Prospects of Machine Translation in the Chinese Context

DUOXIU QIAN
Beihang University, Beijing, China
hkqdx00@yahoo.com

RÉSUMÉ
Depuis les années 1950, la Chine fait partie des pays les plus innovateurs dans la recherche et les applications des outils de traduction automatiques et de la traduction assistée par ordinateur. La première partie de ce travail propose un survol historique de l'évolution de ces techniques en mettant en valeur quelques moments importants de leurs développements dans le contexte chinois. Si les résultats des recherches en Chine sont semblables aux autres pays, on dénote cependant une attention de plus en plus marquée pour la traductions assistée par ordinateur. La deuxième partie expose l'état actuel des connaissances à propos de ces deux outils pour le territoire chinois, pour Taiwan et pour Hong Kong. Par la suite, cet article présente les logiciels commerciaux les plus populaires de traduction du chinois aux autres langues et vice-versa en montrant pour chacun leurs qualités et leurs défauts. Finalement, il est question des perspectives d'avenir de la traduction automatique et de la traduction assistée par ordinateur en Chine. Pour le bénéfice mutuel de ces deux domaines de la traduction, les recherches chinoises devraient opter pour une coopération plus étroite avec le reste du monde.

ABSTRACT
China has been among the several leading countries in the research and applications of Machine Translation (MT) and Machine-aided Translation (MAT) ever since the 1950s. The first part of this paper is a historical sketch of MT and MAT in the Chinese context, highlighting some important stages in its development which have laid the foundation for later achievements. It is shown that the research of MT in this region is similar to that in other parts of the world, with the attention gradually turning to MAT. The second part deals with the state of the art of MT and MAT research and applications in Mainland China, Taiwan and Hong Kong, respectively. Then popular commercial software dedicated to the translation from Chinese into other foreign languages, and vice versa, are introduced, with an appraisal of both their merits and demerits. Finally, prospects of MT and MAT in the Chinese context is discussed. It is suggested that, for mutual benefits, MT and MAT research in the Chinese context should cooperate with the outside world more closely.

MOTS-CLÉS/KEYWORDS
machine translation, machine-aided translation, chinese, commercial software

1. Introduction

Following the United States and the former Soviet Union, scholars in China started research in Machine Translation in 1957, shortly after the founding of the People's Republic of China in 1949. China demonstrateed its achievements in machine translation for the first time in 1959 and thus joined the exclusive club in this field (Dong 1995; Fu 1999).

In the 1980s and 1990s, Hong Kong and Taiwan also began earnestly to do research in MT and they have so far made some remarkable achievements.

2. The State of the Art

In this part, the state of the art of MT in the Chinese context, i.e., Mainland China, Hong Kong and Taiwan, will be dealt with respectively.

2.1. Mainland China

There are some distinctive features in the development of MT in China. Firstly, MT research and devolopment in China has been chiefly sponsored by the government since the very beginning. Early in 1956, "Machine Translation /Mathematical Theories for Natural Language" was already an item listed in the Government's Guidelines for Scientific Development. Later it was among the major national scientific and technical projects, such as "The Sixth Five-Year Plan", "The Seventh Five-Year Plan" and "863 Plan". Though there was a ten-year stagnation in MT research in China from 1966 to 1976, it was not because of the shortage of fundings, but because of political and social upheavals during the Cultural Revolutions. Secondly, scholars from various fields and institutions have been involved in the development of MT in China ever since its start. The collaboration, which is common in this field, among people from computer science, mathematics, linguistics have spurred on the development of MT greatly in China. It was also in this period, not necessarily under the impact of the 1966 ALPAC Report, that people in China realized that machine-aided translation is more feasible.

In the mid-1970s, MT research regained its original momentum and resumed its rapid growth, with the collective efforts of many ministries and institutions, the Institute of Linguistics of the Chinese Academy of Social Sciences being the lead. A five-year long collaboration yielded some rudimentary systems and helped to train many researchers, who would continue their work in places all over China. In the meantime, researchers were sent abroad or recruited to do postgraduate study in this area. National conferences or seminars on MT were held regularly and related journals were published.

The 1980s and 1990s witnessed the second important phase in the development of MT in China. During this period, two milestone practical systems came into being. One is the KY-1 English-Chinese Machine Translation System developed by the Academy of Military Sciences in 1987, which won the second prize of the National Scientific and Technical Progress Award and was later further refined into TRANSTAR, the first commercialized MT system in China. The other is

the "863-IMT" English-Chinese Machine Translation System, which won the first prize of the National Scientific and Technical Progres Award and has brought about tremendous profits. These two systems are the children of collaborative efforts of various institutions and people. Another system worth mentioning here is the "MT-IR-EC" developed by the Academy of Posts and Telecommunications, which is very practical in translating INSPEC titles from English into Chinese. Not mentioned here are many other efforts made in this period, including the joint program between China and Japan, which introduced MT in the Chinese context to the outside world and helped the fostering of talents and transmission of technology and the accumulation of resources. Consequently, some Japanese-Chinese MT systems came into being, such as those developed by Tsinghua University, Nanking University, and the China University of Science and Technology. In the mid 1990s, for the first time the world over, a research group led by Professor Yu Shiwen at the Institute of Computational Linguistics of Peking University, constructed a quite reliable evaluation system of MT (Yu, 1993).

