The Research and Development of Machine Translation in China

Aiping Fu
Institute of Linguistics
Chinese Academy of Social Sciences
100732, Beijing China
E-mail: fuap@linguistics.cass.net.cn

Abstract

This survey of the situation regarding the R&D of machine translation in China concentrates on how the design and application of MT systems could be on the shadow of the characteristics of Chinese Language. After a brief historical overview and the investigation on R&D environment, some technical features are described, including commercialization, Chinese-English language pair and a multi-level transfer in English-Chinese MT systems.

1 A brief overview on the MT development history

Natural language information processing (NLP) in China started from the research on machine translation (MT). As early as in 1956, MT was identified as one of the research subjects in China's first Science Development Plan. The project was entitled "Machine translation, rules of natural language translation and the mathematics on natural languages". In 1957 the Chinese Academy of Sciences (CAS) engaged in a study on Russian-Chinese machine translation system and, in 1959, CAS conducted a series of experiments successfully — nine sentences with different pattern and relatively complicated structures were translated. The research then proceeded to English-Chinese language pair. Until the mid of 1960's, there had been some six research institutions and universities involved in MT research (Liu Y., 1984).

After more than 10 years of standstill, China's MT recovered in mid 1970's. English-Chinese was the first language pair for the new studies, which were led by the Institute of Linguistics of the Chinese Academy of Social Sciences. Two experimental systems were developed later as the result of the studies. The first was a translation system for titles of technical documentation, followed by the full text translation system for S&T documentation (Liu Z., 1981). On this basis, a number of projects were launched and carried out, focusing on other language pairs, including French-Chinese, Japanese-Chinese, German-Chinese and Russian-Chinese. An experiment for translating Chinese into a cluster of languages including English, French, Japanese, Russian and German was also carried out, using the ARIANE of GETA in Grenoble, France (Feng, 1984).

The first commercial English-Chinese MT system, named TranStar was put in the market in 1988 (Dong, etc. 1986). More than 10 MT products became available in market since then. The English-Chinese language pair is targeted in these new systems, and to a less extent to other pairs such as Japanese-Chinese and Chinese-English. Web-oriented translation service has been tried by some producers. During the same period, there was also a fast development in machine translation in Taiwan. Among the existing products and services in market, Behavior Design Co.'s translation service has been regarded as a successful case (Chen, etc. 1991). It should also be mentioned that in Hong Kong, MT systems are also developed for Chinese dialect and Mandarin (Putonghua) (Zhang, 1998).

2 Research and development (R&D) environment

In China, the initial support for MT research came from the Government. Funds was provided through the 1950's National Science Development Plan and the following National Five-Year Economic Development Plans, as well as from National Natural Science Foundation and the 863 High Tech Project started in late 1980's. Since the later part of 1980's, industry and business sectors have begun investing in this field.

The cooperation between research institutions and business companies has speeded up MT development process, and has shorten the time for bringing research achievements into marketable products and services. International cooperation was also emerging. During the period of 1987-1993, a joint project for five Asia language translation system (Japanese, Chinese. Malay, Thai and Indonesian) was financed by Japan. A number of research institutes and universities were involved in this research project. The technical exchanges and training provided through this cooperation were certainly instrumental in furthering the development of MT in China.

The Chinese researchers have established their academic association – Chinese Information Association of China (CIAC). The Natural Language Processing Committee and Computational Linguistics Committee of CIAC have been organizing national seminars since 1991, once for every two years. The seminars covered various subjects on MT including system design, language analysis and generation, rationalistic and empirical approaches in MT research

and evaluation of MT output. The events offered opportunities for demonstration of new products and experimental systems. The participants for the seminars include also researchers from Hong Kong, Macao, Taiwan, Japan and USA.

Due to the demands from market, the most common language pair in China's MT systems is English-Chinese[1]. Followed is the Japanese-Chinese pair[2]. At present, the systems for other language pairs, such as German-Chinese (Di, 1995), Russian-Chinese (Li, 1999), French-Chinese are still at experimental stage. There are already products on Chinese-English, but the quality of the translations remains non-remarkable. By the end of 1998 a Chinese product of English-Japanese MT system (Ma A., 1999) was announced in Japanese market. This is China's first non-mother tongue MT system.

Indo-European and Chinese languages differ from each other enormously in terms of morphology, syntax and semantic expressions. Therefore many common strategies and algorithms already developed in Indo-European MT systems can not be adopted directly for the system involved in Chinese. It is for this reason that Chinese scholars have been obliged to study widely the linguistic theories, and develop MT technologies appropriate to the special characteristics of Chinese language.

