georgetown University carried out the Georgetown-IBM

experiment, which is based on the IBM-701 computer technic. The system has no more 250 words and may translate just 49 selected Russian sentences into English — most of those words were from the field of chemistry.

That was the very beginning of the Fully Automatic High Quality Machine Translation (FAHQMT or MT), and the translation demonstration was also reported in the newspapers, and the public was interested in this new technology. Although the system was so simple that it only can be considered as a "toy" with today's standards, the wave of MT began to sweep over the world just since then.

But the good days didn't last too long. In 1964, US National Academy of Sciences founded up the ALPAC (Automatic Language Processing Advisory Community) to investigate the possibility of MT. The major indicators of the investigation included the translation speed, quality, the cost and the demand for the MT. The investigation lasted for about two years. In 1966, the ALPAC Report was made out. The conclusion about MT in this report was "There is no immediate or predictable prospect of useful Machine Translation."[ALPAC, 1966:34], and it also pointed out that the research had met the impassable "semantic barrier".

After this report, researches on MT cooled down suddenly, and around the world, the governments cut down the research money rapidly, from then on the MT research had been stayed in standstill and abandoned for more than a decade.

The recovery of MT: In 1970s, for the great development of computer science techniques, linguistics and Natural Language Understanding (NLU) begin to regain more and more attention once more. Compared to the former researchers, in this time scientists were no longer to be unrealistically optimistic. Computer scientists begin to borrow ideas from linguists and spend much more time and energies to learn about the language itself. The MT research was more rational than before, the idea that MT can replace human translation had been abandoned.

In this period, MT research acquired series of significant achievements. Researchers poured their energies on the problems of how to use computer may assist people to correct or to improve the translation, but not on how to use computer to translate automatically. The Computer-aided Machine Translation System, Human-aided Machine Translation System and Restricted or Controlled Language Translation System were the fruits of this period.

In 1976, the University Montreal, Canada, invented a practical translation system, TAUM-METEO, which served for the weather forecasts. The most significant meaning of this system was not only the performance of its high speed; which may translate 60,000 – 300,000 words in a hour, but also because the translation can be published directly without any post-editing. METEO's success can be considered as the milestone in the history of the MT's development. It stands for the MT has stepped into a new age; it also proved the MT can satisfy the practical needs.

The prosperous times: During 1980s, as the great development of computer science, the linguists also achieved many remarkable achievements; corpus-based machine translation, statistics-based machine translation (SBMT), and exampled-based machine translation were all created in this period.

In the mid 1990s, the empiricism and rationalism was brought into the MT, and this leaded the MT into a new stage directly. Meanwhile, the corpus began to be used in the rule-based MT technique, and scientists begin to mount higher summit of the MT.

At present, many MT systems are working, such as the SYSTRAN system, the METAL system, the WEIDNER system, the ATLAS system, the TRADOS system and the SMART system, etc. Although we may see so many samples of MT, no one system can take the dominating position till today.

After more than 50 years development, although there are still many disabilities, comparing to the former MT systems, we have made great progresses on accuracy, readability and the speed. More and more researchers begin to realize that the dream of using machine to replace human translator may be realized someday in the future, but there is still a long journey we need to cover.

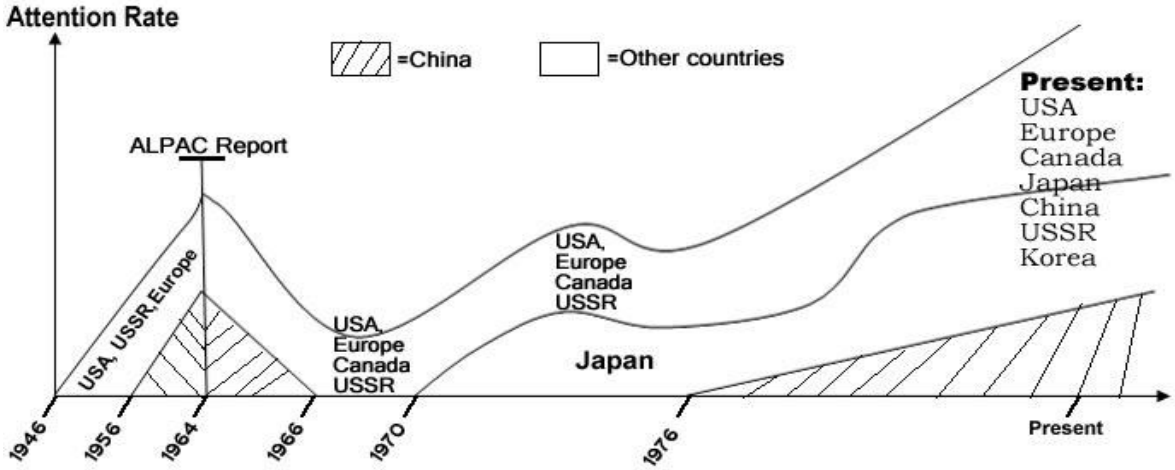

Table 1

This chart reminds me F. D. Saussure, the father of the modern Linguistics had pointed out "…language being what it is, we shall find nothing simple in it regardless of our approach; everywhere and always there is the same complex equilibrium of terms that mutually condition each other."[Saussure, 1959:122]

1.2.2 Research Status in China

Our country is the 4th country that began to carry out the MT system research in the world. As MT is the essential part in realizing the automation of collecting science and technology information, the MT research had been included in the national science development plan since 1956.

In 1957, the research of MT was formally initiated, and the main objective was Russian-Chinese and English-Chinese translation system. After 2 years, in September 1959, using the first high-speed mainframe digital computer built by ourselves, our scientists successfully accomplished the first Russian-Chinese translation experiment.

In that experiment, the translation dictionary only had 2030 entries, and 29 line diagrams to compose the grammar rules. The system translated 9 different types of complex Russian sentences into Chinese, the special trait of this experiment was that the Chinese output mode was not Chinese characters but was code, which was recorded in perforated tapes. Because in 1950s, the computer still can't input Chinese characters.

In 1966 the Cultural Revolution broke out, and it lasted to 1975, the study about MT halted for about 10 years. From 1975, the ice began to melt. After only 3 years, in the second half of 1978, Heilongjiang University established an English-Chinese Machine Translation Group and began to cultivate MT researchers.

In 1981, Chinese researcher Feng Zhiwei built FAJRA System, and carried out a series of experiments, included Chinese-French, Chinese-English, Chinese-Japanese, Chinese-German and Chinese-Russian. That was the first Chinese to foreign languages translation system designed by our Chinese scientist.

In 1983, the first Chinese Distinguishing Word System (CDWS) was created; 1985, Northeastern University of Technology built CTUS (Chinese Text Understanding System); 1988, Tsinghua University proposed the Multi-scan determination algorithm and built TUSMI system and CAAMS (Chinese Ambiguity Analysis Model System).[1]

1988, the first commercial-purpose MT system “译星 1 号” finished. 1993, the commercial Chinese to foreign languages translation system Sino Trans came into service; 1994, the English-Chinese translation system Martrix commercialized. Today’s many translation software and on-line translation service are based on these achievements, more and more universities and companies begin to involve in MT research, and the MT system is being improved rapidly.

In resent years, the English-Chinese translation theory is being perfected without any pause, but there is any mature translation theory can be used to guide English-Chinese MT.

For the MT can’t be completed only by computer programmers, but also need corporations between English linguists and Chinese linguists. And it also need to be supported by many interdisciplinary subject experts to support its research.

But most of the English and Chinese linguists in China have no communications between each others, this situation restricts the development of MT seriously in our country, now, for this reason, we still have any efficient rule can be used in the English-Chinese MT system.

By now, the MT system, most of times in China, is still being considered and being used as Electronic Dictionary. The potential of the English-Chinese MT system is still covered in the dense fog.

1.3 Significance and Objectives of the Study

Every time when people refer to the MT’s value, the question that will be asked for most of time is “Can it translate *Bible* into Chinese?” The answer to this question is NO!! There is absolutely misunderstanding about MT. The purpose of making MT system is not to translate literary works. At least it is not the purpose or we may say it is impossible at present.

In this modern society, our needs for translation are increasing by enormous amount day by day. The files that are waiting to be translated come from various filed. It is impossible to translate everything into high quality translation. We really can’t afford so much money or time to do this. Most of time, we only need the basic meaning of the articles but not to appreciate the poetic imagery of the translation.

The reason is when we human translate an article, we are not only check the meanings of each words from SL (Source Language), but also we can organize the TL (Target Language) freely and chose the better expressions, all of these are all based on the comprehension about

languages. Computer can't perform the same thing as we human now.

Actually, we are not saying that MT is totally useless in translation jobs, the value of the MT systems exist in the following works:

1. To fulfill some initial translations, people may get main ideas of an article as soon as possible.

2. It can be much smarter and more useful than just a simple textual search, especially in the cross-language information searching jobs.

3. It may provide high-quality translations in sublanguage translation jobs, such as the localizations of the instruction manuals, the economic or political news and weather reports, etc.

For these reasons, anyone can ignore the existence of MT, and as early as 1990, the MT system METEO was capable of translating English meteorological data into French with the speed about 45,000 words per day. And till today the METEO system is still being perfected and offering service to the scientists. There are also many other MT systems, such as SYSTRAN, LOGOS, ALPS, ENGSPAN, etc, and all of these systems are still running.

MT is very meaningful. It is a sophisticated and comprehensive systematic engineering. It is a cross subject technology, which is widely supported by Linguistics, Psychology, Psycholinguistics, Brain Science, Computer Science, Philosophy, Logics, Artificial Intelligence, Mathematics, Information, Literature, Arts.

Although we can't use "perfect" or even "good" to describe the MT translations, but it is still can be useful in practice. The most important thing is how to eliminate those flaws but not to criticize it. In this Information age, MT is the necessary tool for people to overcome the language barrier. It must be the most efficient method to realize multi-language communication in future.

Objectives of This Thesis

During the studying, I found that the quality of the translation text is always enslaved by the same problem, which is the English long sentence analysis.

For this reason, the main purpose of this paper is to find methods to improve the quality of the long sentence parsing and the significant Mechanical dictionary

In the end of the paper, the value of this subject and its development tendency will be discussed.

Chapter 2 A Survy of Literatures on MT and Sentence Parsing

2.1 Some Basic Machine Translation Methods

Traditionally, the MT methods can be classified into two categories, Rule-based translation method and Corpus-based translation method. Rule-based translation systems can be divided into three catalogs: Direct Translation System, Interlingua Translation System and Transfer System. Traditionally, Corpus-based translation method can be divided into two different classes: the Statistic-based translation and the Example-based translation method.

In addition to the upper two main categories, more and more researchers have been paying much more attentions to the Multi-engine translation method. In the following all these translation methods will be introduced in details.

2.1.1Rule-based Translation System

All three different rule-based systems have one common feature; all of these systems are driven by a symbol system, which is composed by language knowledge base, to fulfill translation jobs.

1. Direct translation method: Always this is called literal translation, word-based translation or dictionary-based translation. The direct translation method is that the words will be translated as a dictionary does — word by word, usually without much correlation of meaning between them.

Dictionary lookups may be done with or without morphological analysis or lemmatisation. This kind of machine translation is probably the least sophisticated, direct machine translation is ideally suitable for the translation of long lists of phrases on the sub sentence (i.e., not a full sentence) level, e.g. inventories or simple catalogs of products and services. [3]

For this reason, there is an obvious flaw in system. Simple word-based translation can't translate between languages with different fertility (the ratio of the lengths of sequences of translated words is called fertility, which tells how many foreign words each native word produces). For example, it is easy to translate the “hybrid firewall” into Chinese **“**混合防火墙**”,** but impossible to translate the sentence “Microcomputer is very small in size” into Chinese.

