TELKOMNIKA, Vol. 11, No. 12, December 2013, pp. 7125~7130
e-ISSN: 2087-278X

A Survey of Machine Translation Methods

Li Peng
Foreign Language School East China Jiaotong University, Nanchang330013, China
e-mail: pengli666@tom.com

Abstract

With the international exchanges become more frequent, human translation can't meet the need of society, and with the developing of computer technology, machine translation becomes feasible. Reviews the history of machine translation and analyzes the main methods of machine translation, finally puts forward of the suggestions on machine translation construction.

Keywords: *machine translation, rule-based approach, corpus-based approach*

1. Introduction

Stepping into Information Age, language, as the information carrier, has become the most significant means for human to communicate. But it has been considered as the barrier of communications between people from different countries. The problem of converting a language into another quickly and efficiently has become a problem of common concern for humanity.

Machine Translation is the process of converting a natural source language into another natural target language by computer. It is a branch of natural language processing and it has a close relationship with computational linguistics and natural language understanding. With the rapid development of the Internet and the integration of the world economy, how to overcome the barrier of language become the common problems of the international community [1]. Translation by people can't meet the demand of the society, so the use of machine translation to help people to obtain information has become an inevitable trend.

2. Researches on Machine Translation

The history of machine translation research can be traced back to the forties of the 20th century. American scientist W. Weaver and British engineers AD, Booth proposed the idea of machine translation. In 1954, Georgetown University collaborating with IBM completed, for the first time, the Anglo-Russian Machine Translation with the IBM-701 computer. It proved the feasibility of a machine translation. Thus studies on machine translation began.

In 1964, in order to evaluate the progress of the research on machine translation, Automatic Language Processing Advisory Committee (ALPAC), began a two-year comprehensive survey and testing on machine translation. In November 1966, the Commission published ALPAC report. The report provided a comprehensive denial of the feasibility of machine translation, and recommended to stop the financial support of the machine translation project.

From the 1970s, with the increasing frequency of the exchange among countries, machine translation is urgently needed by society. Meanwhile, the development of computer science, linguistics research, particularly the increase in computer hardware technology and applications of artificial intelligence in natural language processing promote the recovery of study on machine translation. Machine translation projects began to develop. A variety of practical and experimental systems has been introduced .such as, Weinder system, the EURPOTRA multilingual translation system, TAUM-METEO system.

With the universal application of the Internet, the acceleration of the integration process of the world economy, machine translation enters a new development level.

Received April 1, 2013; Revised July 26, 2013; Accepted August 26, 2013

3. The Main Methods of Machine Translation

The machine translation process can be simplified to three stages: the analysis of original-language text, the transformation from original-language text to target-language text and the target-language generation. Seemingly, the core issue of machine translation is the accuracy of the translation. In fact, from the aspect of technology, the core issue is the methodology adopted by the machine translation system. From the methodological level, the machine translation system can be divided into rule-based machine translation and corpus-based machine translation. Rule-based translation systems can be divided into three catalogs: literal translation method, interlingua-based method and transfer-based method. Traditionally, corpus-based translation method can be divided into two different classes: the statistic-based translation and the case-based translation method. In rule-based machine translation, translation knowledge base consists of dictionaries and grammar rules. In corpus-based machine translation, the application of the corpus is the core. Knowledge base consists of a divided marked corpus. In other words, the rule-based machine translation is in the scope of rationalism, and corpus-based machine translation is in the scope of experience [2].

3.1. Rule-based Approach

1) Literal translation method

Early machine translation system is basically using rule-based approach. In 1954, the world's first machine translation system IBM701 completed the world's first machine translation with literal translation. Literal translation is called direct translation, word-based translation or dictionary-based translation. The literal translation method is that the words will be translated as a dictionary does- word by word, usually without much correlation of meaning between them [3].

The so-called literal translation is to transforming the word or sentence in the source language into the corresponding word or sentence in the target language. When necessary, adjust the word order. The literal translation is generally designed for a particular language pair. It is not versatile.

Systran is a typical literal translation system. At the beginning, it only translations Russian into English. Later it improved and it can realize the translation among different language. Systran has a great influence on the development of machine translation. Nowadays, there are still many translation systems adopting literal translation.