Most Chinese MT systems are rules-based especially those commercial systems. In order to improve the quality of translation, some researchers established very detailed and refined descriptions on the usage of words, so as to do disambiguity by use of lexical properties. Starting from early 1990's, the empiricism began to influence MT research in China. The researchers have tried more and more corpus-based approach and, as result, some example-based and statistic-based systems were experimented. Now a combined rationalistic and empirical philosophy in MT design has been proposed and explored (Huang, 1999; Chen, 1998).

Due to language divergences, Chinese MT could not use direct approach to the task of translation. Most of our systems adopt transfer approach—an hybrid transfer which is not as same as what people are familiar with in Indo-European Language MT (we will talk about this in next section of this report). Some have used an interlingua method (Xiong, 1998), but only limited to experimental stage.

Along with the development of systems, we also worked in the evaluation of MT systems. An automatic translation evaluation system was put into test in 1991 (Yu, 1993) examining English-Chinese MT output. This system was designed on the basis of a set of 'isolated testing points' in accordance to lexical meaning, morphology, syntax as well as semantics. The system gathered more than 3000 sentences as a testing set in which those testing points are properly distributed. The system can automatically score translations output from MT systems, so as to evaluate their performances. A number of Chinese MT systems have been evaluated by this system. In addition, there has been artificial evaluation tailored with user-specific requirements (Fu, 1995). Other practices of evaluation include the National 863 Project that organizes each year an evaluation of texts translated by MT systems of China.

Whether it is 'readable' or 'understandable' is a crucial factor in determine the quality of translation and evaluating the Chinese text output by an MT system. Unlike English, that is confined with 'subject-predicate' framework and grammatical agreement in sentence structures, Chinese language has parataxis in sentence construction and is more flexible in word order. Owing to the characteristics of Chinese language, the 'readability' and 'understandability' are less restricted or, more error-tolerant, than it is the case in English and other Indo-European languages. In English-Chinese MT, as long as English sentence structure is fully explored, most Chinese translations can be made 'readable' or 'understandable', because of the most Chinese recipients' 'parataxis ability' in reading.

This 'parataxis' in Chinese language has helped MT systems in producing acceptable translations from foreign languages to Chinese, even without support from the precise Chinese linguistic models, which are still very challenging research subjects for Chinese language. With a cost of 'machine accent', some MT systems have become marketable products in China.

3 Technical features

3.1 Commercialization

When MT Products first came to market, the producers and end-users had over expected their performance. Connected them to pocket electronic agenda in the hope to assist children to learn foreign languages, for example, showed how unrealistic the expectation was. Frustration followed, naturally. Today, MT market in China has become much more realistic and sensible.

Most present MT products are PC-based and are used primarily for translation of technical documents. They are often supported by a set of terminology databases (sometimes a system may have more than ten such databases) to facilitate translation of documents in different fields. With proper 'training' to the databases on particular translation subjects, some products can be used with information retrieval systems for specific domain to produce 'draft' translations. With proper post-editing by people, these translations may be published (case already exists). Other examples include translating technical documents for enterprises, with the help of Translation Memory. Those task-specific MT products are likely to achieve satisfactory, as is showed from China's experience. It is especially true when the experts from the particular fields or companies are involved in the construction of the terminology databases and updates, and in adjusting linguistic rules of the systems.

China has also developed a few systems that are web-oriented. Clearly, these systems are far from mature in dealing with the very broad domains and diverse writing styles encountered on Internet web pages. As the number of Internet users are mounting

rapidly in China and many of them have barriers in reading foreign languages, these Web-oriented MT tools also possess a quite large group of users. Although the translated web pages are often found with broken sentences and grammatical errors, for Internet surfing, they can help in some way.

3.2 Situation regarding Chinese-English MT development

In the research of machine translation from our parent language to foreign languages, Chinese-English language pair has been most frequently involved. Compared with English-Chinese pair, the reversal MT system has progressed slowly. This is due to the fact that it is more complex in analyzing Chinese sentences. Transferring Chinese to English is equally complicated.

