Automatic Translation: Overcome Barriers between European And Chinese Languages

WONG Fai, MAO Yuhang, DONG QingFu, QI YiHong
Tsinghua University (China)

Speech and Language Processing Research Center,
Tsinghua Univ., Beijing 100084, P.R. China
Email: myh-dau@tsinghua.edu.cn

ABSTRACT

This paper describes the differences between European languages and Chinese language, the urgency and possibility of realizing machine translation between the languages of Chinese and Latin's. Three different levels of realizing machine translation are enumerated: online dictionary, word-to-word translation and text-to-text translation. The difficulties and the ways to realize machine translation are studied and analyzed. It is pointed out that the rule-based machine translation is most promising and feasible.

Keywords: machine translation, online dictionary, word segmentation, and ambiguity resolution

Introduction

Language is a basic tool of human communication and contact. Communication among the people of different languages relays on translation. Manual translation is not always possible at the real-time, and/or may be expensive. The use of machine automatic translation may offer a cheaper and faster service. This is the reason why machine translation is always very attractive.

The exchanges of culture, science, technology and trade between Europe and China are fast developing. The obstacle of language difference influences on efficiency of exchanges. Machine translation of European languages to Chinese is even more urgent. It has been discovered that the constitution of human language is very complicated when one really starts to develop machine translation. It is not so simple to realize automatic translation of human language by computer.

However, there is a close "consanguineous" relation between different European languages, the vocabularies and grammar constitutions are similar. It is relatively easy to design machine translation among European languages, nevertheless, it has taken several decades to realize some machine translation systems among European languages and put into market. The systems, such as

the series of SYSTRAN, may be acknowledged as quite successful and have been practically used.

In next section, the differences between Chinese and European languages are studied and pointed out. The different level of translation methodologies are analyzed, and described in the second section, and varies developed translation tools are reviewed in the last section.

DIFFERENCES IN LANGUAGES

For languages of different origin, there are big differences in vocabularies and grammar constitutions. It is more difficult to translate between such kind of pair languages.

Chinese is one of the languages, that is quite difference from the west languages. To realize the automatic translation between the European and Chinese languages is very difficult and time-consuming. In following, we have classified the main features between the languages of European and Chinese in differences:

European Languages	*Chinese Language*
➢ Thousands of word is consisted of limited alphabet.	➢ Words are consisted of thousands of characters
➢ In a sentence, word series are separated by spaces.	➢ In a sentence there is no space between words and characters
➢ Word morphological changes.	➢ No morphological changes
➢ There is gender difference in nouns.	➢ No gender difference in nouns.
➢ With article before noun in sentence.	➢ Few usage of article before noun
➢ The adverbial modifiers of time and location are placed at the end of a sentence.	➢ The adverbial modifiers of time and location are placed at the beginning of a sentence.
➢ Few usage of classifiers to modify nouns.	➢ Usually nouns are modified with classifiers.
➢ Usually modifiers of subject or object are placed behind.	➢ Usually modifiers are placed before.
➢ Expression of tense and voice explicitly.	➢ Expression of tense and voice implicitly.

The divergences mentioned above must be dealt with carefully when one is going to design a translation system of European language to Chinese or vice versa by computer.

DIFFERENT LEVELS OF AUTOMATIC TRANSLATION

Machine translation methodologies can be realized in several levels according to the form of translation and the accuracy of translation required. Followings describe the different approaches, in sections, the benefit and the difficulties.

1. On-Line Dictionary Approach

This approach is used for looking up a single word from a computer. To build up such a system, the fundamental construction is to create an electronic dictionary. This can be achieved either from a ready-made electronic corresponding dictionary, or import the content from a paper dictionary, which, as a result, causes a lot of work and manpower in typing and verification. In additions, some programming work is necessary to perfect the user interface.

A proper designed on line dictionary may be useful in varies situations. With the aids of moues tracking technology, for example, user can easily look up the meaning of a word by using the mouse pointer. Thus, a fast reading environment can be constructed for monolingual user. The advantage is that this can be used with varies Windows operating systems (Chinese, English, French, Portuguese, etc.) and applications (in any kind of word processor, Internet browser, and etc.). On the other hand, the system should be able to distinguish the language of look up word, i.e.: Chinese or European. Then, the right translation can be carried out with the minimum degree of user involvement. Such kind of systems has been developed in our center: Portuguese-Chinese, French-Chinese, etc.

