Corpus-based Machine Translation: Its Current Development and Perspectives

Zhou Dajun and Wang Yun
Department of Basic Courses, Naval Aeronautical and Astronautical University, Yantai, China
Email: e_zdj@hotmail.com

[Abstract] Corpus technology was introduced to rule-based machine translation (MT) in the late 1980s. Corpus-based MT mainly includes statistic-based MT and example-based MT – the former lays emphasis on statistic model from mathematics, the latter inference through example translation from machine learning. The semantic-based method will become the trend in statistical MT development, while the perspectives for corpus-based MT system is to combine the latest research fruits of theories and technologies of various subjects concerned and to develop multi-mode corpus.

[Keywords] machine translation (MT); corpus; statistic-based MT; example-based MT; perspective

A Review of Corpus-based Machine Translation

Corpus and Machine Translation

Corpus is a large-scale database with tremendous collective linguistic information in real use, which is provided for retrieval by computers for research. The first corpus was established in Brown University, U.S.A., in the late 1960s (Zhang & Zhang, 2010, p. 55). Much progress has been made in corpus research and application in the past decades. Current studies of parallel or multi-language corpus can be categorized into three aspects: the first is the alignment technology of parallel language material, with various strategies and approaches provided by scholars and with numbers of programs and tools of alignment parallel or multi-language material; the second is the application of parallel language materials, such as statistic-based machine translation, example-based machine translation, and parallel language dictionary compiling; the third refers to issues of parallel language corpus design and management, and its material collection and coding (Chang, et al., 2003, p. 28).

Corpus used in translation is one of the focuses of corpus application research. Machine translation (MT) is a technology to translate from one natural language in character or speech form into another by means of computer programs (Zhao & Liu, 2010, p. 36). MT was initiated in the 1950s and entered into a prosperous period in the end of 1980s, which characterizes practicality of many translation systems in various fields. An English-French translation system TAUMMETEO developed by the University of Montreal, Canada, in 1976, is a typical example, which can provide high-quality translation of weather forecast (Shao, 2010, p. 28). A typical MT system adopts a transfer-based translation strategy, which consists of 3 procedures: 1) analyzing a source language and form representation of the language; 2) transforming representation of the source language into that of the target language; 3) generating the source language translation version from the target language representation (He, 2007, p. 191). Traditional MT has two defects: one is that traditional MT regards words as its basic translation unit; that is, the machine first segments sentences of a source language into words, which are transformed into those of target language, and then those words are connected according to grammatical structure rules of the target language; the other is that traditional MT does not pay much attention to contexts. Peer-to-peer studies of corpus-based MT do attempt to get over the defects of the traditional MT systems and improve efficiencies and accuracy of the MT systems.

With the development over more than 50 years, MT systems have performed certain functions in some fields. However, the current systems have not reached the effect of translation as expected. In its earlier period, MT research was conducted from the viewpoint of natural linguistics, thus creating MT systems based on such linguistic rules as lexical rule, syntactic analysis rule, transformation rules, and target language generation rules. As these rules were summarized and developed from experiences of

linguists, there exist some deficiencies in the analytic rules. For instance, manual writing of those rules demands a large quantity of workload, and the rules are too subjective to keep consistency (Wang, 2003, p. 33). Since 1989, MT has entered a new stage in which corpus methods are introduced to the rule-based technology, including statistic-based and example-based methods, and the method of turning corpus into linguistic knowledge bank through language material processing, etc. (Feng, 2010, p. 28). The past years have seen the prominent achievement in MT systems.

Corpus Used in Machine Translation

There are three kinds of corpus concerning MT: parallel corpus, multi-language corpus, and comparable corpus. Parallel corpus collects original text of a language and its translating text of another language. Multi-language corpus, designed according to the similar standard, is a compound one composed of multi-language material texts, which are original versions without their translating ones. Comparable corpus collects a certain language, such as an original English text, as well as an English translated version of other language texts (Xiao, 2007, p. 25).

Since the middle of 1990s, a group of translation researchers have applied corpus to descriptive translation research for the purpose of discovering the essence of translation texts as communicative media (Baker, 1995, p. 243). In 1995, the Translation Research Center of College of Science, Manchester University, UK, under the leadership of Professor Backer, established the first comparable corpus in the world – Translational English Corpus (TEC). Single and multi-languages corpuses have played a prominent role in the research and development of language information processing in the past few years. Especially in the MT research, many new approaches based on multi-language corpus have been proposed. For instance, example-based or store-based MT methods can be adopted to improve MT quality with the aligned bilingual material. In addition, a bilingual dictionary and translation pattern can be acquired from a bilingual corpus through a statistic model to improve the traditional MT method (Chang, et al., 2003, p. 28). Many larger parallel corpuses have been established, such as the English-Norwegian Parallel Corpus of Oslo University and the Conference English-French Parallel Corpus of Canadian Parliament, to provide powerful tools for translation research so that translation natures and various restrictions in translation processes can be objectively and scientifically observed and studied.

