The Development History of Machine Translation

Xiaolei Wan

Guagndong University of Science and Technology, Dongguan, China.

Abstract

The development of machine translation technology has been closely associated with the development of computer technology, information theory, linguistics and other disciplines. From early dictionary matching to regular translation based on dictionary knowledge of linguistics experts, to statistical machine translation based on corpus, with the improvement of computer computing power and the explosive growth of multilingual information, machine translation technology has gradually stepped out of the ivory tower and started to provide real-time and convenient translation services for ordinary users.

Keywords

Machine Translation; Social Science Translation; Rhetoric; Linguistics.

1. The development road

The research history of machine translation can be traced back to the 1930s and 1940s. In the early 1930s, French scientist G.B. Alchovny came up with the idea of using machines for translation. In 1933, Soviet inventor П. П. Luo Yang, designed the translate one language into another language machine, and in the same year on September 5, registered his invention; However, due to the low level of technology in the 1930s, his translation machine was not made. In 1946, the first modern electronic computer ENIAC was born, and soon after that, the pioneer of information theory, American scientist W. Weaver and British engineer A. D. Booth put forward the idea of automatic language translation using computers in 1947 when discussing the application scope of electronic computers. In 1949, W. Weaver published the Translation Memorandum, which formally put forward the idea of machine translation. After 60 years of ups and downs, machine translation has experienced a tortuous and long development road, which is generally divided into the following four stages in the academia. In 1954, Georgetown University, in collaboration with IBM, completed the first Experiment of Machine translation in Britain and Russia with IBM-701 computer, demonstrating the feasibility of machine translation to the public and the scientific community, thus opening the prelude to machine translation research. In 1964, in order to evaluate the research progress of machine translation, the National Academy of Sciences established the Automatic Language Processing Advisory Committee (ALPAC), and began a two-year comprehensive investigation, analysis and test. In November 1966, the Committee published a report entitled "Language and machine" (ALPAC report), which completely rejected the feasibility of machine translation and recommended that funding for machine translation projects be discontinued. The publication of the report is a blow to the burgeoning field of machine translation, which has come to a near standstill. Coincidentally, during this period, the "Ten-year Cultural Revolution" broke out in China, and these studies basically stalled. Machine translation has entered a slump. From the 1950s to the first half of the 1960s, the research on machine translation has been on the rise. For military, political and economic purposes, the two superpowers of the United States and the former Soviet Union provided a large amount of financial support for the machine translation project, while European countries also paid considerable attention to the machine translation research due to geopolitical and economic needs, which led to a sudden upsurge of machine translation. Machine translation, though just at the beginning stage, has entered an optimistic period of prosperity.

2. System partition

The machine translation system can be divided into two categories: rule-based and corpus-based. The former consists of a dictionary and a rule base. The latter is composed of a divided and annotated

corpus, which requires neither dictionaries nor rules, and is dominated by statistical laws. Machine translation system develops with the rise of corpus linguistics. Most of the world's machine translation systems adopt rule-based strategies, which are generally divided into grammatical, semantic, knowledge-based and intelligent types. Different types of machine translation systems consist of different components. In the abstract, the processing of all MACHINE translation systems includes the following steps: the analysis or understanding of the source language, the transformation in a plane of the language, and the generation of the target language according to the rules of the target language structure. The technical difference is mainly in the transformation plane.

3. Type words

From the machine translation experiment of Georgetown University to the system of the late 1950s, it basically belongs to this kind of machine translation system. Their characteristics are as follows: (1) The establishment of a bilingual dictionary is centered on the transformation of vocabulary. In translation, the purpose of sentence processing is to determine the corresponding equivalent words in the target language; If one word of the source language corresponds to several words of the target language, the machine translation system itself cannot decide which one to choose, but can only output all possible choices; Language and program are not divided, the rules of grammar and the algorithm of the program mixed together, the algorithm is the rule. Due to the above characteristics of the first type of machine translation system, the quality of its translation is extremely poor. Moreover, the design of such a system is a very trivial and complicated work, and there is no room for expansion after the system is designed, and the modification affects the whole body, causing great difficulties to the improvement of the system. The research focuses on morphology and syntax, which are represented by the above irrelevant grammars. Most of the early systems belong to this type. The grammatical system consists of three parts: source-language analysis mechanism, source-language-to-target language conversion mechanism and target-language generation mechanism. The source text analysis mechanism analyzes the input source text, which can be divided into lexical analysis, grammatical analysis and semantic analysis. This analysis gives you some form of internal representation of the source text. The transformation mechanism is used to transform an internal representation that is relatively independent of the surface expression of the source text into an internal representation corresponding to the target language. Target language generation mechanism realizes the transformation from internal representation of target language to surface structure of target language. Most of the machine translation systems established since the 1960s are of this type. Their characteristics are as follows: (1) To put syntactic research in the first place, first use coded structural marks to represent the structure of the sentence in the original language, then convert the structural marks of the original language into the structural marks of the target language, and finally constitute the output sentence of the target language; (2) For polysemous words, special treatment must be carried out, according to the context to choose the appropriate meaning, not allowed to list a number of translated words package; (3) grammar and algorithm separately, under certain conditions, the syntax is within the limits of a certain category, can make the grammar by a given algorithm to calculate, and can be described by the given algorithm for the corresponding formula, so don't change the transformation algorithm can also be in syntax, in this way, the grammar you can write and modify the algorithm. The type 2 machine translation system is a step ahead of the type 1 machine translation system in terms of both the quality of translation and the convenience of use.