On line dictionary can only translate for a single word, or at most word phrase. It is not convenient enough for use in translating a text. As a result, this kind of translation tool cannot fulfill the need of someone who would like to translate a whole document in batch.

2. Word-to-Word Translation Approach

This translation approach can provide user a rough translation for sentence, paragraph, and even a whole text. This kind of word-to-word translation system is not so difficult to realize. It only considers the morphological analysis during the translation process, especially for European languages, which are rich with morphological changes according to cases, tenses, gender, etc. Another consideration in this translation approach is dealing with the Chinese word segmentation. Unlike European languages, there is no delimiter between Chinese words. Comparing with text translation system, this approach is much simpler since there is no any syntactic analyzing of the source text, and no word order rearrangement at target text generation. It does not consider any disambiguation of polysemy in original word nor process the morphological changes at generated text. The result of such translation may not be understandable in some cases.

When translating Chinese to European, the maximum match algorithm of backward and forward may be used in segmenting the Chinese words. The text may be wrongly interpreted if improper segmentation in ambiguous phrase of overlap type.

The translating results of word-to-word translation system are still not good enough. But it can be applied to translating the document of trade and commerce, or some kind of technology terms. This kind of system can provide a rough reading and is worth to develop with limited funds and

manpower, if the requirement of translation accuracy is not the primary consideration.

3. Text-to-Text Translation Approach

The most challenging objective of machine translation is the translation of a whole sentence. Where the system includes the analyzing processes not only the morphological analysis of individual words, but also the analysis on semantic and relations in the context of other words, hence to carry out the sentence translation, even to the whole paragraph.

Designing the knowledge database (lexical and grammatical information) is the fundamental task in machine translation. Where the database structures for lexical information and grammatical data, in a large degree, is quite different, and must be designed separately. Not only the syntactic but also the semantic information of the lexicons should be processed, in advance, and recorded in the database. Which includes the rule information for morphological analysis: inflectional morphology and derivational morphology. European languages have relatively rich inflectional morphology, providing indicators of subject-verb and adjective-noun agreement and marking case relatively explicitly. So is the derivational morphology; carefully exploiting the regularities may greatly reduce the size of database.

Ambiguity resolution is another important issue for lexical ambiguities (category ambiguities, homographs and ploysemes, etc.), and structural ambiguities. We have done an arduous work of counting, analyzing and classifying in ambiguous information. Many scholars have explored and studied in these problems. We can fully utilize the achievements of their researches to give help in formulating.

Although the form of expression of human language varies but the type of sentence pattern are limited in any languages. It is possible to find out the sentence pattern of different kinds of languages and their corresponding mapping. This is absolutely necessary for realizing machine translation.

In summary, to realize a text-to-text machine translation requires a series of processes: word segmentation (for the case of source language is Chinese), morphological analysis, ambiguity resolution, phrases construction, sentence pattern matching and target language generation. In our point of view, rule-based translation model gives the most promising and feasible result.

DEVELOPED TRANSLATION SYSTEMS

In our center, we have successfully developed some practical translation systems of different approaches. Following gives a briefly description for each of them:

1. Chinese to English Translation System (THCE 2.0)

This system is our first research activity in machine translation. The key technology applied in this system is the word segmentation for the Chinese text and the disambiguation of ambiguous phrases. After nearly ten years of refining, we have successfully developed an effectively parsing module for Chinese. With the help of this module, many machine translations from Chinese to European languages are feasible.