2) Transfer-based method

With the improvement of literal translation method, the transfer-based method appeared. The transfer-based method is to analyze the sentence structure, and on the basis of word-to-word translation, generating the target-language text according to the different linguistic rules of different languages. There are three dictionaries available: the source language dictionary, the source language-target language bilingual dictionary and the target language dictionary.

The first stage of the translation is to analyze the input text for morphology and syntax (and sometimes semantics) to create an internal representation. By using both bilingual dictionaries and grammatical rules, the translation will be generated out from this representation. As shown in Figure 1:

Table 1.

SL = Source Language
TL =Target Language
IR = Internal representation

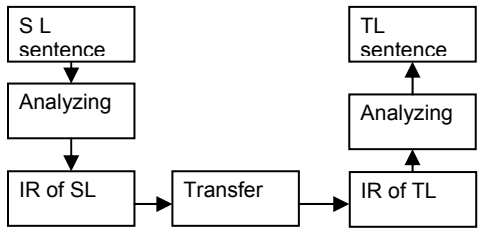

Figure 1. Transfer-based Machine Translation

3) Interlingua-based method

Along with the development of literal translation method and transfer-based method, interlingua-based method came into being. Firstly, the source language is analyzed, and converted into an Interlingua firstly, which is an abstract language-independent representation and apply to all languages. And on this basis, the source language is converted into the target language.

The Interlingual-based method can be considered as better alternative choice, specially compared to the literal method and the transfer-based method. That is, there are two translation process phases: from the source language to the intermediate language and from the intermediate language to the target language.

The advantage of Interlingua-based method is that it provides an economical way to realize multilingual translation. With interlingual system, it becomes unnecessary to make translation pairs between each pair of languages in the system. Instead of creating N*(N − 1) language pairs (N is the number of languages in the system), the system only needs 2N pairs between the N languages and the only interlingua [4]. As shown in Figure 2:

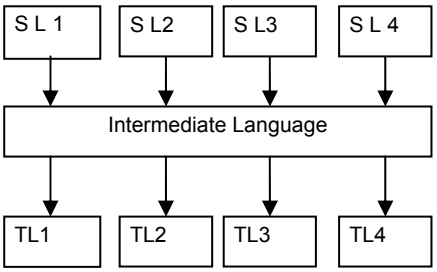

Figure 2. Interlingua-based Machine Translation

However, it is too difficult to define a kind of proper Interlingua. The ideal of interlingua-based machine translation is limited in a very narrow specific domain.

Throughout the course of the development of machine translation, it can be seen, the main method of machine translation has been a rule-based approach. The thought of rule-based machine translation technology is generally accepted, mature, and it is so far the most widely used method. Rule-based machine translation system is to analyze, judge, and choose the morphology and semanteme of the source language, then rearranged, finally generate the equivalent in the target language.

Rule-based machine translation methods need the cooperation between the programmers and linguists. Firstly, they jointly develop data standard, determine the translation algorithm, the representation on knowledge of language and translation, and then the programmers program translation algorithm, linguists prepare the knowledge of language and translation to drive translation algorithm run. The programmers and linguists are indispensable.

Researchers made innovative improvements on the rule-based approach. They made a deeper analysis on language, then convert and generate on this basis. In other word, the translation is not only in the surface (word sequence), but in a deeper structure of sentences (e.g. syntactic structure, semantic structure or knowledge representation) level.

The performance of the translation system is usually constrained by two factors: First, the design of the algorithm is reasonable. Second, the knowledge of the language is rich enough. The main bottleneck lies in the latter.

With the deep research in machine translation, the problems on rules exposed. The biggest problem is on knowledge-obtaining. Manually written rules of the linguists can't meet the actual needs in the application.

Manually adding rules makes rule base larger, and system performance improvement is more difficult. Because the larger the rule base, the more conflicts between rules appear---the so-called "seesaw phenomenon". Although to some sentences the translation effect is better, but to some other sentences the translation effect is worse. On the other hand, the new-adding rules are used to deal with very individual language phenomena. There is little

improvement to the whole system performance. These simple methods can not deal with slightly more complex sentences.

With the development of research, people come to realize: To get a high-quality target-language text through machine translation, to a certain extent, the computer must be able to understand the source language sentences. At the same time, the artificial intelligence has been greatly developed in the 1970s, a variety of knowledge representation and knowledge reasoning theory and algorithms have been proposed by the researchers. People's understanding to natural language and machine translation has been a qualitative leap [5].

