Original Paper

Machine Translation of Concise State Sentences for Quantifier Phrases

Wu Min[1*]

[1] Department of Language Information Processing, Information Engineering University Luoyang Campus, 471000, China

Received: September 22, 2019 Accepted: October 19, 2019 Online Published: December 30, 2019
doi:10.22158/jetss.v1n2p138 URL: http://dx.doi.org/10.22158/jetss.v1n2p138

Abstract

With the development of deep learning methods, the machine translation system based on deep neural network has reached a very high accuracy, but for some daily Chinese phenomenon machine translation system is still not able to translate correctly. In this paper, we study a sentence that often appears in Chinese spoken language, that is, a simple state sentence composed of quantitative phrases, and improve the existing machine translation system. The external helper program constructed in this paper is compatible with the current mainstream network translation systems, greatly improving the translation effect of these translation systems on the concise state sentences composed of quantitative phrases.

Keywords

Deep learning, NMT, concise state sentence, HNC

1. Machine Translation System

1.1 Development of Machine Translation Systems

With the further development of economic globalization and the rapid development of the Internet, the communication between countries in the world has become more and more close, and the consequent demand for language translation has become greater and greater. Although the accuracy of manual translation is not low, it is difficult to cope with the growing demand for translation. Therefore, machine translation technology has been paid attention to by various countries and achieved certain achievements. In 2018, Microsoft announced that the translation quality and accuracy of its English-Chinese machine translation system can be compared with the translation system of human translation (News translation from China to Britain is comparable to human, and Microsoft machine translation has made a new breakthrough, n.d.).

The development of machine translation technology can be divided into four stages: the first is the pioneering period, from 1947 to 1964. In 1954, Georgetown University completed the English-Russian machine translation experiment for the first time with the assistance of IBM, which kicked off the study of machine translation. Then in 1956, China began to include this research in the national development plan of scientific work. The second phase, from 1964 to 1974, was a period of very slow development, called stagnation. In November 1966, the ALPAC committee published a report that totally denied the feasibility of machine translation, and recommended that funding for machine translation projects be stopped. This report was released to the booming machine translation, and machine translation research came to an impasse.

Domestic related research also stopped also for political reasons, in the period from 1975 to 1989 is the rapid development of machine translation research, investigate its reason is because of the national science and technology intelligence communication increasingly frequent, the language barrier problem between countries become more and more serious, the traditional human translation cannot meet the demand of such a huge, all countries in the world urgently needs a computer for language translation, machine translation during this period by the attention of the research institute, has greatly improved its machine translation technology.

This is also true in China, where machine translation has been included in major research programs such as “863” and conducted collaborative research by multiple research institutes, bringing China’s machine translation into a period of rapid development. It has been called the new period of development since 1990. With the global popularization of the Internet and the integration of the world economy, people have a greater demand for machine translation technology, which fundamentally stimulates the development of machine translation technology.

Emerging technologies such as deep learning have significantly improved the effectiveness of machine translation from a technical perspective. In China, a series of machine translation software, such as Transstar and Yaxin, have also appeared, and these commercial software have reached the level that users can use. The current translation is based on the deep learning method with better effect of machine translation system, mostly based on neural network to realize deeply, such as the Google translation, its kernel is GNMT, is based on the depth of the dozens of layer of neural network to realize, the current worldwide by the consistent high praise, and was heavily used to carry out daily translation, the recognition accuracy in all kinds of language is more than 90%. However, the translation effect of this method for specific sentence patterns is not very satisfactory. This paper attempts to improve the translation effect of this system for specific sentence patterns by adding knowledge of specific sentence patterns on the basis of Google translation.

