Analysis and Prospect of Natural Language Processing Research from Machine Translation Perspective

Shuqi Fan

Beijing University of Posts and Telecommunications

1030103385@qq.com

Keywords: natural language processing; machine translation; development process; prospect

Abstract: With the rapid development of computer natural language processing technology and the accelerating globalization process, machine translation has ushered in a boom. Many ideas that were previously difficult to implement can turn from theory to reality. On the one hand, it can be seen that machine translation relies on traditional theory of natural language, and on the other hand, thanks to the development of natural language processing technology. Under this opportunity, machine translation has developed in an all-round way and has been a world-wide prosperity. This paper attempts to sort out the development process of machine translation and analyze the technology at each stage. Based on these, the article puts forward the prospect of natural language processing and machine translation at the end.

1. Introduction

1.1 Machine translation

Machine translation, also known as automatic translation, is the process of transforming one natural source language into another language of natural target by using computers. It generally refers to the translation of sentences and full texts between natural languages. It is a branch of Natural Language Processing.

With the rapid development of economic globalization and Internet, machine translation technology plays an increasingly important role in promoting political, economic and cultural exchanges. From the dictionary matching of the early stage to the rule-based machine translation based on linguistic expertise, and to the statistical machine translation based on the corpus, with the improvement of computers' computing capacity and the explosive growth of multi-language information, the technology of machine translation gradually steps out of the ivory and starts to provide the real-time and convenient translation services for ordinary users.

1.2 Natural language processing

Natural language processing is an important direction in the field of computer science and artificial intelligence. It studies various theories and methods that enable effective communication between humans and computers in natural language.

Language is a unique feature of human beings. Human thoughts and wisdom are based on language and are circulated through the recording of language and characters. Therefore, it is also an important and even core part of artificial intelligence.

Obtaining the meaning of natural language is not the ultimate goal of natural language communication between human and computer. Understanding the intentions and thoughts of natural language is the most important thing. The former is called natural language understanding and the latter is called natural language generation.

However, achieving either natural language understanding or natural language generation is very difficult. From the current theoretical and technological status, the universal and high-quality natural language processing system is still a long-term goal. But for certain applications, practical systems with considerable capacity of language processing have emerged.

It is precisely because of the wide variety of ambiguities that exist in natural language texts and

at all levels of dialogue that it is very difficult to achieve natural language understanding and natural language generation.

The current problems can be roughly divided into two categories. The first one is that current grammar analysis cannot accurately determine the impact on statements and the relationship between them. Therefore, there are no clear rules for the different meanings of the same sentence on different occasions. The second type is that the knowledge needed to understand the sentence cannot be stored in the computer completely, which leads to higher limitations.

2. The development process of machine translation

2.1 Rule-based machine translation: RBMT

The source of machine translation dates back to 1949, when researcher Warren Weave formally proposed the concept of machine translation. In 1954, the world's first translation machine, IBM-701, was officially launched. Although there are only six grammar rules and 250 words, it marks the birth of machine translation and has epoch-making significance.

However, in 1966, due to the mistake in the report of the Automatic Language Processing Advisory Committee that machine translation was not worth investing, the development of machine translation in the United States fell into stagnation for decades.

From the 1950s to the 1980s, machine translation was translated literally according to the dictionary. Although there are syntactic rules to correct, the translation results are still not directly applicable, because the language is complex and ambiguous. It is impossible to exhaust all, so the method was obsoleted after that.

The following example shows more clearly that rule-based translation has poor feasibility.

The Seniors were told to stop demonstrating on campus.

The first solution of this sentence:

1) The seniors were demonstrating and were asked, on campus, to desist.

The second one:

2) The seniors were demonstrating and were asked to desist on campus (although they could prove elsewhere).

The third one:

3) The seniors were demonstrating on campus and were asked to desist.

A sentence with such a simple concept can have such a complicated rule system. If the amount of machine translation rules is used, it will be an amazing astronomical number. Therefore, rule-based machine translation ideas are difficult to become mainstream.

2.2 Example-Based Machine Translation: EBMT

After the failure of the Rule-based machine translation method, machine translation was hit hard, but Professor Nagao Masato of Kyoto University in Japan proposed Example-Based Machine Translation, which is to use enough examples to match the sentences that need to be translated. Even if the sentence to be translated is not exactly the same as the example sentence, the translation can be done by replacing the different words. But this theory is naive and as bad as the effect of the Rule-based machine translation, so it has not received much attention.

2.3 Statistical Machine Translation: SMT

IBM once again led the development of machine translation. In the 1993 "Mathematics of Machine Translation" paper, five statistical models based on words were proposed. The statistical model uses parallel corpus in principle to perform statistics on a word-by-word basis. For example, the word "太阳", although the machine does not know its English, but after the statistics, as long as the word "太阳" appears in the sentence, the word "sun" will appear in the corresponding English. With this statistical method, it is possible to automatically understand the meaning of words without manually maintaining the machine translation system.

In fact, limited by the poor computational power and the inadequate samples of training, the

similar concepts that Warren Weave proposed at the earliest were not applied. The most important machine translation parallel corpus now is mainly from the United Nations, because its resolutions and announcements are available in various languages.

There are several methods for statistical machine-based translation.

