Advances in Social Science, Education and Humanities Research, volume 233
3rd International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2018)

Problems and Countermeasures in Machine Translation*

Jingyuan Xie
College of Humanities
Tianjin Agricultural University
Tianjin, China

Chunyan Ma
Graduate Admissions Office
Tianjin Agricultural University
Tianjin, China

***Abstract*—Computer translation is an interactive translation of natural languages by means of computer devices. Both cultural and linguistic differences have proved some problems existed for the traditional translation. Therefore, the automatic English to Chinese translation system by means of corpus may be an effective solution.**

Keywords—computer translation; manual translation; corpus; electronic machine translation; implementation method; postgraduate

I. INTRODUCTION

The conception “translation machine” was first proposed by Russian inventor Prtr Smirnov-Troyanskii and Armenian-French man Georges Artsdrouni in 1933. It's been over 80 years. With the advances in computing science, statistical and machine learning methods are widely used in the field of natural language processing. Recent years, as a mew discipline and technology, machine translation is a typical interdisciplinary involving linguistics, computer science, it is a research in the fields of artificial intelligence; in mathematics, it is a research in the field of mathematical logic and algorithm. So what is machine translation now? Machine translation, also known as automatic translation, is the process of transforming one natural language into another by computer, and the future of machine translation is very promising, although there are still many problems and difficulties.

II. CONNOTATION OF MACHINE TRANSLATION

Machine translation (MT for short) refers to the mutual translation between natural languages by computer and the transfer the text content from one natural language to another by software.

First of all, machine translation is the product of structuralism, which is essentially a word-to-word translation method. In the 1960s, Chomsky's generative translation grammar and the success of computer linguistics made computer science develop rapidly. People analyzed the translated articles from the whole point of view, and the researchers of machine translation gained encouragement and confidence from it. First of all, machine translation is the product of structuralism, which is essentially a word-to-word translation method. In the 1960s, Chomsky's generative translation grammar and the success of computer linguistics made computer science develop rapidly, the researchers of machine translation gained encouragement and confidence from it.

However, without the participation of "human", the machine translation is untenable only by text analysis. Because translation is a human-centered cross-language communication activity, the author of the text is inevitably influenced by political and cultural factors in his living environment. The source language culture is infiltrated into the text, and the translator also infiltrates the target language culture into the translation process. Translation is not only a kind of cultural construction for a certain purpose, but also a kind of complex psychological activity, a kind of thinking processing activity.

The translator needs to use their own inherent knowledge and experience to analysis and processes, and to obtain the meaning from the stylistic content, and then to choose the target language expression in accordance with the language habits of. But the machine can't understand the original text,

What does it can do is mechanical transcoding, it can't break through the barriers of thinking to translate the articles compare to human. The authentic complete translation must be modified and processed manually.

In addition, machine translation can be misleading, so machine translation can only be machine-aided translation, which also conforms to the computer as a human auxiliary tool.

The development of machine translation depends on the development of linguistics, that is, how to translate the vocabulary and grammatical structure of the source language into the target language. Different scientists in different countries translate according to different criteria, even if the scientific literature cannot be translated the same. So the translator must be proficient both in two languages and the two cultures and their differences, which I'm afraid the machine will never reach. Human memory has its own limitations, dictionary reference books or computers are necessary. In addition, machine translation can be misleading,

*Project No.: 1) Tianjin Agricultural University, Postgraduate Key Curriculum Construction Project, No. 2017 KYC0032.
2) Tianjin Agricultural University, the Science Development Fund Project, No. 2016 SYB04.

so machine translation can only be machine-aided translation, which also conforms to the computer as a human auxiliary tool.

The translator always inadvertently use the knowledge, experience accumulated inherent in the brain. After updated constantly, the translation work is getting better. And because the machine could not handle the new emerging or special expression, artificial modification and processing should be necessary.

III. DIFFICULTIES IN THE MACHINE TRANSLATION

Nowadays, machine translation is undoubtedly a recognized technology direction with wide application prospect. In order to achieve the successful translation between Chinese and English, we need to collect a large number of pairs of Chinese and English sentences, and then use the computer to count and learn the translation knowledge from these pairs of sentences. You may feel that machine translation is not difficult, is not to collect enough words and examples?

Actually, it's not easy for machines to learn translation knowledge. Human language has great complexity. First, many words and expressions are ambiguous, vague, and relevant to a particular application environment. Even the same sentence, it has different meanings in different contexts. Secondly, the word order of different languages is also different. Furthermore, it may have many correct translation methods for the same sentence. This increases the uncertainty of the machine learning process. Therefore, an excellent machine translation system should master the knowledge of word translation, phrase translation, grammatical structure translation, and semantic translation and so on.

