China's MT in the 1970s and the 1980s

-- from revival to prosperity

Zhendong Dong

1. Resurgence in the late 1970s

1.1 Beginning of resurgence

MT research and development in China can be dated back to as early as the mid-50s. China was the fifth country in the world which started MT. In 1959 China demonstrated its first successful experiment in a Russian-Chinese MT system. China started its MT research in the same way as that its counterparts did, but it suspended, though also around the mid-60s, in a way with unique "Chinese characteristics". The great chaos caused by the unprecedented Cultural Revolution ruined almost all the fields of China's scientific research. Thus MT suffered from a 10-year "quietness" from the mid-60s to the mid-70s.

MT research in China revived in the mid-70s. Institute of Scientific and Technical Information of China (ISTIC) led China's MT revival. In 1975 the Institute proposed an ambitious project of MT research and development. They expected to apply their MT system in the publication of its periodicals of technical abstracts and tried to push the application of MT in the provincial institutes of scientific and technical information throughout the country. Thus as the leading institute, they assembled and organized a joint MT research team. The team members were mainly selected from the institutes of scientific and technical information of some ministries or cities. They invited Yongquan Liu and Zhuo Liu, China's MT pioneers, from Institution of Linguistics, Chinese Academy of Social Sciences, as the supervisors of the project. Hence China's MT research in the mid-70s was characterized by its clear-cut application orientation and broad collaboration. Two research teams led separately by Yongquan Liu and Zhuo Liu were formed. To begin they selected 10000 English-Chinese bilingual aligned titles of technical abstracts from INSPEC database as the training data. They chose 200 English titles from the training data as the close testing data and 200 unseen titles as the open testing data for each experiment. Both the systems were rule-based and direct approach. They used constituent analysis grammar. With the limited capacity of the computer they used at that time, they had to build a small-size dictionary and had to endure very slow tuning and testing. In 1978 ISTIC sent two young scholars, Zhiwei Feng and Ping Yang to GETA, Grenoble University and they were trained with ARIANE-78 under the instruction of Bernard Vauquois. They developed a Chinese-to-many MT system based on ARIANE-78, called FAJRA. Nearly two years later, in the early 80s, Feng and Yang came back to China with their MT system. The most significant impact of the GETA's MT was the techniques of second-generation MT, as Vauquois claimed, featured by the separation of linguistic data and programs.

What merits special mention is that 1979 saw China's first national conference on MT. The succeeding annual conferences in the early 80s resulted in the set-up of a special interest group on MT which was affiliated to Society of Information of China and then to Society of Chinese Language Information of China. In this period of revival a special MT quarterly was published for the first time in China.

1.2 New comers

In 1977 two MT teams were formed in two universities, Heilongjiang University (HU) and Harbin Institute of Technology (HIT) in the capital of China's northern province. HIT team was led by Professor Zhen Wang of the Russian department and Professor Kaizhu Wang of the department of computer science. They worked on Russian-Chinese MT. As for the Heilongjiang University MT team, there was an interesting story in its the background. The team had a combination of Zhendong Dong and his two colleagues of the English department and three teachers of the math department. Half a year before the set-up of the team, the three English teachers had been entrusted to do some translation for a gold prospecting company. They translated six thick volumes of manuals about gold mining equipments. In the translation they found so many parts of the manuals were almost of the same wording and same technical terms. After each of the three teachers finished his own part of the translation, they found it was even more difficult to gain consistency in the terminology. A common idea came to their mind when they were discussing the consistence. "Is it possible to use a computer to help us?" But at that time they had not the slightest idea how a machine can do the job, or what is so-called machine translation or automatic translation. Hence they went to the provincial library and they found some very old material in some back-number journals. Then they went to the math department for more detailed information. To the question of "can the computer really do translation", the math teachers' answer was: "yes, it can as long as you can formulate the rules." Thus in the middle of 1978 a part-time and cross-discipline MT team was set-up in Heilongjiang University. As a Chinese saying goes: "newborn calf is not afraid of the tiger". They plunged into a completely new horizon with strong enthusiasm and hope for future success. Soon after the first tiny but encouraging experiments, Heilongjiang University team gained some funds from ISTIC and the team frequently visited ISTIC and did experiments there. So the team made fast progress in their research. HIT team also gained financial support from the former Ministry of Aviation. The two MT systems were both rule-based and direct approach, but Heilongjiang University team's system placed more emphasis on semantics. The revival of MT and some early eye-catching progress invited enthusiasm and hope. In the early 80s more new comers cropped up. Two research teams were formed in the Institute of Operations Research of Military Academy of China, one was a team of information retrieval research and the other was a team of MT research. At the end of 1981 two key members, including Zhendong Dong, of Heilongjiang University team, joined the new MT team in the Military Academy.