Let's look at this example. We know that there is little confusion in English as regard to 'word'. In Chinese, however, there is neither a clear definition on 'word', nor a fixed formal indication. Between Chinese morpheme and word, word and phrase, phrase and sentence, and between sentence and sentence group, there is also no clear 'boundaries'. A same category of Chinese words can be used for different grammatical functions without any morphological change. For instance, a Chinese noun in the same word form can be used for subject, object, predicate, attribute and adverbial. Considering the underlying structures of Chinese sentences, we would find that one syntactic structure often expresses different semantic contents, while one semantic expression could be realized by several syntactic structures. A contemporary Chinese textbook (Qian, 1990) has listed 16 semantic expressions which could be realized by either subject-predicate construction or verb-object construction in syntax. This many-to-many correspondence between semantic and syntactic is widespread in Chinese, which causes big uncertainty on its analysis for machine translation. So far there is no effective solution developed to the problems discussed above which are only part of the difficulties we have met in analyzing Chinese text.

Even though the analysis of source language manages to achieve a correct underlying representation of Chinese sentence, transfer and generation of English sentences remain very difficult. To get a surface string of English sentence from an underlying representation of Chinese sentence, we must add some new information—the number of nouns, the tense, aspect and the voice of verbs, the article for nouns, and the conjunctions, which are not included in the source. This is also a very difficult problem as the knowledge of meaning, context, and even common sense are needed.

Chinese scholars have been nevertheless continuing the research in this field, although the progress has been slow (Wu, 1992; Liu Q.,1998). Some systems can already translate acceptable sentences from Chinese to English.

3.3 A multi-level transfer in English-Chinese machine translation

Most English-Chinese MT systems developed in China have adopted transfer approach which differs from its conventional version based separately on either of the processing levels of lexicon, structure or semantics in a system. Two distinctions exist. In one the transfer has been performed at all possible. processing levels, from lexicon, structure to semantics. In the other, two kinds of transfer rules, general and specific, have been employed. The former focuses on syntactic or semantic correspondence between source and target languages. The later defines the transfer mapping that has to be described on the basis of lexical properties.

Belonging to different language families, English and Chinese have a great distinction both in their vocabulary and grammar. Transferring at several possible levels and on different kinds of rules is motivated by the demand for resolving, in every way, various translation divergent problems between these two languages.

3.3.1 Transfer at lexical level

At the stage of dictionary consulting and morphological analysis, idiomatic expressions and word groups are directly converted from English to surface string of Chinese sentences. After that, the whole expression or word group is reduced to a single node, disregarding its internal lexical structure and retaining in the syntactic structure of the source for the oncoming syntactic and semantic analysis. Obviously this kind of rules is specific since the lexical choices and reconstruction of nodes they define depend on lexical tokens.

3.3.2 Transfer at syntactic level

Sometimes the structural divergent problem between source and target language can be resolved in a lexical-semantic environment. That is to say, some English phrases can be structurally transferred disambiguously to their Chinese equivalents with transfer rule specific to each phrase. Such rules are what we call specific rule built on each lexical token.

Specific transfer rules are usually incorporated with the rules for source language analysis and are performed immediately after the underlying structure of English phrases have been achieved with enough information available for transfer mapping. Specific transfer rules are directional, and carry quite precise message about surface string of target language, including syntactic structure, semantic roles, lexical choice, position of categories that should be deleted or inserted in the sentence structure, and even surface word order. Transfer like that is actually connected to some extent with target language generation, so that sometimes syntactic-semantic construction of an English phrase may be transferred to a shallower surface structure in Chinese (see Figure 1), instead of transferring between source/target representations with equal abstraction. This scheme does not look elegant as it may more or less cause transfer and generation to

be on the shadow of source analysis. But it will also make full use of information obtained from source language analysis, and avoid a severe proliferation of information redundancy possibly occurred in an independent transfer process.

Structural transfer also needs general rules which are based on syntactic constituents. By use of such rules, English sentence patterns can be converted into their Chinese equivalents.

At syntactic processing level of a system, both specific and general rules are employed for structural transfer. The former have priority of being triggered and generally applied for internal structure of phrases, and the later are applied for the structure of sentences. For the syntactic tree of a sentence, specific rules are usually imposed on the leaves or branches near leaves, while general rules on sub-trees or the whole tree.

3.3.3 Transfer at semantic level

Syntactic information is sometimes inadequate for transfer mapping between sentence structures of English and Chinese. For instance, how could we decide syntactic positions of modifiers before their head in a noun phrase of Chinese? Some Chinese grammarians believe that in a Chinese subordinative construction, adjectives tend to be arranged in the order that reveals to what extent they are related in meaning to the noun they modify. Adjectives with the meaning that are closer to the noun will have position nearer the noun, while those that have relatively more distant relationship will be put further (Ma Q., 1995). Such an arrangement of word order can be described by semantic relation of each modifier to its head. For example:

possessive < state or descriptive
< property < shape < NOUN

Among them, "property" could be further sub-ordered by semantic features:

nature or character < sense
< geometric < colour

Note: the symbol "<" means being ahead of .