For these reasons the word-based translation has not been accepted widely today.

2. Interlingua-based translation system: In this system, the source language will be transformed into an Interlingua firstly, which is an abstract language-independent representation. Then target language will be generated out according to the interlingua. The Interlingual system can be considered as better alternative choice, specially compared to the direct approach and the transfer approach. As shown in Table 2:

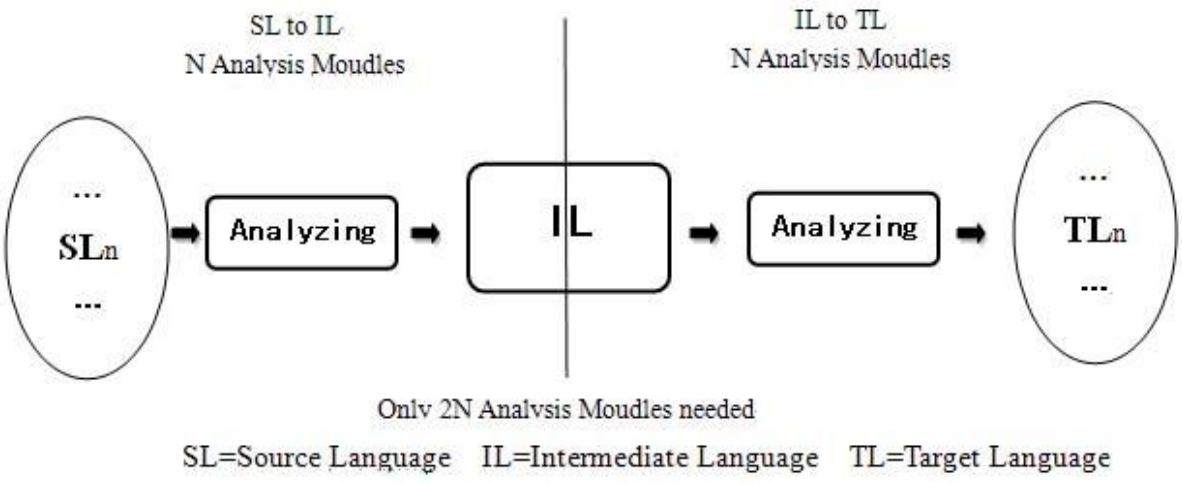

Table 2

The Interlingua-based translation method also has its own advantages and disadvantages. The most significant advantage of this system is that it provides an economical way to realize multilingual translation. With interlingual system, it becomes unnecessary to make translation pairs between each pair of languages in the system. Instead of creating **N*(N − 1)** language pairs (**N** is the number of languages in the system), the system only needs **2N** pairs between the **N** languages and the only interlingua.

The significant disadvantage is that it is too difficult to define a kind of proper Interlingua, which has been agreed by most of linguists now. The ideal SL for Interlingua-based machine translation is limited in a very narrow specific domain.

In the end, after being practiced for a long time, the conclusion about this kind of translation system is“if a meta-language (an interlingua) were to be used for translation purpose, it would need to incorporate all possible features of many languages. That would not only be an endless task but probably a fruitless one as well. Such a system would soon become unmanageable and perhaps collapse under its own weight.”[2]

3.Transfer-base machine translation method：This translation system is quite different from the Interlingua-based system. The Interlingua-based systems use independent

intermediate language to represent a pair of languages, but the Transfer-based systems apply sets of linguistic rules, which are defined as correspondences between the structure of the source language and that of the target language.

The first stage of the translation is to analyze the input text for morphology and syntax (and sometimes semantics) to create an internal representation. By using both bilingual dictionaries and grammatical rules, the translation will be generated out from this representation. As shown in Table 3:

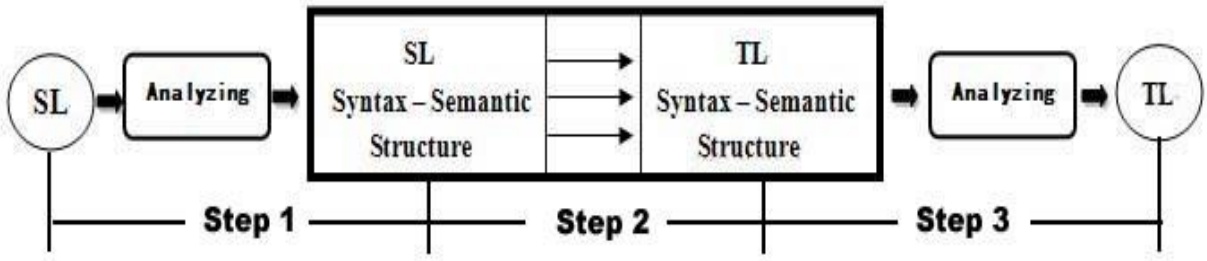

Table 3

Thus, the key problem of this kind of system is how to use dictionaries and rules to realize the syntax and semantic transformations between two languages. As is shown in the picture, Morphological and lexical analysis, such as part-of-speech tagging, happen in step1; lexical and structure transfer happen in step2; and Morphological generation happen in step3.

According to the different analyzing depth, the transfer system also can be divided into two different levels.

The first level is Syntactic Transfer. We often call it Superficial transfer, which may transfer "syntactic structures" between the source and target language. It is very suitable to deal with languages that are from the same language family or of the same type. For example, the quality and speed of the transformations between Romance languages are good, such as English to French, Italian to Spanish, etc. But it is inadequate when it is applied to translate English into Chinese or Japanese, etc.

The second level is Semantic Transfer, or Deep transfer. In this kind of transfer system, the systems often construct semantic representation which is based on SL at first, and then using this created representation, the systems will generate out another series of structure, which represents the meanings of both SL and TL.

Thus, the translation quality relies on the analyzing depth directly. Deeper analyzed, higher quality will be gained. It is possible to get high quality translations with this strategy, and the accuracy of translation can reach 90%, but it is confined to the distance between SL

and TL.

For all the reasons we mentioned in before, this kind of translation theories and practical systems are the successful systems which are being used widely in nowadays.

2.1.2 Corpus-based Translation

At present, the Corpus-based machine translation could be divided into two main classes, the Statistic-based MT and Example-based MT.

For the theory of Statistic-base MT considers the language as a kind of meaningless string, it is not appropriate in analyzing the semantic information and reconstructions of the translations. The results of the experiments are far away from our imaginations. And there are still many problems are waiting to be eliminated, such as the problem, which was proposed by Chomsky, of how to deal with long-distance constraints like subject-verb concord, the Statistic-based MT has been proved that it is not very effective in practice.

1 The Example-base MT (EBMT)

Comparing to the unsuccessful Statistic-Based MT, the Example-based MT (EBMT) has shown us perfect performances, either in experiment or in practice.

The first linguist who proposed EBMT is Japanese linguist Makoto Nagao. His paper that was published in 1984 “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle” can be seen as the start point of the research about EBMT.

It imitates the thinking process of human translation. Makoto Nagao believes that the first step of translation work in human minds is breaking a sentence into phrases and words. The purpose of translating these different parts is to compose these fragments into one long sentence again. Phrasal translations are performed in seeking similar words or phrases previously. The principle of this kind of translation system is encoded to example-based machine translation system, which use enormous Case base as the foundation of the system.

The translation process of this system is in the following. First, a bilingual alignment case base will be established. Input the sentence SS, the system will search for the most similar sentence SS’ from the Source language base; when the SS’ is sought out, then according to the translation TT’, some words or phrases in TT’ will be changed and the translation TT will be done.

The process has been demonstrated in Table 4 and Table 5：

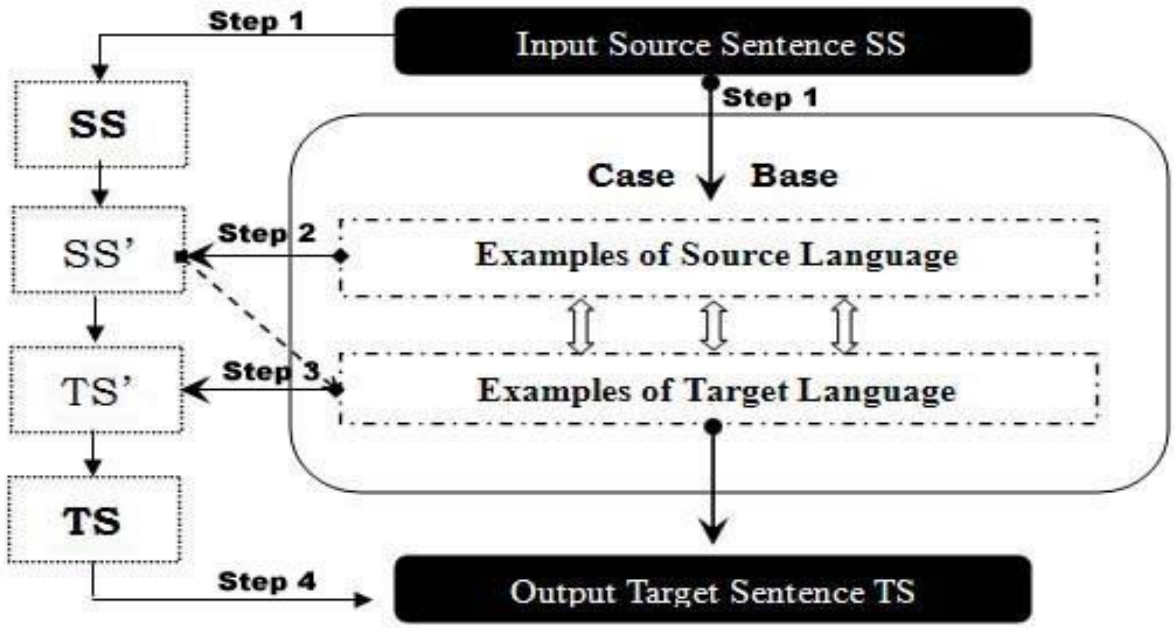

Table 4

For example:

Table 5

In practice, although the EMBT translation quality may be satisfying, there are still three problems are waiting to be eliminated.

1. Alignment of Bilingual text. Sentence alignment is not the only requirement, most of time the phrase alignment or even lexical alignment is also needed necessarily.

2. The measurement of similarity is hard to be defined. For this reason, many ambiguous selection items will be listed out, and this will slow down the system and spoil the translation quality finally.

3. EBMT does not emphasize the semantic analysis, the translation, which is composed by replacing or deleting words in example sentences, always loses some SL's necessary information.

2.1.3 Hybrid Machine Translation

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies.[4]. The approaches differ in a number of ways:

In HMT, the system uses a rules based engine. Statistics are then used to adjust or to correct the translations out from the rules engine.

Meanwhile, statistics is also guided by rules. Rules are used to pre-process data to guide the statistical engine. When the statistical engine finished its work, rules are also used to post-process the statistical output to perform functions such as normalization.

We may see that the HMT is much more powerful than the other systems, the most famous HMT sample system is the PANGLOSS Mark III, which was built in 1992.

Mark III has three translation engines in total, Knowledge-based engine, Lexical-based engine and the Example-base engine, its structure same as to Picture 2.5:

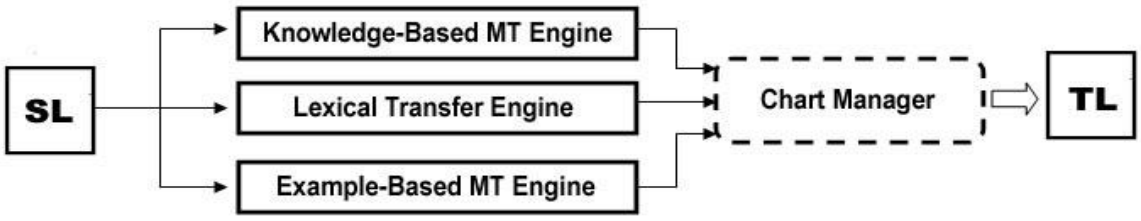

Table 6

When the system acquires the SL, three engines begin to analyze the SL independently at the same time and to convey results into the rule-based chart manager. The chart manager will perform functions and output the best translation text.