1.3 KY-1 English-Chinese system

1.3.1 China's first MT product

In fact the team of Military Academy was formed before Dong joined in. The team invited Zhuo Liu to give its members overall training, including linguistic issues for MT as well as programming. When Dong joined the team they started their new project named KY-1 English-Chinese system. The team made experiments on a FACOM computer with COBOL as its programming language.

KY-1 system started in 1982 and ended in 1987. In the end of 1986 a Japanese MT delegation headed by Mr. Yoshihide Tsuji, executive director of Center of International Cooperation for Computerization (CICC), Japan, visited Military Academy and watched the demo of KY-1 system. The delegates also had some technical discussion with Zhendong Dong and his team. The

Japanese researchers found quite a lot of technical common ground. This event pushed the future collaboration in MT between Japan and China.

Even before the system finished, it was transferred to the Corporation of Software Technologies of China (CSTC), which was a newly established company of software affiliated to the former Ministry of Electronic Industry of China in early 80s. After a year of development for commercialization, in 1988, KY-1 system, now renamed as Transtar English-Chinese MT system, became China's first commercialized PC-based MT product.

Transtar had a general dictionary of 50000 entries and three domain-specific dictionaries with over 30000 entries each. In 1989 KY-1 system was rewarded National Award for Advances of Science and Technology. Transtar was demonstrated in a lot of computer shows and fairs, such as in Guangzhou, Singapore, Hong Kong and Hanover. At the international conference on translation in Hong Kong in 1987, Archtran English-Chinese MT system developed by K.Y. Su of Tsinghua University of Taiwan came into sight and was demonstrated together with KY-1.

1.3.2 Technical features of KY-1

KY-1 system took the rule-based transfer approach. It was characterized by its well-formed linguistic theory and its unique SCOMT with a powerful interpreter that achieved separation of linguistic data and the program.

Linguistically, KY-1 system adopted logical semantics, which was put forward by Zhendong Dong in 1978 when he was developing his first MT system in Heilongjiang University team. Dong's logical semantics for MT had two prominent features:

(a) the separation between the analysis of source language (SL) and the generation of target language (TL);

(b) the transfer was based on semantic roles rather than syntactic elements.

When developing MT system in Heilongjiang University, Dong, proposed the theory of logical semantics and applied it in their MT system as the transfer basis – the outcome of the analysis of source language (SL) text and the basis for the generation of target language (TL) text. Heilongjiang University's system did not adopt syntactic elements, such as subject, predicate, object, adverbial, etc., instead it adopted over 50 logical semantic roles, such as agent, patient, time, duration, source, target, space, etc. Obviously in Heilongjiang University team the analysis of SL was done in a deeper level. The following may be a good example to illustrate the depth of the analysis:

SL text:

its analysis of the data by the computer

Analysis results:

Word	**Semantic role**
analysis	pivot
data	patient
it	agent
computer	instrument

The logical semantics was introduced to KY-1 system in 1982.

1.4 JFY English-Chinese MT systems

JFY MT systems were developed by a MT group of Institute of Linguistics, Chinese Academy of Social Sciences. Professor Zhuo Liu was the group head.

Up till the 1980s JFY referred to 4 MT systems named JFY-I, JFY-II, JFY-III and JFY-IV. The development of JFY-I began in 1976. The first three JFY systems were all experimental ones while JFY-IV was a market-oriented one. Its development began in 1982. JFY systems were rule-based and transfer approach. JFY-VI was characterized by its so-called “background semantic base”. The base was designed on two principles: distributive features and thesaurus-like hierarchy. After several years testing and debugging the system was greatly improved and it finally developed into JFY-V and became a new MT product in China.