1) Word-based SMT

Split the sentence into words, then count the statistics, remember the position of the word in the output sentence, and rearrange the words in the middle step to make the sentence sound more natural. If the machine feels it is necessary to add a new word, it first inserts a NULL tag and selects the appropriate grammatical auxiliary for each tag word. The concept of "relative order" was introduced, which means that the model learns whether two words are frequently swapped. This approach adds more parameters to learning and solves the problem of conflicts of word position.

2) Phrase-based SMT

This method inherits all word-based SMT principles included statistics, reordering, and vocabulary modifications. However, this method breaks down the text into phrases rather than words. This method is derived from the n-gram. Machine learning greatly improves the accuracy of translation by translating a stable combination of multiple words.

3) Grammar-based SMT

It performs a very precise parsing of the sentences - identify the subject, predicate, and other parts, and then create a tree structure of a sentence. Through this tree structure, the machine can learn to convert grammar units between languages. The problem of word alignment has been completely solved.

2.4 Neural Machine Translation (NMT)

In his 2014 paper, Yoshua Bengio first used RNN to automatically capture the word features between sentences, laying the foundation for techniques of deep learning for machine translation. Soon Google invested a lot of human and material resources, using neural network machine translation instead of all Statistical Machine Translation in 2016, and finally neural network machine translation became mainstream.

The most striking feature of Google Translation is the use of attention mechanisms, which first read the statement and pick out a few key words to confirm the semantics. This mechanism ultimately reduces the error rate to 40% of the statistical machine translation system when multiple languages are translated.

However, neural networks require a lot of corpus and it is difficult for humans to understand the specific structure after training. This leads to its imperfections, and humans cannot correct them. Only by using more correct corpora can the neural network be optimized.

In 2018, the machine translation system developed by Microsoft reached the level comparable to human translation in the Chinese-English translation test set of Newstest2017. This marks a breakthrough in neural network machine translation. Its use of Dual Learning and Deliberation Networks is a big highlight. Dual Learning solves the problem of limited parallel corpus and continues to revise and improve based on the difference between results and answers of machine translation. Deliberation Networks mimic the process of human translation. But no matter what neural network needs to be corrected and optimized with reference to humans.

3. The prospect of natural language processing and machine translation

Since 1949, machine translation has undergone several stages of development, and the results of its translation have been developed from unreadable to readable, but there is still a certain distance from full application, and there is still a great room for development. The following is an exploration of its feasible development direction from three aspects.

3.1 Expanding the scale of parallel corpora

The corpus is the basic resource for carrying language knowledge. The parallel corpus is composed of the original text and its corresponding translated text. The development and

application of the corpus open up a new way of machine translation. The corpus can be used for query. We can also classify and analyze the corpus, which is very helpful for translation research. We translate by comparing the original text of the corpus with the sentences that need to be translated. The corpus can provide a large number of reference materials to choose from. The richer the corpus, the higher the similarity with the original text, so expanding the size of the corpus can make the examples available for translation more abundant. The accuracy of translation is also improved.

3.2 Rational use of resources of big data

The expansion of the corpus will lead to a rapid increase in the amount of data. The traditional method of data management is difficult to manage such a large deal of data, which has already exceeded the limit of its processing power. Big data can identify synonyms, terms, new words, industry terms, and use parallel processing tools to automatically generate syntactic categories, which can be seen to be closely related to the accuracy of machine translation, so big data technology is indispensable.

3.3 Cloud computing accelerates the development of machine translation

The corpus is getting bigger and bigger, the data is getting more and more, and the running speed and operational security of the data become a problem that must be solved. Cloud computing, as a technology that is theoretically unlimited in scale and capacity, can be used anywhere in the network. The local computer only needs to send a demanded message over the Internet. At the remote end, there are thousands of computers that provide the required resources on request and return the results to the local computer. In this process, the local computer requires almost no operations, and all processing is done by a computer cluster provided by the cloud computing provider. Cloud translation can collect, transfer, store and utilize massive corpus information. Machine translation with cloud computing can be applied to such things as Google Translate, Bing Translator, etc. Submit the text that needs to be translated locally and use cloud computing to translate and return the translation results.

Machine translation, as a subclass of natural language processing, is also one of the core components of natural language processing. The current hotspots of natural language processing are consistent with machine translation, both of which are deep learning. Natural language processing based on deep learning has made some progress in recent years. The application of human-machine dialogue, question-answering system and language translation has always been a hot topic in natural language processing. The realization of these applications depends on advances in technologies and models. In addition, the field of natural language processing is a multidisciplinary industry, and the future development of natural language processing has far-reaching influence on many related disciplines and directions.

References

[1] Gong Mingzhu. Discussion on the Development History and Prospect of Machine Translation [J]. Journal of Educational Institute of Jilin Province, 2009(7):135-136.

[2] Lu Li. Research and Exploration of Machine Translation and Computer Aided Translation [J]. Journal of Southeast University Philosophy and Social Science Edition, 2002, 4(3):175-179.

[3] Zhao Qing. A Review of Foreign Research Based on Corpus Translation [J]. Journal of Chongqing Jiaotong University (Social Sciences Edition), 2008, 8(3):100-104.

[4] Lin Yiou, Lei Hang, Li Xiaoyu. Deep learning in natural language processing: methods and applications [J]. Journal of University of Electronic Science and Technology of China, 2017, 46(6).