3.2. Corpus-based Approach

The Corpus-based machine translation can be divided into two main classes, the statistic-based machine translation and the case-based machine translation.

1) Statistic-based method

Statistic-based approach treats machine translation as a process of information transmission, using a channel model to interpret machine translation. This idea is that any sentence in a language would be the translation of a sentence in another language, but the probability varies. The task of machine translation task is to find the maximum probability sentences. Therefore, statistic-based machine translation can be divided into the following aspects: model problems, training problems and decoding problem. The model problem is to establish a probability model for machine translation, that is, to define the Calculation method of the probability of sentence translation from the source language to the target language. Training problem is to make use of the corpus to get all the parameters of this model. The decoding problem is to find the maximum probability translation for any sentence of source language on the basis of the known models and parameters.

The mathematical model of statistical machine translation methods is proposed by the researchers in International Business Machines Corporation (IBM). The basic idea of the statistic-based machine translation is that this model only considers a linear relationship between words, and ignores the structure of the sentence. If the word order of the two languages is totally different, the effect of translation may not be very good. If syntactic structure or semantic structure is taken into account besides considering the language model and translation model, the effect of translation will be better.

The technology of statistic-based machine translation itself has also been an evolving process. The statistical cluster translation model framework develops from noise channel model to log-linear model. And the main statistical machine translation model develops from early word-based model to phrase-based model and syntax-based statistical translation model.

Google Translator has been well known, and the technology behind it is the statistic-based approach. The basic operating principle is to search for bilingual web content, and take it as the corpus, then the computer automatically select the most common corresponding word, and finally gives the translation results. Statistic-based machine translation requires large-scale bilingual corpus. The accuracy of the translation model and language model parameters directly depends on the corpus. The translation quality depends mainly on the probability model and coverage of the corpus. Although it doesn't need to rely on a large number of knowledge, it disambiguates and selects the translation text based on the statistical result, the corpus selection and processing is a hard work.

Statistic-base machine translation considers the language as a kind of meaningless string, it is not appropriate in analyzing the semantic information and reconstructions of the translations. The results of the experiments are far away from our imaginations. And there are still many problems are waiting to be eliminated, such as the problem, which was proposed by Chomsky, of how to deal with long-distance constraints like subject-verb concord, the statistic-based machine translation has been proved that it is not very effective in practice [6].

2) Case-Based method

The same as statistic-based method, case-based approach method is: firstly, correctly decompose source language text into sentences, then decompose the sentences into phrases, and translate these phrases into the target language phrases by analogy, finally, merge the phrases into sentences. In other words, the translation process of this method is in the following. First, a bilingual alignment case base will be established. Input the sentence of source language, the system will search for the most similar sentence from the Source

language base, then according to the translation of target language, some words or phrases in target language will be changed and the translation of target language will be done.

This process imitates the thinking process of human translation. Makoto Nagao believes that the first step of translation work in human minds is breaking a sentence into phrases and words. The purpose of translating these different parts is to compose these fragments into one long sentence again. Phrasal translations are performed in seeking similar words or phrases previously. The principle of this kind of translation system is encoded to example-based machine translation system, which use enormous case base as the foundation of the system [7].

To the case-based machine translation, the main source of knowledge is the bilingual case library. The core problem is to maximize the statistics and form a bilingual case library.

Case-based machine translation have a very significant effect for the same or similar text translation, with the increase in the scale of sentences library, its role has become increasingly significant. To the existing text in the case library, we can directly get high-quality translation. To the similar text in the case library, we can get the translation by analogical reasoning, and slightly revise the translation. However, the case-based machine translation requires a large corpus to support, the actual demand of language is very huge. With limited corpus scale, case-based machine translation is difficult to achieve a higher matching rate, so it is often limited to the relatively narrow or specialized fields, the translation effect can meet the requirements.

In practice, some problems should be solved. For example: The measurement of similarity is hard to be defined. To select out the similar sentences, many ambiguous selection items will be listed out. It will slow down the system and spoil the translation quality finally. Besides, the problem of alignment of bilingual text should be considered. Sentence alignment is not the only requirement, most of time the phrase alignment or even lexical alignment is also needed.