The HMT may avoid most of disadvantages existing in the other systems, but it is also facing to the similar problems that belong to the other systems.

2.2 Basic Principles of Sentence Parsing and Some Current Methods

The various sentence structures in articles means the quality of the sentence translations is the most significant factor in the machine translation system. In the following, how computers perform the sentence analysis in practice will be discussed in details, and several sentence analysis methods will be introduced in the second part of this chapter.

2.2.1 The Basic Principles of Parsing and the Process

In practice, the sentences are composed by different syntax elements, so the analysis can’t be accomplished for only one time. It is necessary for the system to divide the sentence

parsing process into stages to analyze the meaning gradually. The core of sentence parsing is to separate complex grammatical structure into different individual segments.

In the following, we will take the ECAT System as the example to illustrate the details of sentence parsing and its process, which is designed by the Chinese Academy of Social Sciences,:

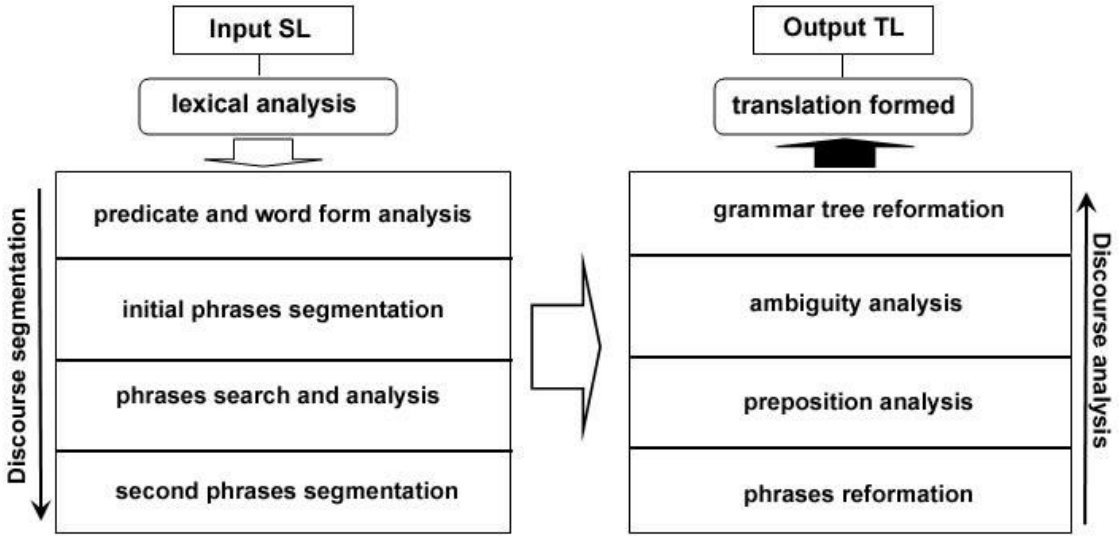

Table 7

Stage One: The discourse segmentation:

Step one, received the SL, the system begins to carry out the lexical analysis to identify the basic unit; step two, the system begins to analyze the predicate and forms of words; step three, to carry out the initial phrases segmentation; step four, according to the results of step two and three, system begins to search for the similar information in data base; step five, accomplished the information search and comparison, system will repeat step three.

Stage Two: The discourse analysis:

Basing on the results that were worked out in stage one: step one, rebuild the phrases; step two, analyze prepositions of the sentence; step three, polysemous words will be analyzed; step four, according to the TL's grammar rules, begin to reconstruct the grammar three.

In the end, output the translation text.

In the following, I will take a sentence as example to demonstrate the process of the sentence parsing system. To cope with the conventional sentence structure expressions, we use the Chinese Corpus Segmentation and Part of Speech Tagging System [俞士汶 1999] as the tagging symbol system.

Symbols are listed in the following form:

Structure name	Mark		Structure name	Mark		Structure name	Mark	
动词短语	vp	荣获冠军	副词短语	dp	不情愿地	介宾结构	JB	按照工程进度
名词短语	np	战斗意志	小句	dj	待人热情	的字结构	DE	无情的
数量短语	mp	三十岁	介词短语	pp	按照这种逻辑	状中结构	ZZ	从实质上看
处所短语	sp	中国大陆	定中结构	DZ	魅力三峡	连谓结构	LW	请他当头儿
时间短语	tp	建国初期	述宾结构	SB	犒赏功臣	主谓结构	ZW	疾病肆虐
形容词短语	ap	绝对可靠	述补结构	SBU	洗干净	联合结构	LH	研究讨论

Table 8

First, input the sentence: “He threw the book to his brother.”

Step One: lexical analysis

He = he; threw = throw; the = the; book = book; to = to; his = his; brother = brother

Step Two: syntax analysis

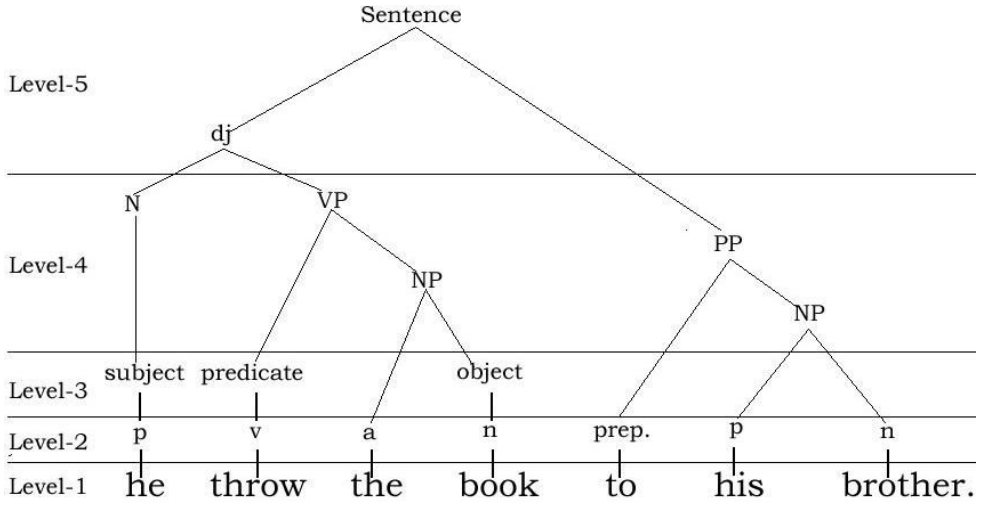

Table 9

Step Three: structure transformation：If the system transfer the sentence structure only by the simple rules, but not to perform the semantic analysis process directly, the result will be similar to the following demonstration：

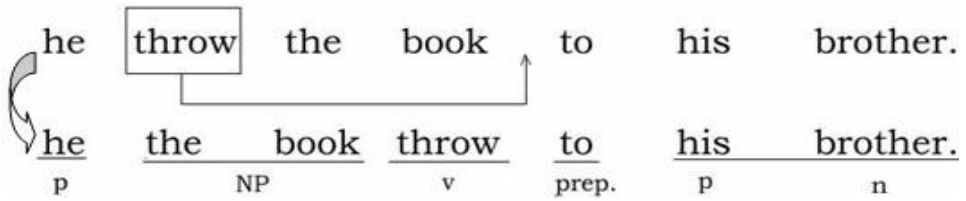

Table 10

Step Four:

he the book throw to his brother. ⇨ He the book threw to his brother.

Step Five: TL words selection:

He = 他 ; book = 书; threw = 扔; to = 给; his = 他的; brother = 兄弟

Step Six: In this step the system, according to the TL's grammar tree, will carry out Grammatical transformation, necessary information will be attached into the translation, and the result will be corrected for the last time: "He 把 book threw to 了 his brother."

Step Seven: Adjust the terminal text and replace all words and phrases into TL："他_把_书_扔_给_了_他的_兄弟。"

Step Eight: Output the translation text: in this step, some necessary adjustments, such as deleting spaces from the text to adjust to the written form of the TL, will be performed. And then the final result will be: "他把书扔给了他的兄弟。"

Although the process is easy to be described on paper, it is still too hard to be realized in the form of computer program in practical situations. We may see this easily according to the following sample:

[1]Text-to-Speech Translator："他把书交给他的弟弟。"
(Address: http://www.oddcast.com/home/demos/tts/tts_tran_example.php)

[2]Yahoo online-translator："他投掷了书给他的兄弟。" (Address: http://fanyi.cn.yahoo.com)

[3]Jin Qiao online-translator："他投掷方面的书对他的兄弟。"(Address: www.netat.net)

[4]Microsoft online-translator："他把这本书给他的弟弟。"
(Address: www.microsofttranslator.com)

I also tested a special interlingual translation, the English-Japanese-Chinese translation, in which the Japanese was used as the intermediate language between English and Chinese. This test was performed by the on-line MT system, which is designed by Japanese experts (http://honyaku.nifty.com)：

The text (SL): He threw the book to his brother.

(1)English-Japanese: 彼は本を兄弟に投げました。

(2)Japanese-Chinese: 他对兄弟投了书。

According to all of these translations, the main difference between different MT systems exists in the process of lexical analysis. The results of semantic and grammatical

structure analysis are insufficient. Especially comparing to the TL, MT translations take on so dramatic changes and mistakes in semantic and grammar structures that readers can't get the original meaning of the SL, if ask them to read the translations only.

For this reason, the sentence analysis in MT can be considered as the most significant barrier in MT research. Whether MT system can succeed or not is rely on the solution of this problem.

2.2.2 The Definition of Long Sentence and Its Classifications

The long sentences which run to several lines can be found in English articles. Comparing to the simple sentences, long sentences always have more than one sub clauses and Non- predicative Verbs. The relations between these clauses and verbs are always confusing. Ignoring any interrelation will directly lead readers into wrong comprehension about the entire meaning of the whole sentence.

Long sentence may be very helpful for us to express complex and accurate ideas, but it also may bring terrible troubles to MT system. The problem of how to analyze long sentences correctly is still remained to be solved

1. Complex Sentence

The structure of the complex sentence is "Main Clause +Sub Clause", the main clause is similar to simple sentence, eg. The fact is….. And there are always several sub sentences behind the main sentence to be the grammatical constituents. The relationship between the main and the sub is narrow. Without understanding the sub clause' meanings, to understand the sentence meaning will be very difficult. And the complex sentence can be divided into three types: the Attributive Clause, the Adverbial Clause and the Noun Clause.

Example:	She is the girl	who I saw in the garden yesterday.	
	Main Clause +	**Subordinate Clause**	**= Sentence**

Table 11

2. Compound Sentence

In one compound sentence, there are two or more simple sentences. And all these sentences are connected by coordinating conjunctions, such as and, but, or, so, etc. The structure form is "…simple sentence+ coordinating conjunction+simple sentence…". The significance of each simple sentences are equal, the relationships between them are parataxis but not hypotaxis.

Example: I'd better take an umberllar, for it is going to rain.

Main Clause + Conjunction + Main Clause ⊨ Sentence

Table 12

3. Mixed Long Sentence

In practice, to express one or more subjects clearly, people always mix compound sentence and coordinating sentence together. In these structures, the sentences have more than one parataxis structures, and some of these parataxis coordinating sentences have their own hypotaxis structures.

Example:

The editorial director of one dictionary says that the noun "clonee" may sound like a good term, but it's not clear enough.