From the 1990s onwards, MT in China has undergone a rapid growth. Many commercial systems are available on the market. All these systems share some common features. For example, most of them are equipped with very big multi-disciplinary and domain specific dictionaries, operatable through the network, and user-friendly. New technologies, such as human-machine interface, began to be developed. So in a sense, MT in China is not far behind in its PC products and network system development. The dominant technology strategy and guideline of MT in the Chinese context then were not very different from those adopted in other parts of the world. They are mainly transformation-based, rule-based and very practical, many of which are still in use today.

In recent years, substantial efforts have been made in developing MT in China. In 1999, Yaxin CAT 1.0 was publicized. It is China's first all-in-one computer-aided translation system, which combines translation memory, human-machine interaction, and analysis. Now Yaxin CAT 3.5 is commercially available and has brought about substantial profits to its developers and users. Since early 2004, its developer and publisher, SUNV, has become listed in Hong Kong Stock Market.

There are now several academic organizations active in MT in the Chinese Mainland. For example, the Chinese Information Processing Society (CIPSC) has been organizing several international and national conferences since it was founded in 1981 and is an active participant of international exchanges in recent years. However, there still leaves much to be done in the fact that the exchanges are mainly done in Chinese, while little effort has been made to have the conversation conducted in English in order to be recognized by a larger audience beyong the Chinese context.

At Beijing Foreign Studies University, a research project on the design and construction of a bilingual parallel corpus has been going on for several years, with one of its goals being to shed some light on the bi-directional translation between Chinese and English (Wang 2004).

2.2. Taiwan and Hong Kong

MT in Taiwan got started in 1985, when Tsinghua University at Taiwan and Yingqun Computer Company decided to cooperate on the development of English-Chinese MT System. Though it is a late comer, Taiwan has managed to catch up with the rest of the world in the wave of MT research.

Since then, Taiwan researchers have been very arduous in learning and exchanging new ideas with the outside world (Su 1997). Many institutions have decided to put in efforts for the translation between Chinese and English or Japanese. Now, they are a member of the Asia-Pacific Association for Machine Translation and the International Assocication for Machine Translation. In general, the research and development of MT in Taiwan is quite full-fledged, with TransWhiz being one of its representative products that are commercially available.

In Hong Kong, early efforts were made at The Chinese University of Hong Kong in the later 1960s (Loh 1972) and there is strong indication that with the rapid advances made in the areas of computer science and computational linguistics, and aided by the government's promotion of technology, interest in machine translation will get stronger and stronger (Chan 2001).

It is also worthy of note that, in this region, much attention has been paid to the teaching and study of MT over the past years. The world's first MA program in Computer-aided Translation was offered by the Department of Translation, The Chinese University of Hong Kong in 2002. The enthusiasm demonstrated by students admitted into this program in the past two years is symptomatic of the growing demand of the society at large. It is believed that such a program will attract more and more talents to join their hands in this field.

In December 2004, the first International Conference on Translation Software: The State of the Art was held in the Chinese University of Hong Kong. At the Conference, representatives from leading producers of CAT software demonstrated their latest achievements in this field, such as Tradod, SDLX, Yaxin CAT, Huajian, etc., and the opinions of users of such software were exchanged. The dialogue among producers and users of CAT software will surely prove meaningful to its further development.

2.3. Commercial Software on the Market

Now that commercial software of various kinds are available on the market, MT is a familiar thing to the public. On the Chinese market, there are various CAT software devoted to many language pairs, for example, between English and Chinese, Japanese and Chinese, German and Chinese, French and Chinese, in free-lance version or coporate version, with or without multidisciplinary dictionaries. It is estimated that there are at least 1,500 kinds of CAT software, with more than 20 developed in the Chinese context and of relatively lower price when compared with western ones. The good news for the users is that some of them have online versions providing free service and users have many choices to suit for their specific needs. The following is a brief introduction of some of the most popular and typical ones.