1.2 Obstacles to Machine Translation

Although machine translation system can replace human translation in specific fields to meet people’s needs, there are still many shortcomings in machine translation system. These problems are often caused by the complexity of language. Taking Chinese as an example, for example, the phenomenon of component incompleteness in Chinese can be divided into ellipsis, implication and empty language (Rong, 1989), which cannot be easily distinguished, and the translation effect of these sentences with component incompleteness in machine translation system is often very unsatisfactory. In addition, there are many more complicated language phenomena in Chinese, such as idioms, xiehouyu and ancient Chinese, which cannot be translated correctly by current machine translation systems.

These special language phenomena are the biggest bottleneck of machine translation system.

In addition, the non-correspondence between languages is a difficult problem to be solved in machine translation. Some typical words that strongly embody Chinese culture, such as “morality and morality”, have no corresponding words in English, or their corresponding words have no similar national meaning (Mao & Gao, 2011). These phenomena are great challenges for machine translation.

1.3 NMT

At present, the mainstream machine translation system is still based on deep neural network technology. This kind of machine translation system is called neural network machine translation system (NMT), and the typical representative is Google translation. Several translation software in China are also popular, such as Youdao dictionary, Jinshan translation master and so on. However, since neural machine translation only USES a single neural network to realize the conversion between natural languages, there are two shortcomings in simplifying the translation process: (1) compared with statistical machine translation, neural machine translation is more sensitive to sentence length; (2) the end-to-end implementation process fails to explicitly utilize linguistic knowledge to achieve better translation performance (Zhang, n.d.).

2. A Concise State Sentence Formed by a Quantifier Phrase

2.1 State Sentences in Concept-level Networks

The theory of concept-level network is a theoretical system for the understanding and processing of natural language, founded by Huang Zengyang, institute of acoustics, Chinese academy of sciences. HNC theory by the space of natural language, language concept space concept Lenovo grain as the main line, to establish a kind of new mode of processing natural language and expression, this model not only can be applied to computer understand natural language processing, such as machine translation, intelligent retrieval, automatic abstract, man-machine dialogue, and other fields, and many other aspects of the research of natural language, such as the study of language ontology, the study of language cognition are rich.

HNC designs three semantic networks for abstract concepts. These three semantic networks are the three clusters of abstract concepts and the basic classification of abstract concepts. They are basic concepts, primitive concepts and logical concepts. Semantic network has a hierarchical structure, each layer has a number of nodes, called concept nodes.

Compared with structural semantics, which describes the relationship between linguistic components from the perspective of aggregation and combination, HNC theory USES sentence classes and sentence patterns to represent the relationship between linguistic components. Among them, the other class is the sentence semantic type, the division is based on “chain effect”, namely the main body of the concept of primitive: six first-level node function, process, transfer and effect, relationship, status, they are expressed the basic side of things, natural language is always a sentence expression effect chain of one or more links.

The concept of concise state sentence in this paper refers to one of the seven basic sentence classes in the concept-level network according to the general theory of sentence knowledge (Miao, 1999)]. In HNC theory, state, as a link of generalized effect chain, supports a special class of sentences, namely state sentences. In addition to the state itself, only one essential element remains in the description of the state, and that is the embodiment of the state. In addition, the content of the state can be presented separately in the sentence. Therefore, the definition of a state sentence in HNC theory is a sentence composed of “state expresser”, “state” and “content of state”. Such as “Xiao Ming is ill”, “Zhang SAN lives a leisurely life. And so on are state sentences.

A concise state sentence is a special type of sentence in a state sentence, which is characterized by no feature semantic block, that is to say, it is a special state sentence without “state” component, consisting only of “state expresser” and “content of state”. For example: “miss li is very beautiful” and “it’s a very simple question”, “Her boyfriend is 1.8 meters tall”, etc. are simple state sentences.

Through simple analysis, it can be found that there are frequent omissions in concise state sentences, such as “her boyfriend is 1.8 meters tall”. The word “height” is omitted from the sentence. Although this expression is often used in daily life, the key word “height” is omitted, making it difficult for machine translation to translate correctly. “Her boyfriend is one meter eight” is obviously wrong, so translating a simple state sentence is still a challenge for existing neural network-based machine translation systems.