Sentence Structure:

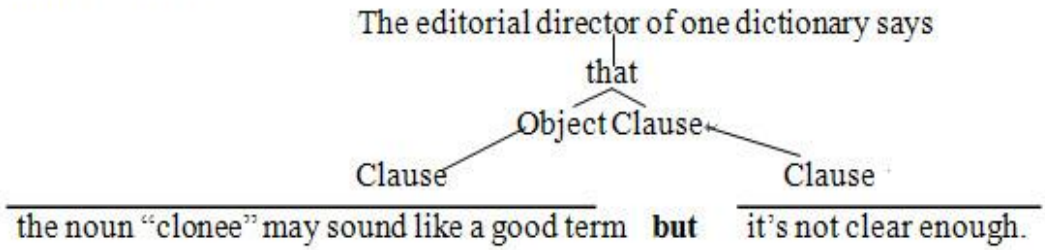

Table 13

2.2.3 The Difficulties in Long Sentence Parsing

English sentence basic forms can be concluded into the following five types:

Basic Sentence Patterns：All of the other sentences are composed by these five structures:

1）SV: Subject + Verb e.g.：I think. 我思考。

2）SVP: Subject + Verb + Predicative e.g.：He is student. 他是学生。

3）SVO: Subject + Verb + Object e.g.：She studies English. 她学英语。

4）SVOC: Subject + Verb + Object + Complement

e.g.：Time would prove you wrong. 时间会证明你是错的。

5）SVOiOd: Subject + Verb + Indirect-Object + Direct-Object

e.g.：Her mother made him new shoes. 她妈妈给他做了一双新鞋。

The truth is that to build up rules to distinguish all there 5 simple sentence structures is very easy, but to practice these rules in MT system will be difficult, especially when these five structures appear altogether in the same sentence, and be connected by relative

pronouns to impact each others meanings. Especially the task that how to identify which sentence subordinate to the main and which is not is too hard to be resolved.

And the details about these difficulties are in the following:

1. How to Identify Main Clause and Sub Clauses

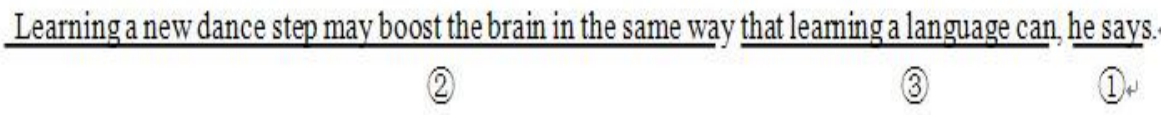

Table 14

In the example sentence, the true logical order is ①-②-③； the main clause ① is at the end， and all structures can be seen as the objects of the main verb “says”, the attributive clause, which is introduced by “that”, is the clause of the sub clause which is introduced by “learning”.

Then, the questions that how to make the computer to define the relationships between different clauses, how to make the system to adjust the clause into the TL’s order correctly have become the highest summit to overcome till now.

2. How to Identify the Correlation Between Clauses

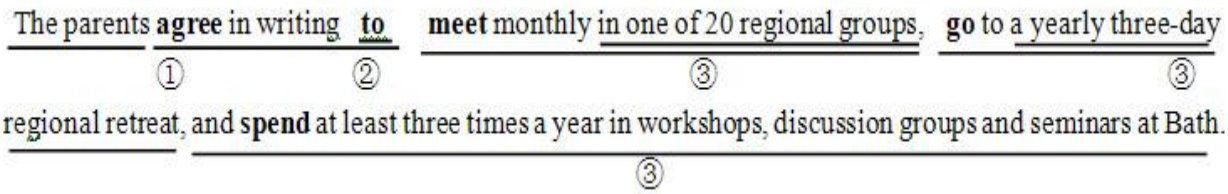

Table 15

In the sample sentence, the predicate and its combined preposition are not stay together, actually the preposition “to” is combined with 3 different VP (verb phrase).

How to make the MT system to judge the sentence as the structure of “①+（②+③）+（②+③）+（②+③）” is the key problem, too.

3. Problem of Identifying Conjunctions Functions and Its Translation

For the conjunctions are the key elements in composing the long sentences, its functions can be concluded as following:

a. To connect the words and phrases into sentence.

e.g. Now I must go ***or*** I shall be late for the party

b. To be the start sign of the sub clauses:

e.g. He phoned me ***that*** he wanted to see me.

c. Used to refer to the other parts of the sentence, and sometimes also syntactically connect sentences or elements.

e.g. He saw the woman ***who*** has a red car.

If the MT system can't process conjunctions correctly, the translation quality will be influnced significantly. For example:

	MT translation:	**Human translation:**
a.	现在我必须去或我为党将是晚	现在我必须走否则的话我会迟到。
b.	他给我打电话，他想看我（并列结构）	他打电话说他想见我一面。（宾从结构）
c.	他看到这个女人谁有一辆红色轿车。	他看见了有一辆红色汽车的妇女。

4. The Sentence Structure Long-distance Correlation Analysis

In English, the sentences, which should have been locked together tightly, are sometimes divided by other syntactic elements, such as parenthesis or other clauses. For this reason, to pay much more attentions to the semantic relation is necessary, but MT translation system still cannot solve this problem as perfectly as human.

Example:

Concrete pillars for highway bridges， as we all know, that previously only had steel rods inside

① ② ③ ④

are now enclosed in steel.

⑤

Table 16

MT system's translation：高速公路的 bridges，众所周知混凝土桩，仅那以前有的钢标尺里面在钢现在被附寄。

Actually, the appropriate structure relation is: ② qualifies ①，④ qualifies①，⑤ and ① composed an passive relation.

So the reasonable translation structure order should be : ③，④+②+①+⑤

And the translation is:正如我们所知，那些以前只不过里面仅有些钢筋的公路桥混凝土支柱现如今都被包裹在了钢铁之内。

How to distinguish the Modifier and the Headword, and how to merge them into translation, till then, all of these two problems haven't been solved perfectly by any existing MT system.

5. How to Identify and Translate the Nominative Absolute

There is one special structure form in English, which is called nominative absolute. Its

structure formation is：

Structure 1: Pronoun/Noun + Participle：Pronoun or Noun is the logical subject of the structure.

Structure 2: Pronoun/Noun + being + adjective phrase
+ adverb phrase
+ participial phrase
+ infinitive

For the nominative absolute has its own independent subject which is different from the main clause, and it is always divided from the main by comma, so the MT systems have been performed inefficiently on identifying and analyzing the semantic meanings of the nominative absolute.

Examples:

1. The condition being favorable, he may succeed.

MT translation：目前有利的条件，他可能会成功。

Human translation：若条件有利，他或许能成功。

2. She had to walk home with her bike stolen.

MT translation：她走在她的被盗自行车回家。

Human translation：因为她的车子被偷了，她只得走回家。

6.How to Process the Inverted Word-order, Ellipsis and the Emphasizing Structure

The problems of how to process and translate inverted word-order structures, ellipsis sentences and the emphasizing sentence structure, are still being puzzles to scientists. And the difficulties are not only how to correct the word order, but also related to identifying the logical subject of the main sentence.

Example1:

Down jumped the man from the second floor when the policeman pointed his pistol at him.

MT translations:

（1）Yahoo online-translator：“下来，当警察把他的手枪指向他，跳跃了从二楼的人。”

（2）Jin Qiao online-translator：“当郡跳从当警察的时候三楼男人削尖他的手枪阿特他.”

（3）Microsoft online-translator：“下跳该名男子从二楼当警察指着他的手枪。”

Human translation:当警察用手枪指着这个男人的时候，他从二楼跳了下去。

Example2:

So unreasonable was his price that everybody startled.

（1）Yahoo online-translator：“很不合情理的是使震惊大家的他的价格。”

（2）Jin Qiao online-translator：“因此，不合理的是他的价格，大家都吓了一跳。”

（3）Microsoft online-translator：“如此不合理是他的价格，每个人都吓了一跳。”

MT translation：因此，不合理的是他的价格，大家都吓了一跳。

Human translation：他的要价太离谱以至于令每个人都瞠目结舌。

Now, we may see clearly that MT systems can’t identify the semantic meanings of pronoun, it also can’t identify the subject of the inverted structure. One of these systems even made mistake in identifying the part of the speech. For this reason, it is quite necessary to improve the algorithm analysis on this kind of structure.

2.2.4 The Basic Idea of Processing the Long Sentence

The processes of analyzing long sentence are equal to the processes of semantic analysis and sentence reformation.

The first step is to parsing the sentence and to divide the SL into several single translation units. Then the system will transform these individual units into TL. In the following, with accordance of the result and the rules, the system will identify the semantic relations between all these units and reorganize the words and phrases order.

According to the grammar rules of TL, the system will check over all the translation and attach some necessary semantic elements to the translation before it output the final result.

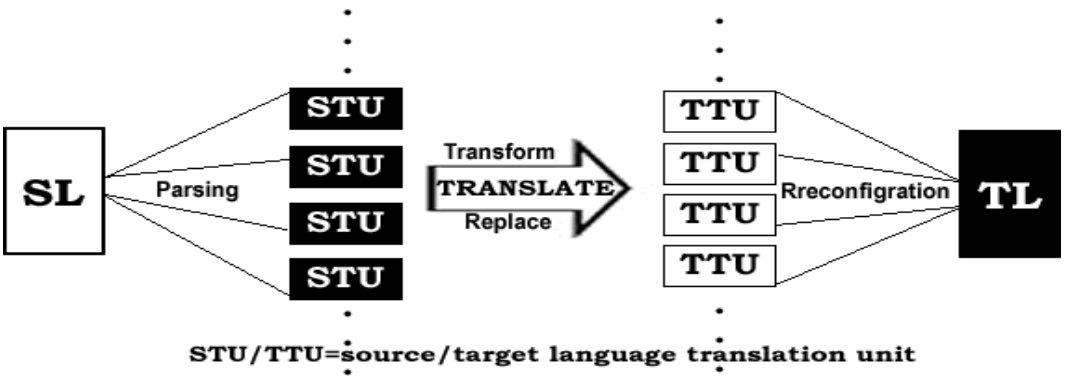

Table 19

This means in the SL semantic parsing stage, the more precise result we get, the more accurate units’ translation we may gain. And the final results will be much better.

2.2.5 Current Methods of Sentence Parsing

1.The SAM system: SAM was constructed by Roger Shank and other researchers in Yale in1975. It can pick up English short articles and output digest of articles in English, Chinese, Russian or Spanish.

The structure of this system has been demonstrated in the following picture:

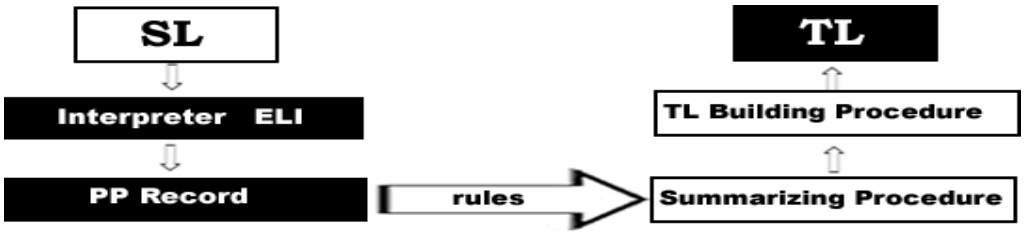

Table 20

The ELI is the language processor which is based on Mechanical dictionaries. The sentence analysis is realized by proving the predictions about SL. The procedure of identifying the definitions of words are consisted by two procedures: the predicting process and acting process：

If the result of the predicting procedure is proved reasonable, the acting procedure will carry out; the purpose of the acting process is to build up the handled language structure. The prediction can be divided into the following aspects: words, functions, specialized vocabulary, phrases, end and error.