In China, there are such corpuses as General Chinese-English Parallel Corpus established by the China Foreign Language Education and Research Center, Beijing Foreign Studies University, and the

Chinese-English Parallel Corpus co-developed by Computational Linguistics Research Institute of Beijing University, Computing Technology Institute of Chinese Academy of Sciences and Intellectual Technology Laboratory of Tsinghua University (Xiao, 2007, p. 25). Nowadays, corpus-based multi-language and multi-mode translation models of various scales and types are set up; Google's is one of the most well-known. Corpuses with only thousands of sentence pairs can be applied to a certain special translation field, while a majority of corpuses usually have millions of sentence pairs. There is no doubt that corpuses on broader scale provide higher possibility of optimized translation. The introduction of the internet search enables the corpus size to be indefinitely enlarged.

Corpus-Based Machine Translation Methods

According to ways of knowledge acquisition, MT can be divided into rule-based methods and corpus-based methods, and the latter can also be divided into example-based methods and statistic-based methods (Zhao & Liu, 2010, p. 36). In these methods, aligned bilingual material can be directly used to improve quality of automatic MT and promote the human-machine interaction in machine-auxiliary translation; furthermore, a translating model can be acquired from the bilingual corpus through a statistic model to improve the traditional translating methods, which are quite time-consuming and error-prone, by acquiring translating models from bilingual corpus. It is predicted that corpus-based MT systems may greatly excel in performance of the third generation of MT systems and become the embryo of the fourth one (Li, 2004, p. 59).

Example-based MT and statistic-based methods have greatly promoted MT development. Both of them are based on larger-scale corpus; therefore, the corpus becomes a focus of MT research. Both

example-based method and statistic-based one are data-driven, but they do not repel each other, as they solve problems from different perspectives. Example-based methods conduct reasoning based on translating examples from the perspective of machine learning. Statistic-based methods, however, lay emphasis on mathematical static models. The objective of the researchers is to combine different methods with their complementary advantages and establish an MT system with a combination of multiple methods (Shao, 2010, p. 35).

Promoted by the two methods in the recent years, many new theories and approaches have appeared. As a result, translation quality has been improved, and, furthermore, as translating knowledge can be automatically acquired from larger scale corpus without writing rules with manual power, the development period of the MT systems has been much shortened and MT application expanded. More researchers have been involved in the MT research because of lower threshold of the research (Liu, 2009, p. 149). Since 1999, a breakthrough has been made in the statistic-based MT method, which is now in a rapid development. Actually, it is a period at present with many MT methods blended and paralleled (Shao, 2010, p. 26).

Main Methods of Corpus-based Machine Translation

Example-Based Method

The idea of example-based method was proposed by Nagao Makoto, the Japanese MT expert, in 1981 and published in 1984. He points out that human do not translate through deep linguistic analysis, and their translation process is that, first, input sentences are correctly segmented into pieces of phrasal fragments; then, these fragments are translated into those of another language, and, finally, complete sentences are constructed with the fragments, each of which is translated by analogy (Feng, 2010, P. 33). Therefore, examples should be stored in the computer and a mechanism established in which similar example sentences of a given sentence can be retrieved. Example-based Machine Translation (EBMT) is a method to search from the bilingual example bank the most similar translation examples of the source language sentences to be translated, and then to complete the translation by regulating those examples (Zhao & Liu, 2010, p. 37).

The operating principle of EBMT system is that, first, the main knowledge source is the bilingual translation example bank in which there are two fields: one stores source language sentences, the other corresponding translation. When a source language sentence is input, the system compares it with the field of source language sentence in the example bank, finds out the most similar sentence, simulates its corresponding translating sentence, and gives output of the final translation. Next, translating knowledge represented by examples and semantic class dictionary are easily added or deleted. The attractive translating strategy is to make a precise comparison by means of larger scale translation example bank to produce higher quality translation and avoid the difficulties of deep linguistic analysis in the traditional rule-based MT method. Three aspects need to be studied in example-based MT; the first is correct bilingual automatic alignment, the second is an effective example matching retrieval and the third is to generate corresponding translation of the source language according to retrieved examples.