Yaxin CAT series (http://www.sunv.com) were first publicized in 1999 and have undergone such a rapid growth that it is equipped with Translation Memory technology, 74 speiclaized dictionaries, and many modules. It also occupies more than 50% of the domestic market. The developer, Beijing Sunway Software Co. Ltd., claims that, Yaxin CAT Version 3.5 is a professional translation platform of a new generation, the biggest advances include: high quality, high efficiency, convenience and practicality, auto memory, intelligent analysis and human machine interaction. It takes advantages of complementary roles played by the computer and people, lets the computer

assist men to finish the work fast, can help the enterprise and individual fully utilize resources, lower the costs greatly, double improve working efficiency. Translation solution in electronic form offered by Yaxin CAT 3.5 is suitable for enterprises and institutions, scientific research institutions, news publishers, translation companies, localization companies, full-time or part-time translator, science & technology translators in particular.

Huajian CAT software (http://www.altlan.com) is produced by Huajian Electronics Co. Ltd. in China for the translation between Chinese and English and between other major Asian and European languages. Features of Huajian products include: (1) webpage full-text translation; (2) document full-text translation; (3) selected-text translation; (4) mouse-tracing translation; (5) menu translation; (6) e-mail translation; and (7) expandable dictionary.

Kingsoft AI Translation tools (http://kingsoft.com.cn) are produced by Kingsoft in China and can translate between Chinese and Japanese, Japanese and English, and Chinese and English. Versions include Kingsoft Express 2000, 2001, 2002, 2003, and 2004. Special features of this software include: (1) instant translation; (2) editable dictionary; (3) web translation; and (4) permanent software localization.

It is clear that the tools included in such software are of multi-purpose with their on-line dictionary, computer-screen translation, text translation, etc. However, one must be aware of such promotional claims, because it is common knowledge that high quality translation can only be achieved through human intervention.

3. Bottlenecks in MT and MAT Research in the Chinese Context

Such rapid growth and remarkable achievements, however, don't mean that the technologies involved are quite mature.

The history of MT research and development indicates that MT and MAT require the collective efforts of people from various fields. In the past, induction was done manually and was time-consuming and very costly. It is also problematic because consistency is very difficult to arrive at. Once some new rules are added to improve the translation of certain sentences, it would be very difficult to handle other sentences which didn't present any problems before the addition. New errors would appear when new formations are made, which has led to the growing complexity of the system and the growing difficulty in maintenancing. This has been a universal bottleneck for MT system in the past several decades.

For Chinese, another problem is word segmentation, which is the first yet a key step in Chinese information processing. So far, there has not been a perfect solution though many advances have been made (Sun 2001; Yu 2002). On the one hand, research conducted at Peking University demonstrates that there is no need for an absolute definition of word boundary for all segmenters, and that different results of segmentation shall be acceptable if they can help to reach a correct syntactic analysis in the end (Duan et al. 2003). On the other hand, the testable online tool it has developed cannot yet segment words with ambiguous meanings in most cases (http://www.icl.pku.edu.cn/icl%5Fres/segtag98/).

4. Prospects of MT and MAT in the Chinese Context

With the rapid growth of Internet Technology, the future of MT and MAT research and development is quite promising and more profits could be made. But as it is pointed out in the previous section, the quality of MT translations has not been substantially improved. One thing that is clear is that MT is not only a problem of language processing, but also one of knowledge processing. Without the accumulation of knowledge and experience over the years, it is hardly possible to develop an MT system which is practical. The short cycle of development at present is the result of many years' hard work and the accessibility of shared resources.

Looking forward, there is still a long way to go before MT can truly meet the demands of the users. Generally speaking, things to be done for both MT and MAT research and development between Chinese and other languages should include the following:

(1) Though the notion of a "text" has been lost because the translation tools now available operate primarily at sentence level (Bowker 2002: 127), the analysis of the source language (Chinese in most cases) should be done in the context beyond the present sentential level which is isolated, and based on the comprehension of the original. Future analysis should take the sentence cluster or even the entire text into consideration. While analysis today seeks to find out the syntactic relationship tree, semantic relationship of the concepts involved at most, future analysis should be on the textual meaning instead. Once this is arrived at, meaning transfer could be done more accurately than the present systems do (Dong 2000).

(2) Basic research needs to be deepened and strengthened, especially the construction of common-sense database. Some scholars even suggested that a knowledge dictionary should be built up to facilitate comprehension-based analysis, such as the Zhi Wang (Knowledge Network) developed by Dong Zhendong, a leading Chinese scholars in MT, and his colleagues (Dong 1999; http://www.how-net.com), which has shed some light on the comprehension-based analysis and explorations of disambiguation.

(3) The stress of research and development should be more and more on the parameterized model and a corpus-based, statistically-oriented and knowledge-based linguistic approach. Efforts should be made to develop a method for semantic disambiguation and an objective evaluation of it. Automatic learning (acquisition, training) strategies of the computer and a bi-directional system design should be strengthened. A more user-friendly feedback control function should be developed so that the user can adjust the behavior of the system.