PP Record is designed to identify the person, location and the complex relationship, it also may identify and replace the pronouns into concrete nouns.

And the internal rule system may match the frameworks that constituted in the ELI procedure to the Information Structure that has been recorded in MT system, and to fill up the possible omitted semantic parts.

In the summarizing procedure, the system will receive the language structures from the rule procedure and reorganize these structures into TL structure.

Finally, the TL Building Procedure will identify the Headword and structure of TL, and according to the TL's grammar to transfer the structure into the progressive language structure.

2. The ECAT system: This MT system is designed by the Chinese Academy of Science，it has 10 SL parsing programs, 1 TL analyzing program and 4 Mechanical dictionaries. The system's working procedure's details have been introduced in the Chapter 3.1.

But what is worthy to pay attention to is the special sentence segmentation system. The Chinese segmentation on sentences is totally different from the English segmentation,

which is based on phrases and verbs. And there are two ways: Segmentation by Keywords and Segmentation by Conditions.

Segmentation by Keywords:

The keywords in this system are: 1. interpunction and symbol 2.preposition 3.conjunction.

The first kind, the interpunctions and symbols are absolute segmentation signals.

The second and the third kind are ruled by the segmentation programs, and the segmentation signals are mainly pronouns, articles, numerals or verbs. Identifying all these signals is the only way to distinguish sub clause in English sentence.

For example:

/To help my disabled classmate /, / I spend an hour working /in his house everyday/.

Actually, if we want to realize this kind of segmentation in the MT system, the Mechanical dictionary must be enlarged into enormous scale. And this seems impossible for the ECAT system, neither the hardware condition nor the possibilities of building up such giant dictionary.

Chapter 3 Long Sentence Parsing System Structure and Current Problems

3.1 Introduction to the System Structure

According to what we've know in before, the MT system can be divided into two main parts: the Source Language Segmentation part and the Target Language Construction part. The system structure has been shown in the picture:

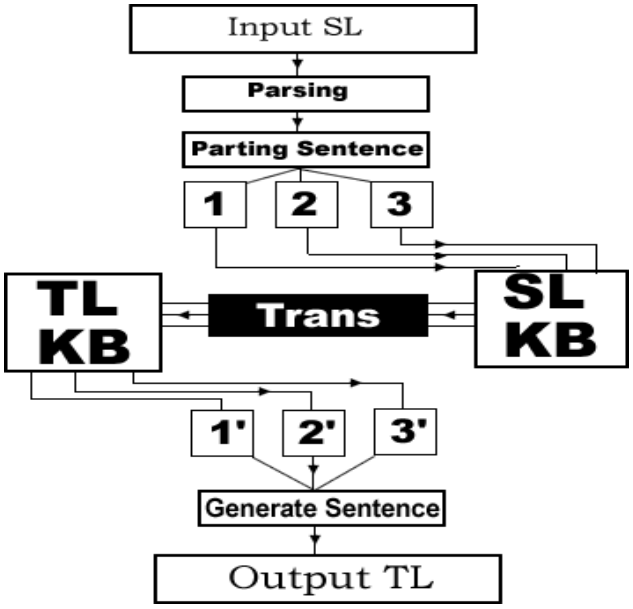

Table 21

In the MT systems, the translation quality is directly related to quality of SL segmentation. With high segmentation quality, the knowledge base of both SL and TL must be integral, the knowledge bases are the most significant elements for the system to analyze the sentences into correct translation units, and it also influences the reconstruction of TL.

High quality SL segmentation, complete knowledge base and TL generate system are considered to be the foundations of high quality translation.

In the following, we will use the ECMT translation system as example to illustrate what are the key elements in MT systems:

3.2 The Mechanical Dictionary:

In front of the paper, we've known that the translation quality relies on the correct sentence parsing, thus the mechanical dictionary can be considered as the decisive factor of

the MT system. In the following, how mechanical dictionaries composed and the inner forms of them will be discussed.

Actually the mechanical dictionary is not same to what we have known, but just a program file, stored and invoked by MT system, which is a set of concluding the descriptions about word features just like container. To compose the dictionaries, the first step is to confirm features and Feature Value (expression forms of word features in computer system).

According to the rules of English grammar, the ECMT (English-Chinese translation system) classified the feature values into 4 main categories.

Category One： Word Category. The abbreviation of this category is "cat". And according to the English grammar rules, all the word classes are expressed by 3-letter symbols. For example, "nnn" stands for noun, "dem" stands for the demonstrative pronoun, etc.

－det：determiner

－prp：preposition

－vrb：verb

－adj：adjective

－adv：adverb

－aux：auxiliary verb

－num：numeral

－rel：relative pronoun

－dem：demonstrative pronoun

－tit：title

－lfp：left quotation punctuation

－enp：end punctuation

Category Two: PCAT (Phrase Category): It refers to the types of phrases.

－S: sentence

－CL: clause

－NP: noun phrase

－VP：verb phrase

－BP：adverb phrase

－PP：preposition phrase

……

Category Three: Function Category. It may display the language unit's functions in sentences.

－Main：main clause

－Head：head word

－Comp：complement

－Advl：adverbial

－Obtj：directive object

－Sbuj：subject

－Subf：main subject

－Pred：predicate

－Attr：attribute

－Cnjt：Conjunctive

－Null：Null

……

Category Four: Logic Category. This category expresses word tense, person, number, degree, etc. The details of this category are in the following:

－The tense：tense={Pres, Past, Futr, Pres-part, Past-part,……..}

Pres = present

Past＝past

Futr＝future

Pres –part＝present-participle……

－The person: Its expression form in computer program is: person = {1,2,3}

－The number: Its expression form in computer program is: number = {sing, plur}

－The degree: Its expression form in computer program is: degree = {comparative, superlative}

……

In practice, all English words' feature values will be listed in the dictionary. In the following is a mechanical dictionary example:

For examples:

－concrete:(cat: nnn, }

－know: (cat:vrb, tense: pres, person:1, number: sing)

－the: (cat:det)

－are: (cat:vrb, tense:pres, person:2,3, number:plur)

At present, most of the MT system's dictionaries are similar to the upper demonstration, there are only lists of word category and function category. The other two feature values, the PCAT Value and the Function Value, are always needed to be identified by the computer algorithm programs in the procedure of parsing sentence.

For this reason, the apparent limitation of this kind of dictionaries is: without words' PCAT Value and the Function Value, the program has no efficient and accurate signs to carry out sentences parsing. and the consequence is semantic errors.

Then we may say the improvement on Mechanical dictionaries is the key to open the gate. Without full-functional mechanical dictionary, we may never ask MT systems to identify sentences precisely. Mechanical dictionary shouldn't be designed as word-checking note book, it is the core of the MT system's brain. To communicate, human being uses phrases and strings of semantic meanings to identify and parsing languages, but not only relies on single words' meanings or dictionaries in our hands or in our brains.

How use limited word structures to identify unlimited semantic structures has become the key problem.

3.3 Form Analysis

Before the procedures of identifying SL's words and parsing sentences, the initial step is to finish two tasks: the lemmatization and tokenization.

3.3.1 The Tokenization:

The token is the single word in sentence. Tokenization means to identify and list out all of these single words one by one. Although the written form of English has Space, there are still many special token forms need to be analyzed before list it out, for example:

——Seriate Numbers Form：Such as "02/04/2010", "1,000,20", "85%", "4/3", etc. Each of these forms needs to be identified as separate unit.

——Abbreviations Identification: All of the abbreviations in English should be identified as separate words, such as “Mrs.” , “Prof.”, “Dr. Tom”, etc.

——The word forms with hyphens: Such as three-year-old, one-third, etc.

——The phrases or idioms with space: Such as, just like, and so on, etc.

——The token with abbreviation need to be identified as separated words:

And many other word forms are considered as singles meaning units by us, most of them are abbreviations. In MT systems these kinds of word forms are also need to be reverted into complete forms.

Such as “let’s = {let us}”, “he’s = {he is} or {he has}”, “S’d={S would} or {S had}”, “can’t = can not”, “sister’s = of sister”, etc.

3.3.2 The Lemmatization

Lemmatization means to identify the inflexions in SL and to revert all of them into the original word forms, such as comparative degree, plural form, past participle, etc.

When all the inflexions are converted into their original forms, the system will label all these words with information of forms:

strongest → {strong + adj + superlative}

caught → {catch + vrb + past-part} or {catch + vrb + past}

does → {do + vrb + present + person=3}

After these two procedures, the MT system then has the basic capacity to parsing sentences on the deep grammatical levels. With the high quality Form analysis, to analyze the grammatical structures deeply and accurately will be easy. Furthermore, it is possible for the MT system to get the semantics of sentences.

3.4 Clause Structure Recognition:

The problem of clause recognition means how to identify the clauses from the complex sentence.

There are two main characteristics used to identify the Head of sentences: the lexical feature and sentence feature. The lexical feature relays on the Form analysis to get single word forms and the information labels of words and phrases.

The sentence feature can be divided into the following 5 aspects: sentence structure, function information, verb information, interpunction information and special situation. Then let’s look at the sentence feature in details:

(1) Features of sentence structure:

①Whether the present word is the beginning word of the sentences or not.

②To separate strings of words and phrase from the sentences in before and back. The emphasis will be concentrated on the verb phrases, commas and relatives, which are the significant signals in deciding the HEAD of the sentence.

(2)Function Information:

① If the present word is "if/ that/ what/ who/ where/ when/ why/whose/ whether/ how/ while…" , then according to the Dictionary's word category to get the information of the word.

② If the present word is "which", check the left word of "which" to find whether there is an "at/in/on…" exists or not, and according to the result to get the information of "which" or "at/in/on...+which".

(3) The number of Verbs

To identify all verbs in sentence, and to count all the verbs that exist in before and after (the result can be Zero), then use this result as the evidence value(证据数值) to identify the translation units.

(4) Interpunction information:

①If there was a comma and it is the only comma of the sentence, the comma can be used as the signal of dividing sentence; if there were more than one comma in sentence, whether there is a verb or VP between two commas can be the signal of parsing sentences.

②If the present interpunction is Colon or Inverted Comma, then the present interpunction and the first word on its right side will be considered as one WORD, and the WORD will be used as the signal of parsing sentence.

(5) Special Cases:

①If the present word is "and" or "or", then seek the verb on its left and right side.

②If the present word is "say", the word features can be used as the signal of parsing sentence directly.

Finally, using the same method and according to the result of upper part, to identify the ending of the sentence becomes easy.

In this paper, I just follow the designs of Xavier[1] in the procedure of marking clauses. The Xavier's idea is:

-First, to find out whether the head of the identified clause is the head of the multiple clause.

-Second, according to every possible head of clauses, the system may identify all possible layers of clauses.

-Third, using the grading function to grade all possible clauses by, and according to the marks of each possibility to decide the most appropriate clause identification.

-Fourth, using the predictions of before and by judging the phrase characters from each clause's beginning and ending, the identification of clauses will much more accurate than before.

Chapter 4 A Tentative Idea of Improving Long Sentence Parsing System and the Clause Identification Sample Demonstration

4.1 The Improvement in Mechanical Dictionary

Adding the semantic and grammatical values into the Mechanical dictionary:

Most designs of mechanical dictionaries concern about two main factors:

The first is the hardware limitations. For the system resources is limit, the dictionary can't occupy the system resources infinitely; the second is the complexity of composing dictionaries. If we just hope for the integrality and the functionality in making dictionaries, the dictionary will has an unimaginable enormous scale, which will be far beyond the capacity of hardware.

For these two reasons, all the dictionaries of MT are only use the single words as the basic compile units now. The unit's values only contain the basic category and grammatical attribute, but not contain the semantic and grammatical structures figures and functions at all.