4. Semantics

The research focuses on the introduction of semantic feature information in the process of machine translation, represented by the semantic grammar proposed by Burtop and the case framework grammar proposed by Charles Fillmore. Various theories and methods of semantic analysis mainly solve the unification of form and logic. By using the rules of semantic segmentation, the input source text is segmented into several related semantic elements. Then according to the rules of semantic transformation, such as keyword matching, find out the corresponding semantic internal

representation of each semantic component. By testing the relationship among the semantic elements, the system establishes the logical relationship among them and forms the semantic representation of the whole text. The processing is mainly achieved by looking up the semantic dictionary. Semantic representation is generally lattice framework or concept dependency representation. Finally, the translation system interprets the intermediate semantic representation to form the corresponding translation.

Since the 1970s, some machine translators have put forward a semantic third type of machine translation system. After introducing semantic plane, requires some substantial change in the aspect of language description, because in a predominantly syntactic machine translation system, is the smallest translation unit in the word, one of the largest translation unit is a single sentence, the algorithm of machine translation only consider automatic processing of a sentence, without regard to belong to different sentences of the connection between the word and the word. Type 3 machine translation systems must consider problems beyond the scope of sentences. In addition to sememes, words, phrases, and sentences, they must also study paragraphs and chapters larger than sentences. In order to establish the third type of machine translation system, linguists need to study semantics deeply, mathematicians need to develop the algorithm of semantic representation and semantic processing, and in the aspect of programming, they also need to consider the characteristics of semantic processing.

5. The knowledge

The goal is to equip machines with human knowledge to achieve understanding-based translation systems, represented by the knowledge-based machine translation system proposed by Tomita. Knowledge-based MACHINE translation system uses a large semantic knowledge base to transform source text into intermediate semantic representation, and then uses professional knowledge and daily knowledge to refine it, and finally converts it into one or more translation outputs. The goal is to use the latest achievements of artificial intelligence to realize multi-path dynamic selection and automatic recombination of knowledge base, and to transform different sentences on different planes. In this way, grammar, semantics and common sense can be connected into an organic whole, which can not only inherit the advantages of the traditional system, but also realize the function of self-growth of the system. This type of system is represented by the IMT/EC system developed by the Institute of Computing Technology of the Chinese Academy of Sciences.

6. Statistics-based editing

The general machine translation system Based on corpus-based is the machine translation Based on statistics. Because this field is a new force, statistics is statistical parallel Corpus, and many different statistical models are derived from it. Unlike the rule-based machine translation system, which is composed of a dictionary and a grammar rule base, the corp-based machine translation system takes the application of corpus as the core and consists of divided and annotated corpora to form the knowledge base. Corpus-based methods can be divided into statistics-based methods and example-based methods.

7. Statistics-based machine translation

The statistics-based machine translation approach regards machine translation as a process of information transmission and uses a channel model to interpret machine translation. According to this idea, the translation from source language sentences to target language sentences is a matter of probability. Any target language sentence may be a translation of any source language sentence, but with different probabilities. The task of machine translation is to find the sentence with the highest probability. The concrete method is to regard translation as the decoding process of translating the original text into the translation through the model. Therefore, statistical machine translation can be divided into the following problems: model problems, training problems, decoding problems. The so-called model problem is to build a probabilistic model for machine translation, that is, to define the

calculation method of translation probability from source language sentences to target language sentences. The training problem is to use the corpus to get all the parameters of the model. The so-called decoding problem is to find the translation with the highest probability for any input source language sentence on the basis of known model and parameters.

References

[1] Andrew Fitzmaurice. Humanism and America. New York: Cambridge University Press, 2007.

[2] Andrew Radford. Linguistics: An Introduction. New York: Cambridge University Press, 2009.

[3] Hu shu-chung. Rhetoric in contemporary English. Shanghai: Shanghai Foreign Language Education Press, 2013.

[4] Jin yitong. From "creativity" to "creative translation": the application of translators' creative thinking -- a case study of the English translation of xi jinping on the governance of China [J]. Shanghai translation, 2019 (01): 57-62 95.

[5] Li Heng-wei. High Quality Machine Translation Within reach? N. Journal of Social Sciences, 2018-04-26005.