So far, very few machine translation systems only adopt case-based approach. Case-based approach is only used as one of the multiple translation engines, in order to improve the correct rate of the translation

4. Advantages and Disadvantages of Each Approach

It should be said, whether it is the literal translation method, transfer-based method, interlingua-based method, or statistic-based method and case-based method all have advantages and disadvantages. As in the scope of rationalism, the literal translation method, transfer-based method, interlingua-based method, are rule-based approaches. The typical disadvantage is that the grain size is too large, that is, the computer language can not fully describe the actual infinite language infinite rules.

Statistic-based method and case-Based method are corpus-based approaches, the typical disadvantage is the sparse data. In other words, because of the infinite language, any high-performance computer can not count all usages of the phrase.

With the disadvantages of these approaches appear, since the 1990s, more and more approaches and strategies are integrated in machine translation. In machine translation research in the future, a variety of approaches interact and integrate. Machine translation should be combined rule-based approach with corpus-based approaches and machine translation should be combined with translation memory.

5. Suggestions on Machine Translation Construction

After analyzing the advantages and disadvantages of each machine translation method, some suggestions are put forward.

1) Focusing on the integration of a variety of methods in machine translation

It is difficult to get a satisfactory translation with a single rule method. For example, statistic-based approach focuses on the common phenomenon in language, and ignores the individual phenomenon. It can't grasp enough the flexibility of the language. However, rule-based approach can deal with this problem effectively. In a word, using multi-strategy machine translation method, combining rules, corpus with semantic methods to complete the machine translation system is an effective way to obtain high-quality translations.

2) Strengthening the role of the semantic analysis in translation

To get a high-quality translation, we should introduce and strengthen the means of semantic analysis, study the expression and language comprehension and analyze the deep relationship between the elements and the deep structure of the sentences, so that the computer can translate correctly the original text accurately on the basis of understanding.

3) Introducing linguistic knowledge

Syntax-based statistical method has taken the first step in the introduction of linguistic knowledge in statistical-based machine translation. With the deepening of the study, introduce continually a variety of linguistic knowledge (for example, syntactic knowledge, semantic knowledge etc.) has become a trend.

4) Constructing the of language resources

Massive corpus is important to help to achieve a good translation, but there is a lot of noise which has a great influence on the model. A feasible solution is to study automatic processing methods, at the same time, strengthen the manual processing and handling of the corpus. Although it is time-consuming, but people can permanently benefit from it.

5) Strengthening study on evaluation methods

To encourage competition, promote the benign development, finding a fair and reasonable machine translation evaluation method has become an important project. Evaluation is the driving force of machine translation development, so a fair and reasonable evaluation of machine translation evaluation is very important for machine translation development. Evaluation methods will affect to machine translation research methods. Many of the current evaluation systems focus on improving the indexes, sometimes they deviated from the technology research. Therefore, Studying on a reasonably designed, practical and effective machine translation evaluation method is also an important research direction.

References

[1] Chris Callison-Burch, Philipp Koehn, Christof Monz, Omar F Zaidan. *Findings of the 2011Workshop on Statistical Machine Translation.* Proceedings of the 6th Workshop on Statistical Machine Translation, Edinburgh. 2011; 22–64.
[2] Sebastian Pado, Michel Galley, Dan Jurafsky, Chris Manning. *Robust Machine Translation Evaluation with Entailment Features.* Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Singapore. 2009; 297–305.
[3] John DeNero, Shankar Kumar, Ciprian Chelba, Franz Och. *Model Combination for Machine Translation.* Human Language Technologies. The 2010 Annual Conference of the North American Chapter of the ACL. Los Angeles. 2010; 975–983.
[4] LV Stoimenov, EO Kajan, S Djordjevic-Kajan. Ontology-Driven Semantic Interoperability. *International Review on Computers and Software.* 2006; 1(2): 132–136.
[5] WJ Hutchins. *Retrospect and prospect in computer-based translation.* Proceedings of MT Summit VII, Singapore. 1999; 30-34.
[6] Cameron Shaw Fordyce. *Overview of the IWSLT 2007 evaluation campaign.* Proceedings of International Workshop on Spoken Language Translation, Trento. 2007: 3-27.
[7] Yang Liu, Qun Liu, Shouxun Lin. *Tree-to-string alignment template for statistical machine translation.* Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney. 2006; 609-616.
[8] G Doddington. *Automatic evaluation of machine translation quality using n-grams co-occurence statistics.* Human Language Technology: Notebook Proceedings, San Diego. 2002; 128-132.