The sequence is to leave the jobs of identifying the word's semantic and grammatical functions to the following sentence parsing procedure. We say this is quite unfair. Why? For we are still using the word as the minimum semantic unit to comprehend the meaning of sentences now, why would we ask machines to identify the word's function without giving any details of the word's semantic and grammatical features?

For all the reasons in before, the grammatical features can be added into the dictionary. It will be very helpful for the computer program to search and identify the phrases and clauses efficiently and accurately.

For the limitations about my major and I just began to interest in the linguistics since 2 years before, the time is really too insufficient to consider everything in details. Thus, with accordance of what I already read and understood, the design of Mechanical dictionary structure will be discussed in the following part:

1. The first class of value: All of the structures and the values of words are same to what we have introduced in Chapter4.2. All of these are defined as the first Value Category, we label this category as "sig.1". Its format is:

Sig.1[(cat:cat_name);(pcat:pcat_name);(tense:tense_name);(person:person_name);(number_numbername);……]

2. The second class of value: the function of this part of dictionaries is to record the semantic values and grammatical values. These values are signs of words' semantic meanings and grammatical meanings. This class of value's format can be described in Mechanical dictionaries as the following structure:

sig.2[(gram1_structure); (gram2_structure);…(gramN_structure)]

Thus, each word in the Mechanical dictionary has two different classes of values to express its special function, and the format is:

Word

{

sig.1(cat:cat_name, tense:tense_name, person:person_name, number_numbername….);

sig.2(gram1_structure, gram2_structure…gramN_structure);

}

In the following are some examples of this kind of dictionary's format:

Example1:

concrete

{

sig.1 (cat:nnn, number: sing);

sig.2(gram1_concrete&nnn,gram2_concrete&prp,gram3_of&concrete,gram4_prp&concrete);

}

Example2:

are

{

sig.1(cat:vrb, tense:pres, person:2,3, number:plur);

sig.2(gram1_are&pres, gram2_are&past, gram3_are&nnn, gram4_are&prp, gram5_are&adj);

}

Then, the first value class can be used by lemmatization and tokenization procedures. The second value class can be used in the sentence parsing procedure and searching for the

similar data in knowledge base to test and verify the reliability of the analyzing results. With the **Word-based Level Semantic Sentence Parsing(词层级语义分析)**，the errors that were caused in the procedures of identifying complex sentences' semantic structures, which are now produced by computer programs, will be reduced greatly.

4.2 Right-to-Left Reciprocal Lexical Sentence Parsing Algorithm

4.2.1 The Idea Came From Classroom

I found one interesting phenomenon in non-English major classrooms, which is that the reason of being confused by Long sentences is that students can't identify the beginning or the ending of each different semantic sentence structures, but not because length of the sentence. These confusions lead students into wrong judgments on dividing sentence into different basic semantic units.

Normally, if human want to read and comprehend English sentences, absolutely we need to read from first word at the very left till to the last word at the right end. It just like we are seeking the road in maze, we only know where is the entrance but don't know where exit is. We need to stop at every intersection, which can be considered as prepositions or articles, and then to decide which word is the next station that we need to go. The reason for us to perform this kind of reading process is that we are gifted reasoning capacity to understand the inner links between words and phrases.

Machines or computers have no capacities to reason, the only way to comprehend for them is to carry out the programs lines. If we gave a computer two completely same balls and ask them to choose one out all by itself, we may never see the result even if we wait for thousand years.

Then the question is there: why we must give computer two completely same balls? Why don't we give the exit to the computers and let them to trace where the entrance is? Even we human beings may be comforted when we are told where exit is at very beginning.

4.2.2 Right-to-left Lexical Sentence Parsing Algorithm

4.2.2.1 The Problems of Left-to-right Parsing Algorithm:

According to what we've mentioned in before, we may get the conclusion that the key problem of lower down the efficiency machine translation system is that MT system can't identify the beginning and the ending of semantic structures efficiently and correctly.

Thus, the key of improving the MT system, either in the remote future or at present, is

to build up mechanical dictionary and long sentence parsing algorithm appropriately.

For example:

SL: His first finding, which backs up earlier work at the US National Science Foundation, was that the degree of annoyance was not directly related to the time.

MT 1: 他的第一个结论，而备份早期工作在美国国家科学基金会，是，烦扰程度没有直接关系的时间。（http://translate.google.com/）

MT 2：他第一次发现,早在支持美国国家科学基金会是噪声的程度并不直接相关的时间。（http://dict.cn/fy/）

MT3：他的第一，发现哪一个向上后面在美国美国科学基金会早些时候使工作是那程度的对烦恼不是直接和时间有关。（http://trans.netat.net/index.php）

MT 4：他的备份早期在美国国家科学基金工作的第一个发现是烦恼的程度没有直接关系到时间。（http://www.microsofttranslator.com/）

MT 5: 他第一发现，支持更加早期的工作在美国国家科学基金会，是程度心烦未直接地与时间有关。（http://fanyi.cn.yahoo.com/）

These results tells us that if we ask computer to recognize the translation units from left to right, the most difficult thing is to make the computer understand which word is the beginning of the phrase and which word is the ending.

The mistakes that made by the MT systems can be concluded into the following:

Mistake 1: Wrong lexical analysis: such as the “finding” is translated as Verb but not Noun mistakenly.

Mistake 2: Sentence structure analysis error: such as MT1 to MT 4, the translation results may reflect one significant flaw that the systems can’t identify the ending of each translation semantic unit.

Mistake 3: System can’t identify long distance relationship between two grammatically related words. Such as “**was not directly related to the time”,** “was” and “related to….” should have been translated into one phrase, but the system was designed to analyze word by word from left to right, it will be easy to take error actions in the parsing procedures.

4.2.2.2 Basic Principles Introduction

To solve the difficulties of identifying the endings of the semantic units and confining

the semantic distance, the method of making the system to proceed from the ending to the beginning may be helpful to solve the problems. This means let the system to identify the last word of the sentence firstly.

To illustrate this kind of sentence parsing algorithm, some technical terms should be introduce in the first place:

Buffer: In the beginning, when we input the SL, the whole sentence will be stored in this area; as the sentence paring procedure carries on, the buffer will be emptied finally.

Stack: Stack is a kind of Data Structure. The priciple of storing data is "the first in and the last out". It means the data which was input into the buffer first will be placed in the bottom, and last input data will be put on the very top of the stack. When we need to take out data from stack, we need to take out the last data firstly, and so the first data will be the last one to be taken out. The structure of the stack is similar to the following picture:

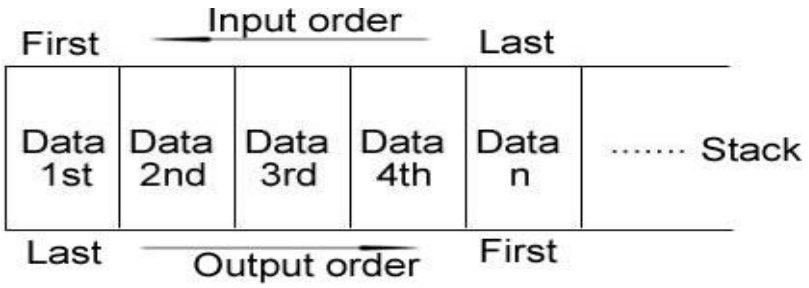

Table 22

Now we use a chart to illustrate how buffer and stack work in sentence parsing system:

When we input the sentence into the system, the situation of both buffer and stack same as the following picture:

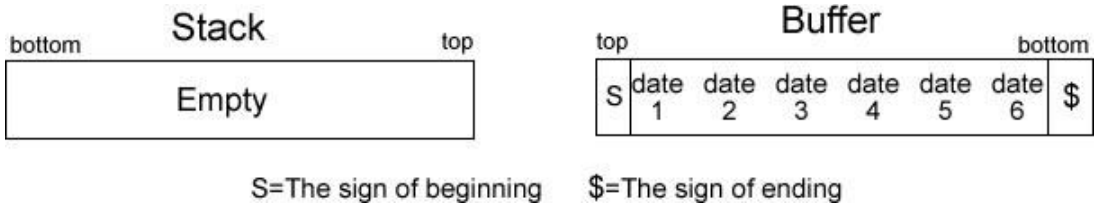

Table 23

According to the rules, the system begins to identify the semantic units and input the result into the stack. As the picture, the first translation unit "data 1" will be moved from the top of the Buffer into the bottom of the Stack, and the situation will be same to the following:

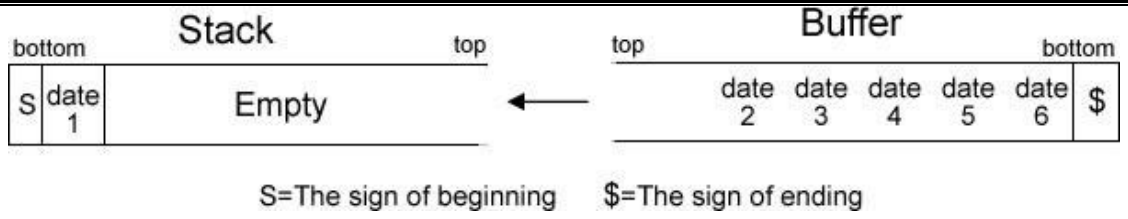

Table 24

Then the procedure carries on till the Buffer become empty and the Stack become full. This is the basic process of the sentence parsing.

4.2.2.3 Right-to-Left Algorithm

This algorithm is different from the present English analysis system algorithm. The system will start from the right side of the SL, according to the last word which has been found, to find which word is the beginning of the undividable semantic translation unit in which the last word exists.

In the following we will take the sentence "**His first finding, which backs up earlier work at the US National Science Foundation, was that the degree of annoyance was not directly related to the time.**" as the example to introduce the algorithm in details.

Step1, the system find the "**.**" and then recognize it as the ending sign of the sentence, put transform it into the sign of "**$**" and insert it into the bottom of the stack.

Step2, find the nearest word of $ is "time", go for the preword "**the**", according to the rules, "**the**" and "**time**" can be put together and be recognized as one structure **NP=Data1='the time';**

Step3, find the nearest word of the Data1 is "**to**", according to the rules

{

If 'Verb' before 'Prep then Verb+Prep = VP

Else goto next word

}

Then decide the "**related to**" to be the VP, for the "**related**" is the P.P form, then seek the next word on the left side to find if there was Noun or Personal Pronoun exists, find no Personal pronoun but "**was**" existing, then according to the rules:

{

If 'Copula' before 'Verb' then Copula+Verb=VP

Else if 'Personal Pronoun' before 'Verb' then Verb=VP

Else 'next word' till 'Copula or Personal Pronoun'

}

……..

Then decide the ‘was not related to’=VP=Data2

Step4, find the nearest word on the left side of the Data2 is a Noun ‘**annoyance**’, according the rules:

{

If ‘Noun’ before ‘Copula’ then Noun+Copula=SP

Else if ‘Personal Pronoun’ before ‘Copula’ then Personal Pronoun+Copula=SP

Else ‘next word’ till ‘Noun or Personal Pronoun’

}

Then decide the “**annoyance**” is the subject of the Data2.

Same as the upper method, we may easily find the “**degree of annoyance**” should be placed together, and it will be the subject of Data2.

……

StepN, find the Start sign “**S**”.

The core of this kind of algorithm is to suppose the very left word is the beginning of the smallest semantic unit, and then try to prove whether this assumption is right or not. It is a kind of method to catch the most certainty to trace the uncertainties.