(4) As it is pointed out by Hutchins (1999) and applicable to MT and MAT in the Chinese context, translation software now available are still expensive. How to develop an efficient system that is of low cost, high reliability and required less work on constructing the translation memory for individual translators is another emerging problem. Besides, translation systems into minor languages and spoken language should also be further explored.

(5) It is necessary for scholars in the Chinese context to learn from and exchange with others. The close collaboration, led by Professor Yu Shiwen, Institute of Computational Linguistics, Peking University between Peking University and Fujitsu has been fruitful. They have managed, to a great extent, to produce a tagged corpus of 13,000,000 Chinese characters in order to find out some

statistical rules and parameters (available at http://www.icl.pku.edu.cn/WebData_http-dir-listable/ICLseminars/2002spring/).

(6) Attention should be paid to "spoken language translation", which still elludes us and could be a very ambitious project (Somers, 2003: 7).

(7) Attention should also be paid to network teamwork, from stand-alone systems, so that multiple users can share the same resources.

5. Conclusion

According to J. Hutchins (1999), the unofficial historian and librarian of MT, the quality of translation produced by MT has not been remarkably improved in the past 50 years.

With China's entry into the WTO, domestic CAT developers will come across new challenges. How to take this opportunity and face the challenges is a problem that needs to be taken seriously.

Even though MT is not mature theoretically and technologically at present, people can still make do with the poor quality of MT and researchers and developers are contented with making big money with their systems and are not willing to invest more funds and technology as if the unsatisfactory quality of MT had nothing to do with them. They may not have realized that the popularity and profit MT has got so far is based on the rapid growth of the network, when people are faced with linguistic problems and are not very particular about the poor quality of translation done by MT, so long as they can grasp the main idea of the message. So it would be very desirable that MT could make good use of this opportunity and improve itself.

It is not possible to cover every aspect of the problems and possible areas of future MT and MAT research and development in the Chinese context in a single paper. However, this paper has managed to give a somewhat overall picture of the achievements, bottlenecks, prospects and future steps to be taken of MT and MAT research and development. Though it is mainly from a user's perspective, it is hoped that this paper could help in exchanging opinions with the outside world on MT and MAT in the Chinese context.

REFERENCES

BOWKER, L. (2002): *Computer-aided Translation Technology: A Practical Introduction,* Ottawa, University of Ottawa Press.

CHAN, S.-W (2001): "Machine Translation in Hong Kong", in CHAN S.-W (ed.): *Translation in Hong Kong: Past, Present and Future,* Hong Kong, The Chinese University of Hong Kong, p. 205-18.

DONG, Z. (2000): Review of MT in China in the 20th Century, Available at http://tech.sina.com.cn.

DONG, Z. (1995): "MT Research in China", in MAXWELL, D., SCHUBERT, K. AND WITKAM, T. (eds.): *New Directions in Machine Translation*, Dordrecht-Holland, Foris Publications, p. 85-91.

DUAN, H., XIAOJING, B., BAOBAO, C., SHIWEN, Y. (2003): *Chinese Word Segmentation at Peking University*, Available at http://acl.upenn.edu/w/w03/w03-1722.pdf.

FU, A. (1999): "The Research and Development of Machine Translation in China", Paper Presented at the *Machine*

Translation Summit VII, September 13-17, 1999, Singapore.

HUTCHINS, J. (1999): "The Development and Use of Machine Translation Systems and Computer-aided Translation Tools", Paper Presented at the *International Symposium on Machine Translation and Computer Language Information Processing, 26-28, June, 1999, Beijing*.

LOH, S.-C. (1972): "Machine Translation at The Chinese University of Hong Kong.", in MATTHIAS, J.(ed.): *Proceedings of the CETA Workshop on Chinese Language and Chinese Research Materials, 24-25 March*, Washington, D. C.

SOMERS, H. (2003): "Introduction", in SOMERS H. (ed.): *Computers and Translation: A Translator's Guide*, Amsterdam/Philadelphia, John Benjamins Publishing Company, p. 1-12.

SU, K. (1997): *Development and Research of MT in Taiwan*, Available at http://www.bdc.com.tw/doc/twmtdvp.gb.

SUN, M. (2001): "New Advances in the Study of Automatic Segmentation of Chinese Language" in *Proceedings of Conference of the 20th Anniversary of CIPSC*, Beijing, Tsinghua University Press, p. 20-40.

WANG, K. (2004): "The Design and Construction of Bilingual Parallel Corpus", in *Chinese Translators' Journal*, 6, p. 73-75.

YU, S. (1993): "Automatic Evaluation of Output Quality for Machine Translation Systems", in *Machine Translation*, 8, p. 117-126.

YU, S. (ed.) (2002): *CJNLP 2002, The Second China-Japan Natural Langauge Processing Joint Research Promotion Conference Proceedings*, Beijing, Institute of Computational Linguistics, Peking University.