One of the key members in Liu’s MT group, Li Wei was involved in international collaboration in MT relevant to Esperanto. In the summer of 1989, Li, Zhuo Liu and Zhendong Dong were invited by Toon Witkam, head of BSO’s MT group to participate in their MT seminar held in Utrecht, the Netherlands. At that time BSO was the Netherlands’ second biggest software company. Toon Witkam and his MT group attempted to develop a multilingual MT system using Esperanto as the interlingua instead of any formal intermediate language. In the seminar BSO team gave detailed presentation with demo of their DLT (Distributed Language Translation) system. We learned for the first time some initial application of large-scale bilingual corpus and statistics in MT.

1.5 IMT/EC English-Chinese MT system

IMT/EC English-Chinese MT system was developed by a MT group in Institute of Computing Technology, Chinese Academy of Sciences. The group head was Chen Zhaoxiong. In the 1980s the development of the system was in its beginning stage including the conceptual and architecture design. However the system was completed in the early 90s and it has become the most successful and profitable MT product in China. Based on this product China’s biggest MT company was established in the early 90s.

2. MMT project

2.1 Far-reaching impact

The involvement in MMT project was the most significant event in China’s MT history. In 1987 a MT lab was set up by Center of International Cooperation for Computerization (CICC), Japan, and a joint multilingual machine translation project (MMT) was launched by Japan. The project involved Japan and its neighboring countries, including China, Malaysia, Indonesia and Thailand. The project had the following specifications:

1) translation fields
   mainly industrial and technical information;
2) translation languages
   between Japanese and Tai, Japanese and Malaysian, Japanese and Indonesian, Japanese and Chinese;
3) translation accuracy
   80-90% (with pre-editing);
4) term of the project
   6 years (1987 - 1992);

5) elements of research and development
   formulating of basic plan, interlingua, text analysis, text generation, electronic dictionaries, input/output systems, translation support system, integrated system, standardization, operation and dissemination;
6) evaluation and report of the result;

Organized by CICC's MT lab, almost all the Japan's key computer companies were involved in the project, such as Fujitsu, NEC, Hitachi, Toshiba, Sharp, Mitsubishi, etc. They undertook the above R&D elements respectively, for example, NEC was in charge of text analysis system, Toshiba was responsible for construction of technical dictionary.

The organizer of the project on China's part was the Corporation of Software Technologies of China (CSTC), which was a newly established software company. Six Chinese universities and research institutes joined in the project. They were the Chinese Information Labs of Ministry of Electronic Industry, the Institute of Scientific and Technical Information of China (ISTIC), Northeast University, Nanjing University, Beijing Institute of Languages, and Chinese People's University. They also undertook the above R&D elements respectively. In 1987 Dong transferred to CSTC from Military Academy and was appointed as the Chinese technical leader in the project.

The MMT project had a far-reaching impact not only on China's MT research activities, but also on its R&D of natural language processing. MMT was believed to be as a NLP training course, as over 60 researchers joined in so many subject areas of NLP. The project as an incubator nurtured a lot of new MT labs in China. The project helped to accumulate rich NLP resources. Moreover, the project provided Chinese researchers a good chance to enter the international environment. This is especially valuable to China, as it had been kept closed from the rest of the world for nearly 30 years. In the first MT summit held in Hakone in 1987, Chinese scholars suggested hosting the 1992 COLING and had initial discussion with some COLING committee members. The proposal was officially passed by the COLING committee in Budapest in 1988. But unfortunately the decision was changed in 1989 owing to again to unique "Chinese characteristics".

3. Discussion on some technical issues

3.1 On interlingua approach

The interlingua approach sounds nice and economical for multilingual machine translation, especially for many-to-many MT. However, it is very difficult to specify a really practical and effective representation for the interlingua. When adopting the interlingua approach, two elements should be defined, one is the intermediate representation of the semantic relations among concepts in the sentence, the other is the interlingua among the languages processed on the word level. The lesson from the MMT told us the latter element was even more difficult to handle. Actually MMT adopted English, the definitions of Longman Dictionary of Contemporary English as the interlingua. None of the natural languages could serve as an effective interlingua, because the natural languages themselves were ambiguous.