And the process of this analysis procedure is in the following:

Present word	**Condition**	**Result**
1 # time	**time$**	Seek for left word
2 # to	**prep.**	Combined with ‘time’, continue = Data1
3 # related	**Verb +Prep**	Combined with Data1,continue = Data1
4 # directly	**Adv+Verb**	Combined with Data1, continue=Data1
Data1 = ‘directly related to the time.’		
5 #not	**Adv+data1**	Seek for the left word
6 #was	**Copula+not**	Combined with ‘not’, continue=data2
Data2= ‘was not’		
7 #annoyance	**N+copula**	combined with data2, continue=SP
8 #of	**Of +N**	Seek for the left word, continue=data3

9 #degree	**N+of**	Combined with data3, continue=data3
10 #the	**Definite article**	Combined with data3, continue=data3
Data3= 'the degree of annoyance'		
11 #that	**that**	Seek for left word
12 #was	**Copula+that**	Combined with 'that', continue=data4
13 #,	**Comma+was**	if ',' + copula, then seek for next ','
14 # **,**	**$ + , + $**	Seek for the left word
Define 'which backs up earlier work at the US National Science Foundation' as single structure		
15 #finding	gerundial +, +copula	Combined with 'was', continue=data4
16 #first	**adj+gerundial**	Combined with 'finding', continue=data4
17 #his	**person+adj**	Meaningless, then
Data4= 'first finding was that'		
18 #his	**pronoun**	Seek for left word
19 # 'S'	**Starting Sign**	Continue=data5
Data5= 'His'		
Then continue to the single structure 'which backs up earlier work at the US National Science Foundation' as single structure		
1 #Foundation	**capitalization**	Seek for left word
2#Science	**capitalization**	Combined with 'Foundation', continue=data1
3 #National	**capitalization**	Combined with data1, continue=data1
4 #US	**capitalization**	Combined with data1, continue=data1
6 #the		Combined with data1, continue=data1
7 #at	**Prep+the**	Combined with data1, continue=data1
Data1= 'at the US National Science Foundation'		
8 #work	**Verb or Noun**	Seek for left word
9 #earlier	**Adj+N**	Combined with 'work', continue=data2
10 #up	**prep**	Seek for left word
Data2= 'earlier work'		
11 #backs	**backs up**	Check for phrase, continue=data3
12#which		Combined with data3, continue=data3

13# ,		
Data3= ‘which backs up’		

Table 25

Then the whole sentence structure has been stored in the Stack, the system status will be:

Stack

top	Data5= ‘His’
	Data1= 'directly related to the time'
	Data2= 'was not'
	Data3= 'the degree of annoyance'
bottom	Data4= 'first finding was that'

Table 26

In the following, the system will take out all of these data from the stack one by one.

The basic idea of this kind of English sentence parsing algorithm is to use the preposition as the key word to judge whether the word on its right-side is the true beginning of the semantic translation unit.

4.3The Right-to-left Sentence Parsing System Working Process

The only significant difference of the Right-to-left parsing system is the algorithm used in sentence parsing procedure. Without this difference, we may find there is any difference exists, and of course the working processes are quite same to the other systems.

In the following, I combined the new dictionary structure and Right-to-left algorithm into one system, and the working process is in the following:

Step 1): Input the SL

Step 2): Lemmatization and tokenization

Step 3): Search in Mechanical dictionaries, and list results.

Step 4): From left to right, identify all the prepositions and relatives.

Neither of them exists, jump to Step 6.

Step 5): According to the Values in dictionary and the results of step 4, system will carry out the procedure of sentence parsing, and identify clauses.

Step 6): According to the results from upper step, divide the main clause and sub clauses,

and divide each of them into Main Clause Translation units and Sub Clause Translation Units.

Then, reorder all these units according to the relationships between themselves.

Step 7): Translate minimal units into TL and form the parallel bilingual translation text.

Step 8): According to the Values(sig.1; sig.2) to reorganize the minimal units into word groups.

Then search for the sentence structures that are similar to these word groups and the parallel translations.

Step 9): According to the results of Step 5, generate the translations of the main clause and sub clauses independently.

Step 10): According to the result of Step 9, generate initial translation.

Step 11): According to the grammar rules of TL, adjust the translation by adding or deleting words.

Step 12): Output the TL.

4.4 Example Demonstration

To demonstrate working procedures of long sentence parsing algorithm, we use one sentence as example to demonstrate, and the whole demonstration is based on the dictionary structure which has been proposed in this paper:

Example sentence is:

Concrete pillars for highway bridges，as we all know, that previously only had steel rods inside are now enclosed in steel.

Step1：Input the SL: “Concrete pillars for highway bridges，as we all know, that previously only had steel rods inside are now enclosed in steel.”

Step2: The lermmatization and tokenization procedure.

The result is:

“Concrete pillar for highway bridge, as we all know, that previous only have steel rod inside are now enclose in steel.”

Step3：Dictionary consultation procedure.

According to the result from step2, the outcome is:

concrete

{

sig.1 (cat:nnn, number: sing);

sig.2(gram1_concrete&nnn,gram2_concrete&prp,gram3_of&concrete,gram4_prp&concrete);}

pillar

{

sig.1 (cat:nnn, number: sing)(cat:vrb tense:pres person:1 number:sing);

sig.2(gram1_pillar&nnn,gram2_pillar&prp,gram3_nnn&pillar,gram4_prp&pillar,gram5_concrete&prp);}

for

{

sig.1 (cat:prp);

sig.2(gram1_for&nnn,gram2_for&syn);

}

as

{

sig.1 (cat:adv) (cat:cjc) (cat:prp);

sig.2(gram1_as&nnn,gram2_as&adj,gram3_as&syn);

}

we

{

sig.1 (cat:prn, number: 3);

sig.2(gram1_we&nnn,gram2_we&veb);

}

have

{

sig.1 (cat:vrb, number:1);

sig.2(gram1_have&nnn,gram2_have&p.p,gram3_have&to do);

}

Step4: Record pronoun , verbs, prepositions and adverbs:

Preposition{ for; as; that; in }

Pronoun{we}

Verb{ know; had; are; enclosed }

Adverb{ previously }

Step5: according to the dictionary, and using the Right-to-left algorithm parsing the sentence into the following semantic translation units:

Data1=Concrete pillar

Data2=for highway bridge

Data3=as we all know

Data4=that are now enclose in steel

Data5=previously only have steel rod inside

Then according to the grammatical rules, the next procedure is to check the parsing results and reorder the Datas in correct word sequence.

The processes are in the following:

{

#1 gram_concrete&nnn = Concrete pillar

#2 gram_ pillar&prp = pillar for

#3 ……= for highways bridge

#4 ……=as we

#5 ……=we all know

#6……=Non-empty Sign＋,+ as + we + v;

= tag “we” as Non-empty Sign and Non-Subject＋,+we+,+ Non-empty Sign，the nearest semantic relative word of the word “know” is “we”.

=,as we all know,

#7……...=Non-empty Sign＋“,”＋ that + Non-First-person pronoun sign＋Copula＋ Full stop(句结束符号)

= “that” is meaningless null set(无意义的空集)

#8 ……=have steel rod

#9 ……= have steel rod inside

#10 ……=are now enclose

#11……..=enclose in

#12 …….=are now enclose in

#13 …….=in stee

#14…….=According to #6, find there are two verbs in the following “have” and “are”, according to the rule of English, the clause part is “that…….inside”. #14

Then, the sentence structure has been identified by the system into the following structure:

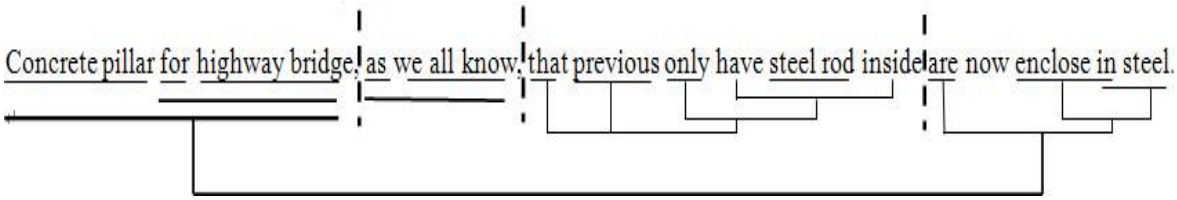

Table 27

Step6: According to the results formed in step5, distinguish the main clause, the sub-clause and the translation units.

Distinguish the clauses:

#1 according to the result of step5-#6, insert sentence=as we all know

#2 according to the result of step5-#14, sub-sentence = that that previous only have steel rod inside

#3 according to #1 and #2 of this step, main clause= Concrete pillar for highway bridge are now enclose in steel

Till then, the sentence has been divided into 3parts. Then according to the results, the system will seek for the meanings of translation units in the mechanical dictionary::

Finding the translation units, the final result is:

#1Concrete pillars： 混凝土＋墩

#2 For： 用于+空符 1＋的＋空符 2

#3 highway bridges： 公路桥

#4 as： 正如＋空符＋一样；

#5 we all know： 我们都知道

#6Non-empty sign ＋“,”＋ that + Non-First-person pronoun sign＋Copula＋Full stop

＝ “that” is meaningless null set, its position need to be adjusted, and its translation is just only a space “ ”

#7 previously： 在以前；以前

#8 only had steel rods： 只有钢筋

#9 inside： 在里面

#10 are enclosed: 被封装

#11 in: 在＋空符＋里面

#12 steel: 钢材

Step7:Then according to the results from step5, analyse the results of step6 .

#1: #1 #2 #3 = 用于#3 的#1 ⟶ 用于公路高架桥的混凝土墩

#2: Non-empty sign＋,+ as + we + v; the “we” has been tagged as non-subject. Non-empty sign＋,+we+,+ Non-empty sign，”know” is Nearest relative word toward “we” = #4 #5= “正如#5 的一样” ⟶ 正如我们都知道的一样

#3: #6#7#8#9 ⟶ 在以前只有钢筋在里面

#4: Non-empty sign＋“,”＋are＋enclosed, identify the “are” is the nearest and relative word toward “closed”, then the structure can be identified as the Passive Sentences(被动句) = #10#11#12 = 被封装在#12 里面 ⟶ 被封装在钢材里面

Step 8: According to the results of step7, reorganize the sequence of translation units:

#1According to the grammar rules: the structure of #2 of step7 will be placed at the head.

#2Non-emptysign＋concrete pillars +insert sentence + sub-sentence + are

= #1#4 =main clause

#3 Will be placed at the bottom of the sentence.

The final translation structure order is：#1#2#3

Step 9: Adjust translation: delete space, insert punctuations of TL.

And the final result is:

“正如我们都知道，用于公路桥的混凝土墩现在被封装在钢材里面，在以前只有钢筋在里面”。

With the examples searching in knowledge bases and result comparing procedure, the qualification of the MT translation can be better and readable.

Although the process seems very easy to be described on paper, it is really hard to build up such analysis system in fact. Without close cooperation between linguists and computer experts, this kind of system is hard to be realized, there is any possibility for only one man to fulfil such exceptionally arduous project.

Chapter 5 Conclusion

5.1 Rational Thinking on MT

MT has developed for decades, and in this period we really have achieved great achievements. Especially in the resent 15 years, various technologies have been used in this field. As the rapid development of the Internet, many practical MT systems emerged, such as On-line translation service systems can translate Web pages, E-mails and some other documents, more and more people begin to realize the importance of MT.

Unfortunately, there are still many significant problems waiting to be solved. For instance, how to make the system to accumulate experiences, to deal with the ambiguity efficiently, to summarize and describe the principle of the characteristics of language, etc. Till now, the ALPAC report is still worthy to be read. We must re-examine the terminal goals of doing the MT research, which is to fulfill FAHQMT (Fully Automatic High Quality Machine Translation). For there are so many differences in cultures, customs, social backgrounds between two languages, the goal of realizing the "true MT system" seems impossible. At present, this is also agreement of the scientists in recent years, we need to refresh our short-term goals and waiting for the reinforcement of technology.