Taking the direction of Chinese - English translation as an example, the system should first master the translation knowledge of words, phrases and grammatical structures between Chinese and English. With this knowledge of translation, the system divides the Chinese sentence into words, phrases, or combinations of grammatical structures (in the process, there are thousands of possible segmentation, each unit also has a variety of translation options), then translate each unit separately, and finally combined to form the final English translation. Ambiguity is a common phenomenon in natural language, which refers to the situation where the same form of language has different meaning.

A. Ambiguity and Meaning Understanding

Ambiguity is a common phenomenon in natural language, which refers to the situation where the same form of language has different meaning. Natural language is full of ambiguities at all levels including vocabulary, syntax, semantics, and pragmatics. Not only should the ambiguity within the same language be considered by Mt researchers, but also that of transfer between two languages. And structure is a common structure ambiguity during language transfer, for example, the more expensive clothes, which can be understood in Chinese as "更贵的衣服" or "更多的贵的衣服". It seems that the ambiguity in the structure is easy to handle. But computer cannot understand it. Someone joked that whoever can solve this problem should be awarded a Nobel Prize. It can be seen that language disambiguation is one of the straitened difficulties faced by MT.

The ambiguity remains to solve due to computer's inability in understanding language. It has been a great vision of MT researchers or even humans to make computer understand nature language. As a complicated process, human translating integrates several thinking activities including understanding, analyzing, selecting, and re-creating. Artificial translator can produce a translation combing their bilingual knowledge, culture, history, and other background knowledge. After considering all aspects of the original text, the translator can also delete or add information, and do some carving and polishing work, which will make the translation better. In short, humans do translation work with the participation of mind. However, the working principle of machine is based on the rules researchers put into the system, and serial binary statistics or non-linear database, which do not have the ability of thinking, judging, or reasoning, so MT system will inevitably cause ambiguous translation versions.

B. Defects of Machine Traslation Methods

Machine translation theory is still being formed, and there is not a perfect approach or mature theory to guide the research of machine translation. At present, rule-based machine translation methods including direct approach, transfer approach and interlingua approach involve analysis on level of word, syntax, significance and texture and generating of target language. However, due ot the lack of in-depth understanding and analysis of semantic and structure as well as simplistic rules, rationalism methods cannot produce satisfactory translation. Empirical machine translation methods, including SMT and EBMT are faced with problems increasing linearly with the expanding scale of training data. Although neural machine translation with its non-linear data processing methods is not bothered by training data, it has other problems to solve, such as data sparseness and data noise, which are also the difficulties faced with empirical methods.

C. Deficiency of MT Research in China

Although China takes the lead in developing practical NMT system, the theory of NMT was put forward by researchers from Canada and there is still a gap between China and leading countries in the field because there is no systematical or mature theory guiding Chinese machine translation; in terms of staff, China lacks stable group of inter disciplinary talents engaged in machine translation research, which requires not only programmers and linguists, but also scholars in cognitive psychology and machine learning.

Chinese linguists, translators, or computer experts engaged in researching English-Chinese translation were independent in their own field and do not communicate well with experts of leading countries in the field, so they cannot

do adequate work in revealing rules of languages and transfer rules between two languages, especially grammar rules of English. In return, efficient communication will also benefit in understanding grammar rules of Chinese for researchers.

IV. THE COUNTERMEASURES IN MACHINE TRANSLATION

The traditional machine translation systems are generally based on the rules of language including the original words of lexical, grammatical and semantic analysis. It converts the language to generate the target language finally. In this process, a large number of language rules, including lexical, syntactic and semantic rules, and even pragmatic rules are used.

A. Some MT Methods and Technologies

There are some advantages and disadvantages in rationalism methods, empirical methods, and neural methods in processing nature languages. Machine translation system performance varies, so does the accuracy of translation. The fusion of rationalism methods and empirical methods, which is multi-engine approach, has appeared since 1990s and it performs better than single approach. Combining traditional and new technologies approaches is one trend.

For rationalism methods , sometimes complicated grammar rules are difficult to analyze and depict, so in this circumstance, statistical models which provide much informal and complex language phenomenon can play a supporting role in handling the difficult grammars.

Machine translation system of empirical approach can integrate rationalism methods to get a more comprehensive processing of language. Rationalism approach can provide rules of syntactic structure and some semantic information, so it will reduce the word order errors and ambiguities in the translations by statistical system, thus improving the translation quality of empirical MT system.

Another fusion is the field of technology, GNMT (Google neural machine translation) is actually a combination of some advanced technologies proposed in academic fields. Based on its own data, computing power, and high-level engineers, Google took the lead to integrate those technologies into its system, which turns out to be much better than traditional methods SMT. Therefore a MT system with strong vitality will fuse more than one technologies.

B. The Corpus Methods

The corpus method, a new theory and method of machine translation, has been taken seriously by many people. In corpus method, analogy is an imitation feasible translation method. It is to set up a database, to store a large number of examples and the corresponding translation structure, it is called the corpus. This translation system can search and input similar sentence structure in the corpus to converse from the source language into the target language. Because of this method is not too deep in syntactic semantic analysis, this method is effective as long as the corpus and search technology and other auxiliary skills do a good job in the class.