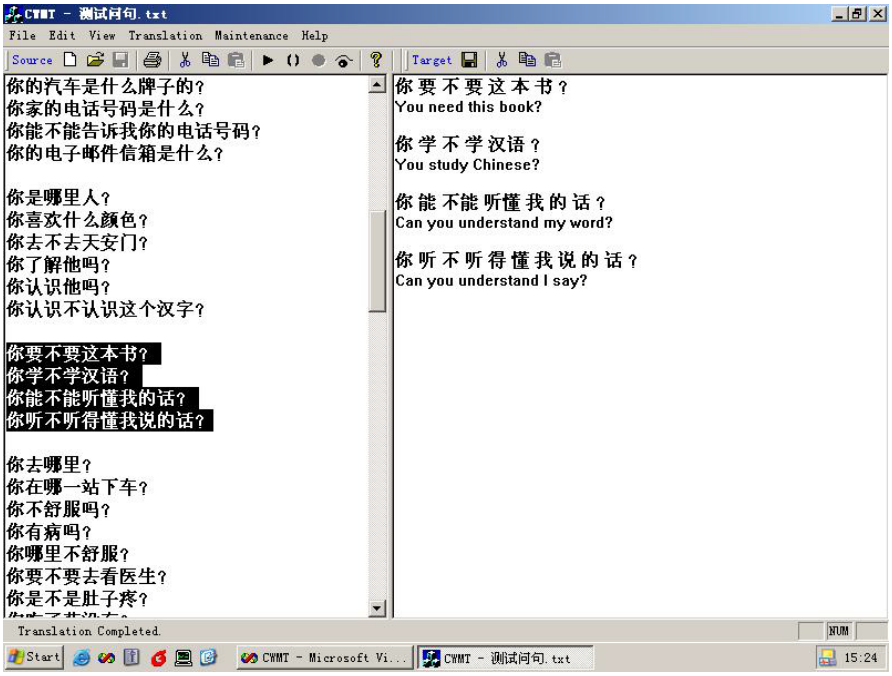

Figure 1 THCE 2.0 Translation System

2. English to Chinese Translation System (THEC 1.0)

THEC 1.0 is another system that accomplished the translation from English to Chinese. The system makes use of the technologies of rules-based and statistical method to complement the weakness of each other's.

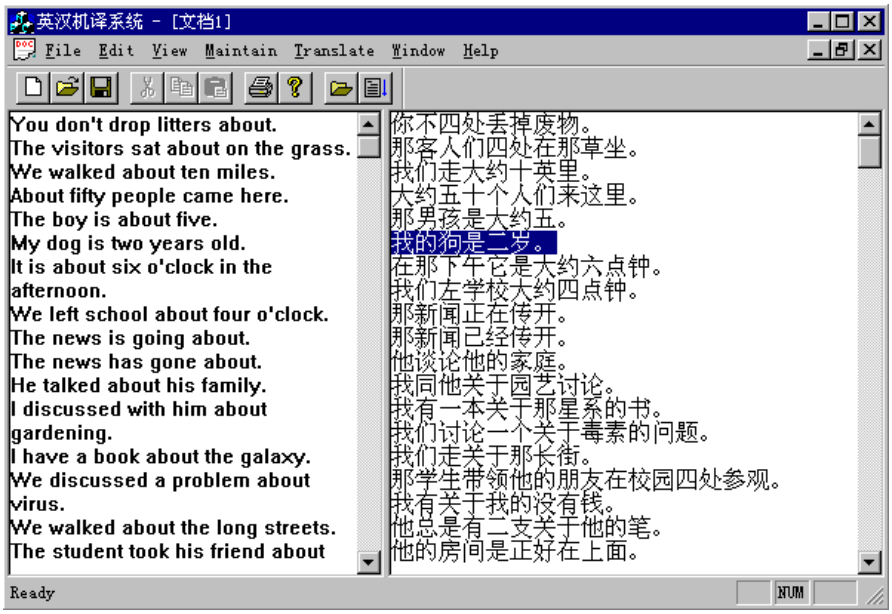

Figure 2 THEC 1.0 System Interface

3. Portuguese/Chinese Word-by-Word Translation System (PCT 2.1)

Considering the coexistence of Chinese and Portuguese in Macau is a unique characteristic. Most of the official documents and legal codes are in Portuguese, there is a great demand in translation of official documents into Chinese. However, there is no such translation software

available in the world.

Figure 3 PCT 2.1 Main Interface

Based on this consideration, Tsinghua University, INESC-Macau and University of Macau had jointly started a research project in Portuguese-Chinese bi-directional machine translation. PCT 2.1 is the achievement of the first phase in this research activity. This system has promised and created an environment to assist users in their daily work.