Example generalization plays an important role in translation (Zhao & Liu, 2010, p. 37). As the exactly same examples may not be found for the source language to be translated, EBMT has to find out the most similar examples by means of semantic dictionary. Once the similar sentences are chosen, then follows the translation regulation by means of bilingual dictionary; so, EBMT is used in full-automatic translation. The functions of the initial EBMT is later extended in many aspects, and the most typical one is the creation of example pattern through the example generalization, which means that some specific words are generalized to some categories. Example generalization has considerably increased the example matching rate and reduced the scale of example bank needed in translation. Theoretically, examples can be abstracted to rules and the rule-based method regarded as a highly abstract result of translation example. However, generalization is not an easy job due to the ambiguity of natural language. In many cases, examples cannot be found to cover the source language to be translated, which is the limitation of

example-based method; therefore, this method is only used as a complement to other mainstream methods of translation systems.

Example-based MT method does not conduct a deep semantic analysis, thus avoiding, to some extent, the difficult process of language analysis. The system based on this method can be extended by increasing examples and words, so it is easily maintained. In addition, it can produce translation of higher quality, as it makes use of a large quantity of translation examples. Nevertheless, there are still many key problems with the method that need to be solved, such as how to construct bilingual alignment corpus, how to computerize the similarity between fragments to be translated, and translation examples in order to find out the most suitable fragments when the matching fragments are retrieved, how to effectively combine the example fragments to form translation texts, and how to increase the coverage of translation examples (Shao, 2010, p. 27).

Some major example-based MT systems nowadays are: MBT1 and MBT2 systems in Tokyo University, the PANGLOSS system of Multi-Engine Machine Translation in Carnegie-Mellon University, USA, TOC and EBM systems in the Japanese Oral Translation Communication Research Laboratory. The Computer Department of Tsinghua University, China, has established an example-based Japanese-Chinese MT system. Example-based technology has also been used in the "Daya" system of computer writing and translation co-developed by Harbin University and Tsinghua University (Feng, 2010, p. 34).

Statistic-Based Method

Statistic-based MT was proposed by IBM researchers around 1990. The IBM system, with several years' development, could directly acquire translating knowledge from the corpus without manual adjustment of rules, which was a smash in the industry at that time. From 2002, the National Institute of Standard Technology, USA, supported by the Defense Advanced Research and Planning Agency, has conducted annual MT testing in which statistic-based MT has excelled the traditional rule-based method and become the mainstream hot point of MT research. With the rapid development of statistic-based methods in the recent years, based on phrase-based model, researchers have proposed many new types of syntactic-based statistic models and gained initial success (Zhao & Liu, 2010, p. 38).

The basic idea of statistic-based MT is that translation from source language sentences into target language sentences is an issue of probability, and any target language sentence has the probability of becoming the translation of any source language sentence, with the probabilities being different; therefore, the task of MT is to find out the most probable sentences. So far, the statistic-based method has experienced three periods: word-based model, phrase-based model, and syntactic-based model (Zhao & Liu, 2010, p. 37). Statistic-based MT, based on bilingual parallel corpus, abstracts a statistic model from the implicated translating knowledge in the corpus with statistical analysis, and then uses the model to translate. Statistic model includes translation model and language model. The function of the translation model is to calculate the possibility of translating one language strand into another one, often represented as a conditional probability.

The language model is used to calculate the possibility of a language strand appearing in the target language, that is, to calculate the syntactic and semantic rationality of the language strand in the target language, usually represented as an N variable model. Compared with rule-based MT or example-based MT, statistic-based MT, based on mathematic theory, presents translating knowledge in the form of probability, and its model is represented as parameters, so translation is realized by translation text retrieval with the parameters. Statistic-based MT acquires language knowledge from the corpus, so there is no need of manual writing of dictionaries and rules. As a result, statistic-based MT system can be conveniently transplanted to different languages and fields. However, statistic-based MT largely relies on the corpus, so the quality of the corpus can directly affect the establishment of statistic model.

The advantages of statistic-based MT are as follows: first and foremost, compared with the traditional rule-based method, a statistic-based MT system has a lower cost of manual work and a shorter period of development. Second, translating knowledge comes directly from larger-scale and authentic bilingual corpus, so genuine expressions frequently appears in the translation. Third, because of the

machine learning with parameters, translation is not related with language itself; the translation model can be rapidly transferred to new languages and new fields (Zhao & Liu, 2010, p. 38).

Statistic-based MT based on phrases has some inherent defects. For example, such problems as the overall reordering in the phrasal level, phrasal non-continuity, and phrasal generative power have restricted the further advance of the method. Consequently, researchers have to resort to syntax because theoretically the introduction of syntactic structure knowledge is helpful to solving those problems. Therefore, it can be seen from the development of statistic-based MT that statistic-based MT based on syntax has become a new trend after MT based on phrase. Observed from the current situation, some statistic-based MT systems based on syntax have obviously excelled those over those based on phrase.