In MMT semantic relations were adopted as the interlingua for sentence meaning representation. The approach proved to be effective, both for source text analysis and the target text generation.

In conclusion theoretically it is practical to find a common basis on the concept level.

However, translation process should not only go deep into the concept level (in the analysis of the source text), but also should come to the surface level of the words and expressions (for the generation of the target text).

3.2 On knowledge resources

The long experience in working on MT R&D shows that knowledge resources are indispensable to MT as well as to NLP. We need three types of knowledge: intralinguistic knowledge, interlinguistic knowledge, and extralinguistic knowledge. Intralinguistic knowledge incorporates a wide range of conventional linguistic areas such as phonetics, morphology, syntax, semantics and pragmatics. By interlinguistic knowledge, we mean cross-language knowledge, such as comparative linguistics, or the knowledge that we especially apply in language translation or recently in cross-language information retrieval. Extralinguistic knowledge is common sense knowledge or world knowledge. Different tasks need different kinds of knowledge. We should be very clear about how to use knowledge, how to use the right knowledge in the right place. We should make full use of various kinds of knowledge. Dictionaries and rule sets of syntax and morphology, though manually coded, may be the best resources of intralinguistic knowledge, while bilingual aligned corpora are the best resources of interlinguistic knowledge. The best approach to developing of MT systems is to integrate all the usable resources. It is not wise to give up all the old accumulation of resources when shifting to a new technique. MT is rather a practical skill than a kind of academic exploration. Linguistic knowledge is essential to human language technology. But we should adopt a pragmatic attitude toward the theory and algorithm.

HowNet, a common sense knowledge resource, developed by Zhendong Dong and his group, was incubated in the MT nest. As Dong stressed on many occasions: “there would not be HowNet if I had not had experience in MT and MMT.” Let me conclude the article with the preface of his new book “HowNet and the Computation of Meaning”:

> “At this moment to look back into the past and recall some of the great events and figures that exercised influence on the development of HowNet will be of significance. In the late 80s, in Makoto Nagao’s frequent academic visits to China, I translated for him and learned a great deal from his rich experience in building semantic dictionary, especially the principles for semantic classification of nouns and verbs. From 1987 to 1992 when I was the chief technical leader of Chinese team participating in the machine translation project among Japan and other four Asian countries, I learned quite a lot from various kinds of MT dictionaries of Japanese IT companies and labs, especially from Japan’s EDR concept dictionary. In 1988 I was invited by Antonio Zampolli to attend the Summer School on computational linguistics. During the 1-month course, for the first time I learned ontology from Sergei Nirenburg’s lecture, and visited the knowledge base built in the lab of Pisa University. In 1993 when I worked in Tokyo, I visited Hozumi Tanaka of Tokyo Institute of Technology, during the whole morning of discussion, he gave me lots of valuable advice and gave me many papers which described taxonomy or discussed semantic categorization.”

March 2007

Bibliography

Dong, Zhendong, 1988, Knowledge description: what, how, and who? Manuscript & Program of International Symposium on Electronic Dictionary, Nov. 24–25, 1988, Tokyo, p.18

Dong, Zhendong, 1988, MT Research in China: In Dan Maxwell, Klaus Schubert, Toon Witkam (eds.), New Directions in Machine Translation, Budapest, Foris Publications, pp.85–92

Tsujii, J., Linguistic Knowledge, World Knowledge and Conceptual Knowledge, in: Proceedings of the International Symposium on Electronic Dictionaries, in Tokyo, November 1988

董振东，1981 逻辑语义及其在机译中的应用，中国的机器翻译，刘涌泉编，知识出版社，北京，1984，pp. 25–45

冯志伟, 1987 自动翻译，上海知识出版社，上海

刘涌泉, 1984，中国的机器翻译，知识出版社，北京，1984