4. Portuguese/Chinese Sentence-by-Sentence Translation System (PCT 2.0)

PCT 2.0 is the continuing of PCT 2.1 in extending the functionality to sentence translation. This system is under developing, and is planning to be distributed on the beginning of next year.

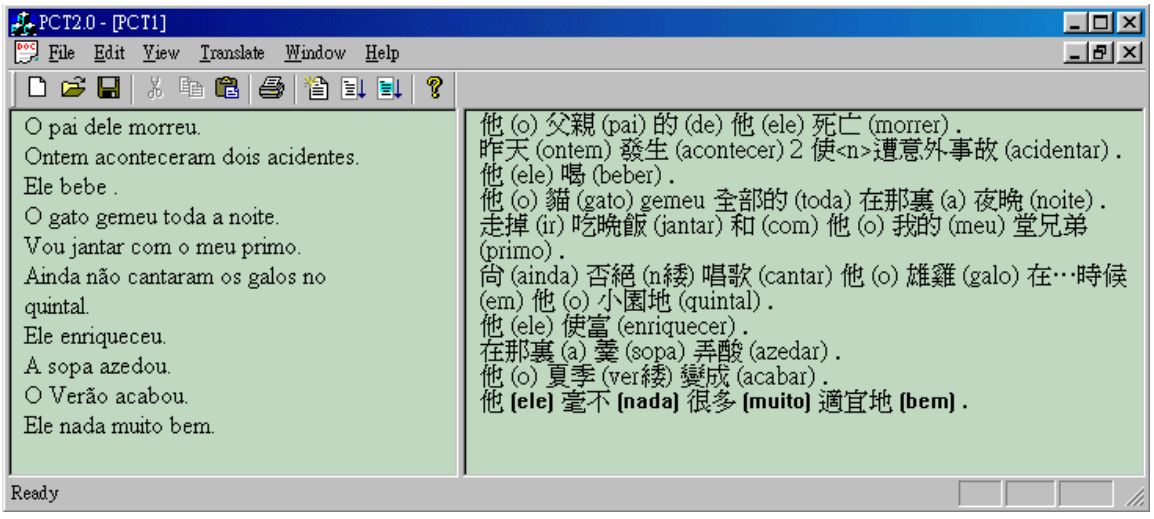

Figure 4 PCT 2.0 Translation System

5. French-Chinese Word-by-Word Translation System (FET 1.0)

This joined project is cooperated with department of national education of France (Ministère

Éducation Nationale). FET 1.0 is the resulting system of the first phase of the cooperation. Similarly, the system has created a translation aid environment.

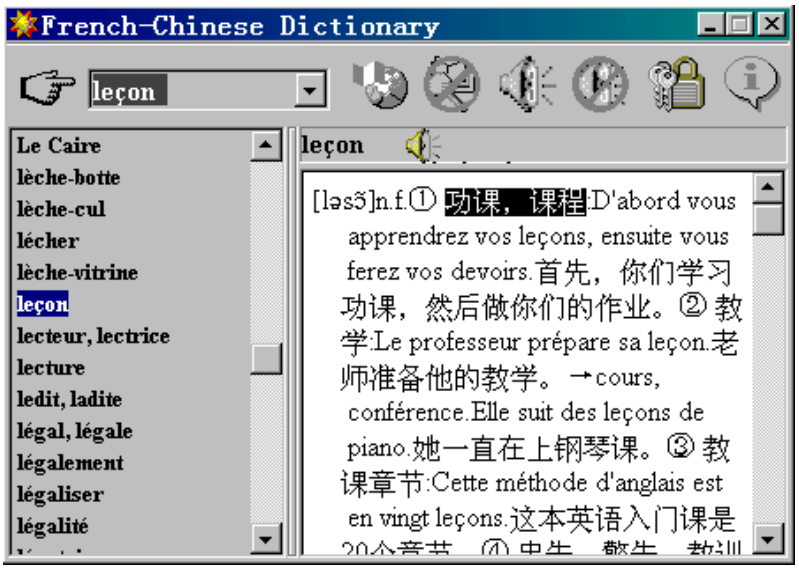

Figure 5 FET 1.0 System Interface

6. Chinese-French Sentence-by-Sentence Translation System (CFMT 1.5)

FET 2.0 is the target system of the research project in Chinese-French translation, and is currently being developed in our center. Chinese analysis and French generation is the main studied. Aiming at translating formal written Chinese, a primarily extensible and practical Chinese-French translation system has been developed. The system adopts post-priority maximum match (PPMM) method to segment Chinese text. The independent of rules database from the program makes the system easy to maintain and extend. The system has been tested by some simple sentences and shown its reasonable.