For instance, the Hiero system and the ISI system, respectively, in NIST MT testing in 2005 and in 2006, and the MT system developed by Computation Research Institute, the Chinese Academy of Sciences are approaching or even have excelled the best system based on phrases. There are different ways to introduce syntactic knowledge into the statistic-based MT system; for example, syntactic knowledge is introduced into the word alignment model so that source language order is adjusted with syntactic knowledge before translation and re-ranked after it (Xiong, et al., 2008, p. 29).

Perspectives of Corpus-Based Machine Translation

So far, language knowledge used in statistic-based MT is limited. It is impossible for some MT problems to be solved if more complicated language knowledge is not introduced. Such problems in translation as syntactic rule validity, reference and discourse have not been solved and need to be further studied. It is believed that with the progress of deeper research, more language knowledge will be effectively integrated into statistic-based MT, which will step up to a higher level (Liu, 2009, p. 150).

Word-based and phrase-based methods do not use any language knowledge but a lexical approach of probability calculation. All the language knowledge is directly represented by lexical probability statistics. Now, statistic-based MT based on syntax has become a research hotspot. In the translation model based on linguistic syntax, syntactical knowledge has been fully utilized and overtaken phrase-based and formal syntax-based methods. It is the corpus, which the translation machine uses and with which it conducts comparative statistics that makes translation result reliable in the levels of word, phrase, and syntax. Therefore, MT technology on a reasonable and effective platform based on the corpus and other relevant technologies has made much progress at both the theoretical and practical levels. Nevertheless, problems concerning correspondence between discourse and semantics still exist. The semantics-based method is seldom used now, and only some work based on word meaning disambiguation slightly improves the current MT performance (Liu, 2009, p. 153). Rational semantic and pragmatic models should be built up in terms of future translation model, and the semantics-based method will be a trend in the development of statistic-based machine translation.

The future development of corpus-based MT systems focuses on two aspects. First, new achievements of social science and information technology will be integrated into the research to improve the performance and quality of the MT system (Li, 2004, p. 62). For example, neurolinguistics, composed of neuroscience and linguistics, will help to further understand the deepest mechanism with which humans process language, and the application of this theory will greatly increase the efficiency of corpus process; moreover, importance will be laid in the development and application of the new artificial intelligence computer technology, and how to flexibly use corpus will be one of the subjects in the artificial intelligence technology research. In the second place, a series of multi-mode corpus for MT will be established and developed, including sentence pattern banks, stylistic feature banks, cognitive knowledge banks, etc., and a series of corresponding tools of language process, analysis, and retrieval will be developed so that the current text-based corpus are expected to be more affluent in content and more flexible in use (Zhang & Zhang, 2010, p.57).

References

Backer, M. (1995). Corpus in translation studies: An overview and some suggestion for future research. *Target*, *7*(2), 242-251.

Chang, B. B., Zhan, W. D., & Zhang, H. R. (2003). Bilingual corpus construction and its management for Chinese-English machine translation. *Computer Assistant Terminology Studies*, *5*(1), 28-31.

Feng, Z. W. (2010). Corpus-based machine translation systems. *Terminology Standardization & Information Technology*, *15*(1), 28-35.

He, L. Z. (2007). A design of translation debase based on Chinese-English corpus. *Modern Foreign Languages*, *30*(2), 191-199.

Li, L. (2004). Corpus-based machine translation. *Shanghai Journal of Translators for Science and Technology*, *19*(2), 59-62.

Liu, Q. (2009). Recent developments in machine translation research. *Contemporary Linguistics*, *51*(2), 147-158.

Shao, Y. Q. (2010). Introduction of some machine translation terms. *Terminology Standardization & Information Technology*, *15*(1), 25-27.

Wang, H. F. (2003). Method and issues of example-based machine translation. *Terminology Standardization & Information Technology*, *3*(2), 33-36.

Xiao, W. Q. (2007). Parallel corpora applied translation studies. *Chinese Science & Technology Translators Journal*, *22*(3), 25-28.

Xiong, D. Y., Liu Q., & Lin, S. X. (2008). A survey of syntax-based statistical machine translation. *Journal of Chinese Information Processing*, *23*(2), 28-39.

Zhang, Y., & Zhang, X. D. (2010). The research on the corpus-based machine translation in business text environment. *Journal of Jilin Normal University (Humanities & Social Science Edition)*, *38*(3), 55-57.

Zhao, H. M., & Liu, Q. (2010). A brief introduction to machine translation and its evaluation. *Terminology Standardization & Information Technology*, *15*(1), 36-45.