The design of the system is: to establish a corpus and the electronic dictionary. The corpus is the core of the system, the translation process is carried out around the corpus to, mainly in the corpus structure and the corresponding translation of sentence structure. It also holds some necessary sentence feature information, such as the required part of speech, parts of speech and other grammatical and syntactic characteristics. The selection of corpus is to consider its completeness, universality and pertinence, and its data structure can be made up of one or more databases with a general relational database. The design of machine dictionary is very important because there are many kinds of words and polysemy in English. General translation software consists of the input of an electronic dictionary, which is a core part of the translation engine.

We can design a corpus to restrict the dictionary - to collect the fixed collocation of words and words, to help the lexical analysis module to disambiguate. The choice of interpretation is often dependent on the word that goes with it. For a polysemous word, the arrangement should be adapted to the field.

Electronic dictionary is different from the corpus, it is a lexical information database, its design is one of the key of machine translation research. It can even generate concept dictionary, to speed up the system operation, to improve the interpretation accuracy because it is machine-oriented, content to have better completeness, applicability and scalability. Corpus and machine dictionary can use the common relational database, expansion; modification and maintenance are very convenient. In addition, it is the key to improve the quality of machine translation by establishing a corpus of words collocation, improving the database of semantic structure and adding background knowledge.

V. THE FUTURE OF MACHINE TRANSLATION

A. The Commercialization for MT systems

Google released Google neural machine translation in September 2016, which is a landmark from statistical translation to neural network translation. Before it, except Baidu, there was no such a large-scale MT system able to apply DNN in the world. Google translation performs better in aspects of updating technology and marketing than Baidu, causing a series of commercial effects.

Translation technology is to serve the public; otherwise it is moon in the mirror and flowers in water. Users are more concerned about the practicality of Machine Translation systems. It is important that the MT system gradually serves better in daily life and meet various fragmented translation needs. Baidu translation also improves its practicality of APP by combing OCR (Optical Character Recognition) technology which can help system to recognize characters in picture, and speech recognizing technology which will facilitate speech and even simultaneous interpreting. Baidu provide users with cameras translation and speech translation. So when attempting to create commercial systems from lab

models, MT system should firstly solve the difficulties in camera translation and speech translation. Besides, it will be greatly helpful for future Mt to develop in aspects of speed, friendly interface, and multi-language services.

B. Compromises for Fully Automatic MT

High quality fully-automatic machine translation cannot be accomplished overnight. The complex feature of nature language system determines that the research on MT system is complex, difficult and long-term. Therefore, a rational and pragmatic attitude towards Machine translation should be held. Post editing is an alternative choice. Sub-language and control language are also being research currently in order to limit the input of MT and make MT systems easier to process into language. In addition, before fully-automatic machine translation systems are available.

C. Collaboration

Firstly, during the past twenty years, affected by the overall trend of artificial intelligence, especially the fruits of machine learning, machine translation have developed in a data-driven approach, automatically learning from a large number of Internet text. Machine Translation is a crossed discipline, depending on progress in many areas; mathematics, linguistics, computer science, cognitive science and engineering.

Secondly, nowadays language experts engaged in the study of English and Chinese research are not working in collaboration with each other; they cannot effectively carry out large-scale cooperation in order to reveal the English and Chinese transformation law. Few people involved in the national authoritative Journal of college or Chinese Journal of research in the field of the application. Machine translation research is an interdisciplinary field; its successful development requires the cooperation of linguist and computers.

VI. CONCLUSION

The article aims to introduce and expound the situation and the future of machine translation. It focuses on the development of machine translation in twenty first century, especially the creation of new approach, neural machine translation.

Entering into 21st century, fierce completion between various MT systems has led to a considerate development of MT no matter in aspects of application or approach. Some rational approach shows strong describing and generating ability in processing nature language and it can solve the transfer difficulties form deep representation to surface representation. Some empirical approach, especially, statistical-based method, using mathematical model, has capacity of acquiring knowledge and predicting trend.

REFERENCES

[1] Boilet C. (1995) Factors for Success and Failure in MT. Some Lessons of the First 50 Years of R&D. Proc. of MT summit V, Luxembourge.

[2] Google. 92016, December， 26） Translation Text. Retrived from Neural Machine Translation.

[3] Krashen, S. (1981). Second Language Acquisition and Second Language Learning. Oxford Pergamon.

[4] Swain, M. and Bowers, W. (1991). (eds.) Applied Linguistics and English Teaching. Macmillan Publishers Limited.

[5] Peter F. B.et al (1990) A Statistical Approach to Machine Translation Computational Linguistics(2).

[6] Spmers, H L. (2003) Computers and Translation: A Translator's Guide. Amsterdam/Philadelphia" John Banjiamins Publishing Company.