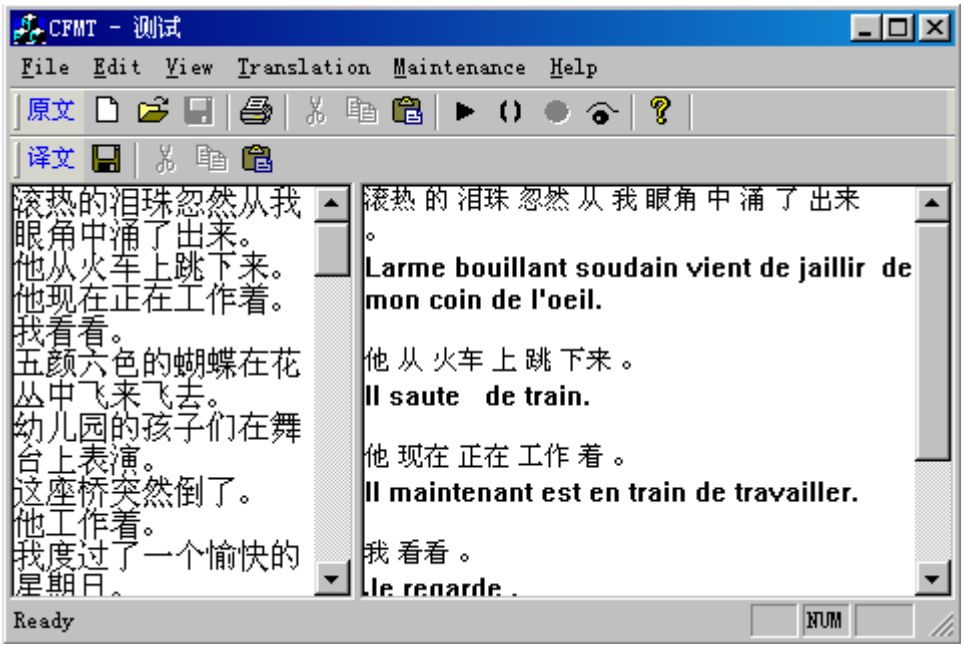

Figure 6 CFMET 1.5 System

CONCLUSION

As the explosive growth of information in the Internet, translation tools have been urgently demanded to resolve the communication barriers between peoples of different nationalities. Especially in China, as the development of the cities continues, it creates a lot of business activities with foreign countries. So it is quite necessary to find a suitable tool for languages translation. However, there are very few commercial, Chinese to European, and European to Chinese, translation systems. This gap should be filled, as there is a demand from the market.

Machine translation is one of the subjects in artificial intelligence. It should be working in someway of imitation of human's brain. We are quite interested in exploring the question of, what is the type of knowledge structure and way of thinking in human's brain? Is the structure of knowledge and thinking in a form of rules or databases? In other hand, how is the generating process of human language in brain? If it is a kind of database, then we may say that human's knowledge and thinking are infinite. But if it bases on rules, then we may conclude that the rules for sentence generating from words and phrases are limited, and the patterns of sentence are finite also. Therefore, what we need to do is preparing a rich lexicon dictionary, a set of rules for source and target languages, and together with algorithms of realization. In this way, we can produce a better system of machine translation.

REFERENCE

[1] Julius T. Tou, An Intelligent Full-Text Chinese-English Translation System, An International Journal of Information Sciences, Elsevier, 1998. pp. 1-18

[2] 姚天顺等 著, 自然语言理解 - 一种让机器懂得人类语言的研究, 清华大学出版社, 广西科学技术出版社, 1993.