2.2 A Concise State Sentence Consisting of Quantitative Phrases

Quantifier phrases are made up of simple state sentences that are more special. “the embodiment of state” and “the content of state” are both quantifier phrases. For example, “a kilo of cabbage costs 50 cents”, “three people in a room”, “one family has one cow”. This kind of sentence pattern, hereinafter referred to as quantifier concise state sentence.

This kind of quantifier concise state sentence first inherits the ellipsis characteristics of concise state sentence, which has increased the translation difficulty of machine translation system. Secondly, quantifier concise state sentence can switch the former and the latter two components and still express the same meaning. That is to say, “a kilo of cabbage costs 50 cents” and “fifty cents a kilo of cabbage”. Either the result is the same or the meaning is the same. This is also a great difficulty for the neural network machine translation system trained by a large amount of data.

3. A Machine Translation System for Quantifier Concise State Sentences

3.1 Opportunities Brought by Big Data

In order to solve the problems raised above, there are two ways to translate concise state sentences of quantifiers. One is to construct an HNC translation system, that is, to analyze the common use of HNC in Chinese and English, and to translate through conceptual correspondence. This method can fundamentally make the translation system "understand" the language. Second, by analyzing Chinese, complete the omitted parts and then translate the complete sentences.

To construct a system of HNC translation methods can fundamentally solve the translation problems, but for a particular language HNC network construction is complex, need a lot of experts and scholars pay great efforts in the field of a particular language, on the basis of HNC network is built, is likely to achieve the HNC correspondence between the different language, thus implementation language translation.

Through the method of Chinese analysis and completion, only the analysis of Chinese is needed, and the resulting sentences can be applied to various languages to be translated, which is more extensive. In addition, with the advent of the era of big data, there are more and more Chinese phenomena in web pages, so the capacity of various existing corpus resources is rapidly increased. Finding omitted elements in corpus resources can reduce the step of manual analysis and greatly improve the efficiency of machine translation system.

3.2 Build a Thesaurus—Assisted Machine Translation System

This paper attempts to build a prototype of a new machine translation system for quantifier concise state sentences on the basis of Google's web page translation and the existing word collocation corpus. The system framework is as follows:

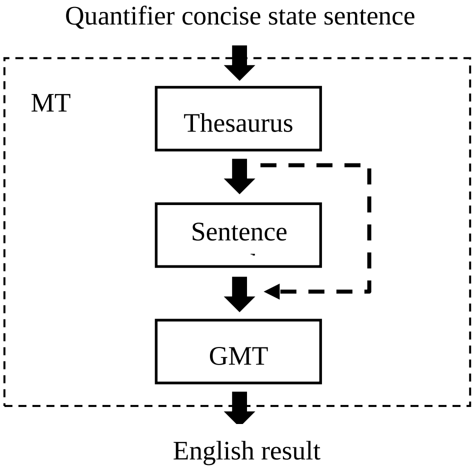

Figure 1. System Framework

The whole machine translation system consists of three parts: the first part finds the collocation relationship and collocation words between the two quantifier phrases and noun components of a concise state sentence by searching the collocation thesaurus, and extracts all the corresponding words of the collocation words. The second part of the sentence adjustment, combined with the existing Chinese corpus, select the most appropriate collocation words and the sentence to be translated to form a clear meaning of the sentence to be translated; In the last part, the complete sentence to be translated is input into the neural network-based machine translation system built by Google (hereinafter referred to as GMT) to get a relatively correct translation result. In theory, this method can improve the accuracy of the translation results, compared with the direct use of Google translation, can improve the translation effect.

3.3 Experimental Data and Process

Condition of experiments, to translate simple quantifier phrase sentence through web crawler access, through to the domestic several big news website in Chinese corpus to crawl, and then use existing software to word segmentation and part-of-speech tagging, every sentence of “quantifier + noun + measure word + noun” to find the part of speech of the structure of the sentence, and then the artificial to filter results, lastly, the 1000 state of quantifier concise words. First, the 1000 sentences obtained were directly input into GMT for translation, and the experimental results were used as a control.