These rules are general which provide transfer mapping from semantic expression of source language to surface syntactic structure of target language, and are usually implemented on the basis of structural transfer discussed in the previous sub-section.

The transfer approach described above—performed at several levels and using specific and general rules—is called multi-level transfer (Figure 1). Having been studied and implemented for many years, this approach has gradually taken shape and been widely accepted by quite a number of English-Chinese MT systems developed in China. It has also been employed in other foreign to Chinese language pairs such as German-Chinese (Di, 1995) and Russian-Chinese (Li, 1999).

4. Development trends

It is widely recognized that there will be a large market in 21st Century in China for MT technologies and products. Some years ago when MT products started entering market, the producers targeted the individual users. Owing the large language barriers of Chinese people in reading and writing foreign languages, there was indeed a strong interest showed by individual users. This did not, however, last very long. The large difference in specific individual demands showed that most of MT products today are not yet capable of meeting all their needs. It will take some years for multiple-use MT system to come.

A realistic MT development approach is, perhaps, to focus on assimilation of information for multi-use systems development, and to focus on information dissemination for specific-users. Individual users for the demand of occasional translation and Web surfing belong to the former, as well as translation of technical document and translation for various kinds of information access. In that case the output of MT should focus on the delivery of principal information without confusion to the recipients. The latter is referring to translation in well-defined specific domain or users, or in controlled languages. Typical examples include the translation of company documentation such as technical manuals, as well as localization of software.

While machine translation is entering many and widening application domains, on the line of technical development it is challenged with more profound and complicated issues. This is particularly regarding precise source language analysis and target language generation. Dong Z. (1999) has proposed two technical MT development subjects: multiple-sentence processing and knowledge-based language understanding. The aim to pursue these two subjects is to establish 'meaningful connections' between the sentences so as to make sure that source text is really understood, and generation of target text is made on this basis. For this, Dong has constructed a Chinese-English bilingual common-sense knowledge database. It is clear, however, to reach these objectives a long voyage for research is still expected. Let us take this as a new MT research paradigm in the coming century.

References

Chen, Qunxiu. (1998) "Research of the Robustness in Japanese-to-Chinese Machine Translation System". *Proceedings of 1998 International Conference on Chinese Information Processing,* pp. 490-498. Beijing: Tsinghua University Press.

Chen, S., Chang, J., Wang, J., Su, K. (1991) "Archtran: A Corpus-based Statistics-Oriented English-Chinese Machine Translation System". *Proceedings of MT Summit III,* pp.33-40.

Dong, Z., Zhang D. (1986) "KY-1 English-Chinese MT System". *Library & Information,* China, Vol.3, No.4.

Dong, Zhendong. (1999) "Bigger Context and Better Understanding—Expectation on Future MT