5.1.1 The Impossibilities and Possibilities of Terminal MT:

For the basic hardware structure of the computer is consisted by the CPU (central processing unit), the EMS memory and the Bus Structure. The CPU can only execute one instruction once time, it need time to execute the instruction and react to the instruction and then carry on to the next one. Just for this reason, the computer may only execute string-like thinking, it means we need to build up a group rules to limit the other group rules.

Scientists found that if we want to make the computer to understand, just understand but not to communicate with human, the daily language spoken by a common people which only includes 20,000 English words, we must have the computer to execute about 100 billion commands per second at least.[微电子与计算机技术：P68]

The undeniable truth is that the development of the computer science is far beyond our imagination. In 2009, November 17, the Super computer - Linpack which is located in the U.S.A Oak Ridge National Laboratory, its speed has reached 1750 Trillion per seconds, and

the peak speed reached 2300 Trillion per second. This means the computer may calculate "1+0" for 2300 trillion times will spend only one second.

But with the high speed is still insufficient in making computer to understand human languages. Human brain is the most complex structure in the nature, the amount of the neurons in the cerebral cortex is equal or even more than the amount of stars in the Milky Way galaxy. What's more is all of these neurons not only have an elaborate division of labour but also each of single neuron can accept the neuron action information and deliver the action information to all the neurons around it at the same time. This is the biggest functional difference between human brain and computer.

If one neuron may spend 0.01 second to accept command and carry out the action, every one neuron may pass the message to 5 neurons cells, it means that only in 0.1 seconds, one single neuron action message can be delivered to and carried out by 12207031 neurons, and the number may grown in geometric progression as the time pass by.

The result is if we want to make out a computer same as our brain, we need to build 10 billion CPU into one computer, this seems impossible, even in the following 20 years.

But there is still something can be expected. Some scientists have begun to research the Fifth Generation Computer since 1990s. This kind of computer is called Neural Network Computers (NNCs). It imitate the neuron links of human brain to build the line structures of the computer, and use optical function materials and biological materials to make the CPU, this kind of computer may fulfill the functions of Human right brain, and it may has the Intuition. (微电子与计算机技术：70)

All in all, if we want to liberate human from translation jobs completely, we must uncover the "black box" of brain. To fulfill this dream, we must make a breakthrough at the aspects of AI (Artificial Intelligence), machine cognitive ability, etc to make the computer have abilities to think, learn, or even abilities of reasoning and judging.

Indeed, there is still a quite long way to go before we accomplish our terminal goal- the FAHQMT (Fully Automatic High Quality Machine Translation).

5.1.2 Contradictions

1. Is there really an answer to the question we are considering?

4 = 3 + 1 ; Without 1, how can we prove 3 > 4 ?!!!

In the philosophy theory, the boundless aggregate is an undefined principle. For the finity can be defined as aggregate, but without

knowing where are the bounds, how can we say it is an aggregate?

And we say finity always and must be included in the infinity. The language of human is the result of long-term development, it was not created by finite rules and principles, it just a kind of signal system[语言哲学：18] to express the outline of the language. For this reason, the Language that we speak in daily life is just an aggregate of the infinite language.

It means if the machine can't "express" even only one word "YES" all by itself consciously, the language of machine will not be considered as "language", but to be the attachment of human language.

Based on this kind of idea, Using MT system to translate is equal to use the Finite to express the Infinite. This is an absolute Paradox obviously. Because using finite characters to conclude infinite object is impossible, neither in practice nor in theories.

2. Can the human brains be replaced by Machine devices?

For the language is the basic trait of human being. Without thinking, without language. Language is the carrier of our spirit, it is the physical form of the human thoughts.

For this reason, once MT system can provide translations flawlessly, it means machine can understand the deep-level semantic meanings of human language, which equals to say that machine has the same mode of thinking as our human. The difference between Computer and Human brain will not exist any longer, and at the moment, machines can be called, frankly speaking, a true meaning "mechanical human" but not only to be called "Robot"! The answer is horrible!

When a man died, people may copy his experiences and his own language customs into disk, and rebuild the man's thought in computer, just like installing software into the computer, to resurrect his spirit. By then shall we decide whether the man has dead or not? For the death means the Extinctions of both flesh and spirit! Flesh dead but spirit may exist continuously may be the by-product of MT researches.

5.2 The Developing Trend of MT and Its Impacts

The meaning of MT research not only lies in gaining short-term economic benefits, but it may also speed up the non-barriers development of human society and technology.

For the forms of human language is impossible to be list all, the ideas of building up complete language-form dictionaries is a impossible mission. It means we still have no capacity to use finite program rules to express infinite language forms. The computer, at

present, only can deliver messages correctly in special language form.

It equals to say that we may use rules to express restricted languages in some certain circumstances. Temporarily we may only use MT systems as assist tools but not complete self-control system to compose natural language.

For example, we may use MT system to search for similar translations, such as the short message translations, the subtitle translations or Web-site information publishing jobs, or to assist human translators,.

At present, the researches about MT have developed for about 70 years, since the first computer was created. The qualification of the MT translations is still far away from our needs. The main reason is we can't find out efficient rules to express languages by existing systems. And we are also being astricted by the technical conditions right now.

But all of these can't be the barriers for the MT research to step forward, although the future seems too far and impossible to be realized, the future is still waiting for us in remote future.

And of course I will continue my research about MT, and my research emphasis in future will be put on building up new dictionary structures and finding rules for the English-Chinese MT systems, such as how to identify the main clause and sub clauses, how to bond two long-distance relative sentence in translation.

At last, I really want to appreciate my tutor Mrs. Ye, she is not only my teacher, but also the teacher who aroused my interests in linguists. And I really hope that I may have chance to take more research in the MT field.

5.3 Conclusion

The true MT, or in another words, the real MT which meet all requirements of human communication has been heard since long time ago, but it has not been seen till now. The reason is we haven't found the best way to "rule" our language, what we have done in past decades were just to find another group of rules, which came out of from language itself, to explain or even to restrict language itself.

The significant problem of today's MT is how to find another new symbol system, whether it is based on our language or not, to explain our languages. For example we may use pictures or patterns to explain the semantic meaning.

In the last, although we may say there are so many experts saying that MT will step into cemetery sooner or later, I still insist that everything is changing, and there will never be a

final station on the trip of technology development.

The idea of using machine to translate was created about hundreds of years ago, comparing to the long history that human passed through, it is too short to get the final conclusion. We should pour constant energies and time to improve the idea, but not abandon it just for it can't fulfill our imaginations right now. As the most important communication aiding tool, it will never be abandoned by the trend of history!

History always repeats itself!

Works Cited

[1]计算机翻译研究，张政，2006：52

[2] A four-valued semantics for terminological reasoning, Artificial Intelligence, 38, 1989

[3]Uwe Muegge (2006), "An Excellent Application for Crummy Machine Translation: Automatic Translation of a Large Database", in Elisabeth Gräfe (2006; ed.), Proceedings of the Annual Conference of the German Society of Technical Communicators, Stuttgart: tekom, 18-21.

[4]Boretz, Adam, "AppTek Launches Hybrid Machine Translation Software" SpeechTechMag.com (posted 2 MAR 2009)

[5] Xavier Carreras , Lluís Màrquez1Boosting Trees for Clause Splitting [C] Proceedings of CoNLL220011Toulouse France. 2001 :73 - 75.

[6] Nieben, Vogel S, Ney H, et al. ADP Based Search Algorithm for Statistical Machine Translation, ACL 36/COLING17,1998.960~967

[7] Josef F, Ney H. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In Proc. Of the 40th ACL, Philadelphia, 2002. 295~302

[8] 赵铁军. 机器翻译原理[M].哈尔滨：哈尔滨工业大学出版社

[9] 易绵竹，南振兴.计算语言学，上海外语教育出版社，2005

[10] 机器翻译研究，冯志伟.中国对外翻译出版公司，2004

Bibliography

[1] 冯志伟《自然语言机器翻译新论》，语文出版社, 1994。

[2] 柯　平《欧美的机器翻译》,《中国翻译》1995 年第 2 期。

[3] 晋薇，黄河燕，陈昭雄，基于语义相似度并运用语一言学知识进行双语语句词对齐，2002，计算机科学，第 29 卷，第 11 期

[4] 姜　新　朱学锋　俞士汶《机器翻译的评价与应用》，《中国计算机用户》1989 年第 9 期。

[5] 姚兆炜《我国机器翻译概况与展望》,《文字改革》1989 年第 9 期。

[6] 吴保民. Matlink 英汉机器翻译系统 [D]. 郑州:信息工程大学 , 2002. 6.

[7] 郭永辉 ,吴保民,王炳锡. 基于规则知识的英语词法分析研究 [J]. 计算机应用 , 2004, 24

[8] 郭永辉 ,吴保民,王炳锡. 基于 GLR 算法的英汉机器翻译浅层句法分析器[J].计算机工程与应用 ,

[9] 李剑 ,郭永辉 ,吴保民等. 基于 EICG 的英汉机器翻译规则的句型转换器设计 [J]. 信息工程大学学报 , 2005, 6

[10] 薄冰. 高级英语语法 [M]. 北京: 世界知识出版社 , 2002.

[11] 黄河燕 ,陈肇雄. 基于多策略分析的复杂长句翻译处理算法 [J]. 中文信息学报. 2002, 16

[12] Mary Dalrymple. The Interpretation of Tense and Aspect in English [C],In Proceedings of 26th Association for Computational Linguistics, 1988

[13] 吕雅娟，赵铁军，李生，单语句法分析指导的双语结构对齐，计算机研究与发展，2003 年七月，第 40 卷，第七期

[14] 邹冰，《现行英汉机器翻译系统存在的问题及解决策略》，东北大学学报，VOI.5，No.5

[15] Dorr, Bonnie J.1994, Machine Translation Divergences:A Formal Description and Proposed Solution. Computational Linguistics, 1994,20(4): 597-633.

[16] Nagao, Makoto, 1989, Machine Translation －How Far Can It Go? Oxford: Oxford University Press.

[17]Muriel V，Marjorie Leon，SPANAM and ENGSPAN: Machine Translation at the An American organization CL，1985，Vol.11，No.2 一 3.

[18」Ashizaki Tatsuo，Adaptation of JICST，5MTSystemfor Work station and PC，s，Proceedings of MT Summit V，Luxembour，July10 一 13,1995.

[19] C.K1shorePaPieni,Salim Roukos，Todd Ward，Wei 一 Jing Zhu，Bleu: a method for automatic evaluation of machine translation，232 一 240，ACL2002

[20] Doddington G，Automatic Evaluation of Machine Translation Quality Using N Gram Co –Occurrence Statistics， R. Nist Research Report，2002

[21] 张剑，吴际，周明，机器翻译评测的新进展，中文信息学报，2003，Vol.17，NO.6

Acknowledgements

At the completion of my thesis, I must express my heartfelt gratitude to all those persons who have given me much help, advice and encouragement in the process of preparing for and finishing my graduate thesis. Without their help, this thesis would have been impossible.

First and foremost, I greatly appreciate my respectable supervisor Professor Ye Huijun who has offered me generous guidance, patient instructions, valuable suggestions and painstaking correction. But for her help, I would not have been able to accomplish this challenging thesis.

My sincere thanks also go to all the other teachers in the Foreign Language Department who have taught me in the past two years. Their insightful lectures have widened my scope of knowledge and will be beneficial to my future study and research.

I should also give my special thanks to my colleague for their support and persistent encouragement. Sincere thanks should also be extended to my parents who have helped me in many ways.