After the experimental data are prepared, the “lexicon assisted GMT” translation system proposed in this paper is used for the experiment: after the concise state sentence of quantifier to be translated, the two noun components are put forward and searched in the collocation thesaurus. Collocation thesaurus is constructed through web crawlers. By crawling the Chinese corpus of several major news websites in China, and then searching the part of speech structure of “noun + collocation words part of speech + noun” through word segmentation and part of speech tagging, the collocation thesaurus is obtained by storing the found results. If there is no such noun collocation in the collocation thesaurus, no processing of the sentence is carried out. If there is only one collocation word in the collocation thesaurus, then fill the collocation word directly into the sentence to be translated, skip the second step, and input directly into GMT for translation to get the translation result. If multiple collocations are returned, the collocations need to be selected. At this point, it is necessary to use the news corpus obtained before again to establish the language model, and then score the language probability model for the complete sentence to be translated composed of different collocate words, return the sentence with the highest score and input it into GMT to obtain the translation result.

3.4 Laboratory Finding

The effect of translating quantifier concise state sentences directly through GMT is not ideal. The effect of translating through “GMT assisted by thesaurus” has been increased by more than five times, to nearly 80%. The specific experimental results are shown in the following table:

Table 1. Experimental Results

	GMT	GMT assisted by thesaurus
Accurate translation	153	761
Inaccurate translation	847	239
precision rate	15.3%	76.1%

In spite of this, the GMT assisted by thesaurus is still insufficient. In the case of inaccurate sentences, unprocessed sentences account for a large proportion. This problem should be solved by expanding the corpus and optimizing the search process. The inaccuracy of sentence translation after processing may be due to the small scale of language model. In addition, there is a very obvious disadvantage of the lexicon assisted GMT translation system, that is, its translation accuracy is directly related to the translation accuracy of GMT. Only when the translation accuracy of GMT is high enough, can the lexicon assisted GMT translation system significantly improve the effect of the actual translation application, otherwise it is useless.

In addition, with the introduction of HNC related knowledge to further improve the effect of translation is quite doable, since state concise sentence is judging by HNC, then attempts through to the two languages are HNC mode of syntax and grammar analysis, through the HNC mode matching to further weaken the wrong translation options, strengthen the correct translation options, boost for HNC concise status according to the effect of translate-on.

4. Conclusion

Through exploration and research, this paper constructs a “thesaurus assisted machine translation prototype system based on Google web page translation”, and verifies through experiments that this system is of significant help in solving the translation problem of “concise state sentences composed of quantifiers and phrases in HNC network”, and the translation effect is significantly improved. This method has good portability and is suitable for any fixed sentence pattern, and can be used to improve the existing machine translation system to some extent. In addition to build the system, this paper also puts forward a feasible scheme, namely the construction of HNC machine translation system, this system is successfully built, theoretically translation effect would exceed the mainstream of machine translation system (NMT) neural network, it still needs to be the natural language processing (NLP) related to the unremitting efforts of scholars, believe in the near future must have a higher accuracy of the translation system.

References

Mao Meilan, & Gao Jiazheng. (2011). On Chinese-English translation of Chinese and common words. *Shanghai translation*, *1*, 52-55.

Miao Chuanjiang. (1999). *HNC theory [C]--Scientific and technological progress and social and economic development for the 21st century*.

News translation from China to Britain is comparable to human, and Microsoft machine translation has made a new breakthrough. (n.d.).

Rong Jing. (1989). Classification of ellipsis, implication and empty language in Chinese. *Journal of Xinjiang University (Chinese version of philosophy humanities and social sciences)*, *4*, 81-87.

Zhang Xueqiang. (n.d.). *Research on machine translation technology based on deep learning*.