Technology". *Proceedings of International Conference on Machine Translation & Computer Language Information Processing, pp.* 17-25.
Di, H., Chai, P., Xu, Y. (1995) "The Conversion and Generation between German and Chinese Clauses Based on Corpus and Rule base". *Advances and Applications on Computational Linguistics,* pp. 264-270. Beijing: Tsinghua University Press.
Feng, Zhiwei. (1984) "Experiment in Automatic Translation from Chinese to French, English, Japanese, Russian and German". *Machine Translation in China,* pp.103-184. Shanghai: Knowledge Publishing House.
Fu, Aiping. (1995) "The User-Oriented Evaluation of MT Systems". *Advances and Applications on Computational Linguistics,* pp. 245-250. Beijing: Tsinghua University Press.
Huang, H., Chen, Z., Song, J. (1999) "The Design and Implementation Principle of an Interactive Hybrid Strategies Machine Translation System IHSMTS". *Proceedings of International Conference on Machine Translation & Computer Language Information Processing,* pp. 270-276.
Li, X., Zhou, Q. (1999) "The Principles of Designing a Syntactic Rule Base of Russian-Chinese Intelligent Machine Translation System". *Journal of Chinese Information Processing,* Vol.13, No.l, pp. 16-19.
Liu, Qun. (1998) "Discussions on the Difficulties of Chinese-English Machine Translation". *Proceedings of 1998 International Conference on Chinese Information Processing,* pp. 507-514. Beijing: Tsinghua University Press.
Liu, Yongquan. (1984) "Machine Translation in China". *Machine Translation in China,* pp.1-14. Shanghai: Knowledge Publishing House.
Liu, Zhuo. (1981) "JFY-II English-Chinese Machine Translation System". *ZHONGGUO YUWEN,* No.3, pp. 216-220; No.4, pp.279-285.
Ma, Aijun. (1999) "The Japanese Generating Model Based on the Transformation from P-tree to D-tree in an English-Japanese Machine Translation System". *Proceedings of International Conference on Machine Translation & Computer Language Information Processing,* pp. 289-297.
Ma, Qingzhu. (1995) "On the Order of Adjectives and their Classification in the Multiple Attributive Structures". *ZHONGGUO YUWEN,* No.5, pp. 357-366.
Qian, Nairong. (1990) *Contemporary Chinese,* pp. 233-251. Beijing: Higher Education Publishing House.
Wu, Weitian. (1992) "The principles of Grammar and Transformation in SinoTrans". *Proceedings of 1992 International Conference on Chinese Information Processing,* pp. 274-279.
Xiong, Wenxin. (1998) "The Processing of Chinese Generation Rules Based on Interlingua". *Proceedings of 1998 International Conference on Chinese Information Processing,* pp. 515-523. Beijing: Tsinghua University Press.
Yu, Shiwen. (1993) "Automatic Evaluation of Output Quality for Machine Translation Systems". Machine Translation, (8)1-2, pp. 117-126.
Zhang, Xiaoheng. (1998) "Dialect Word Processing in Cantonese-Putonghua Text Machine Translation". *Proceedings of 1998 International Conference on Chinese Information Processing,* pp. 499-506. Beijing: Tsinghua University Press.

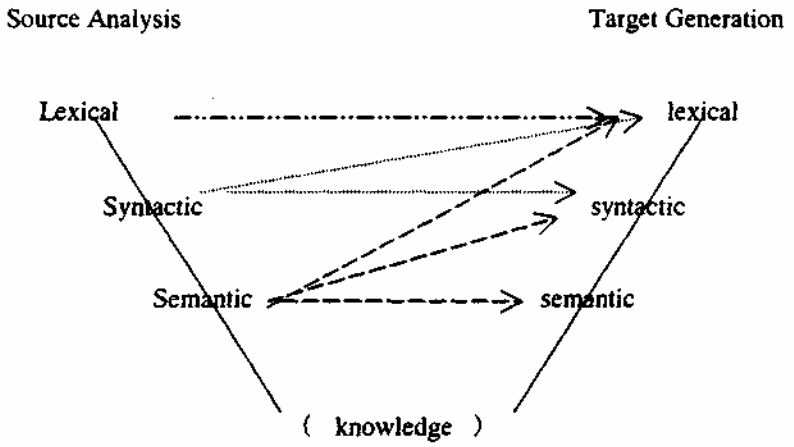

Figure 1

[1] These English-Chinese MT systems are either MT products (marked with * as following) in Chinese IT market or well-developed research projects. They are listed below with their first designers or investors:

*Transtar English-Chinese MT system	Prof. Dong, Zhendong
*JFY(Gaoli) English-Chinese MT system	Prof. Liu, Zhuo
*IMT/EC English-Chinese MT system	Prof. Chen, Zhaoxiong
TYECT English-Chinese MT system	Prof. Wang, Guangyi
TECM English-Chinese MT system	Mr. Liu, Xiaoshu
TH(Tsinghua) English-Chinese MT system	Prof. Chen, Shengxin
*NetTrans English-Chinese MT system	Prof. Wang, Huilin
*SuperTran English-Chinese MT system	Dr. Shi, Xiaodong
*HansBridge English-Chinese MT system	Creative Next Technology Ltd.
*JiShiTong English-Chinese MT system	Moon Computer Company
*TongYi English-Chinese MT system	Tongyi Institute of MT Software
*EastExpress English-Chinese MT systems	Shida-Mingtai Computer Ltd.
*YaXin English-Chinese MT system	Yaxincheng Computer Software Ltd.

[2] Three of them are listed below:
The system developed by Prof. Wang Qixiang of Nanjing University
The system developed by Prof. Hou Fang of Heilongjiang University
The system developed by Prof. Chen Qunxiu of Tsinghua University