A Diachronic Assessment of Research on Machine Translation Methodology

Zhao Ran¹a

¹Tianjin Maritime College, No. 8, Yashen Road, Jinnan District, Tianjin, China
a zhaoran1101@163.com

Abstract: Approaches of machine translation are streamlined into three types based on strategies of knowledge processing, i.e. syntax-based methods, rule-based methods, and corpus-based methods, with their merits and demerits assessed. Consequently, the updated research tendency of the neural networks, as the current mainstream methods of machine translation, mainly comprising computational complexity reducing, words alignment enhancing, prior knowledge and constraints incorporating are stressed. Eventually, the future development orientations of the mainstream methods of neutral machine translation, particularly that of networks integration, data parallelizing and training are prospected.

1. Introduction

Machine translation is to implement the means of computers to transform the text in one language into that of another language. In the 1950s, translation methods based on manual rules was prevalent. Until 1980s, IBM applied statistical methods on the basis of large-scaled corpus, which completely converted the tradition of manual translation rules. At the beginning of this century, based on the massive data on the Internet, Google has developed a statistical machine translation system with multiple language pairs. The translation results are in the leading position in the authoritative evaluation, marking that the statistical method has gradually replaced the rule-based approach, and ushered in the era of machine translation. As an online translation mechanism, Google translation demonstrates the potential value of machine translation in the information age.

Approaches of machine translation have experienced many iterations, each of which has its own merits and demerits. This paper summarizes the main thread and foundation of MT and also analyzes its application and foreseeable prospects.

2. Diversities of machine translation methodology

2.1. Syntax-based machine translation

The model of machine translation based on syntax is founded on the syntactic analysis of source language and target language. Different from other machine translation models, syntax-based machine translation models rely more on deep syntactic analysis of sentences. The process of translation is the transformation of the syntactic tree which reflects the internal structure of a sentence. Kenji Yamada (2001) put forward three translation methods: sequence adjustment, insertion and translation. During the translation process, iterative method is exerted to estimate the parameters of these three operations. The specific steps are as follows:

a. Initialize probability tables of all the three operations of sequence adjustment, insertion and translation:
b. Reset the operation counter;

c. Count the times of three operations based on each sentence pair \(< e, f >\) of the source language in the training corpus,

d. Calculate the probability of three operations

e. Repeat step b and step d until the probability no longer changes significantly.

Statistical machine translation is a combination of statistical machine learning and corpus technology, which has attracted wide attention from all aspects since it is launched and signifies that the technology of machine translation has changed from the rule-based method based on rationalism to the statistical method based on empiricism.

2.2. Rule-based machine translation

Rule-based machine translation system can be generally divided into three types: literal translation system, conversion system and intermediate language system. Their differences lie in the depth of the analysis of the source language, and they all demand a large-scale bilingual dictionary, a large number of derivation rules of the source language, language conversion rules and generation rules of the target language. These three translation systems are briefly introduced as follows.

2.2.1. Literal translation system

The design of this translation system is constructed on the corresponding relationship of lexical modules between target language and source language; therefore, this kind of translation system has the disadvantage of being too much focalized and targeted.

For instance, the English sentence "He likes Lynn" is translated into "Him(He) guata (likes) Lynn" in the English-Spanish direct translation system, while the correct Spanish sentence should be "Lynn him guata". In general, the word sequence of the source language and the target language is differentiated which leads to high probability of low-leveled readability of the sentences translated by the direct translation system, if the knowledge of syntactical structure of the target language is lacked.

Some of the literal translation system also takes into account part of the sentence characteristics and syntactical rules of the target language, and consequently improves the readability of the translation in target language by supplementing certain rules. For instance, the following mapping rules can be added to the English-Spanish translation system mentioned above:

\[ X \text{ LIKE } Y \rightarrow Y \text{ X GUSTAR} \]

With the supplement, the readability of the translation is significantly developed. However, this adapted system displays the characteristics of a conversion system.

2.2.2. Conversion system

The construction of conversion system requires bilingual comparison and complex mapping rules, which is the major difference between conversion system and literal translation system.

In the translation system, the analysis of the source language is independent of the target language, and this process is merely on the syntactic level. A bilingual dictionary is a necessity for bilingual conversion, and the structural differences between the source language and the target language should be taken into account as well.

2.2.3. Intermediate language system:

In the intermediate language system, a direct relationship between the source language sentence and the target language sentence does not occur, while the text of the source language is expressed in the first place with the assistance of the unambiguous intermediate language sentence (which is designed manually), and then the meaning represented by the artificially designed intermediate language is reconstructed and manifested with the vocabulary and syntax structure of the target language, which add an auxiliary languages. At present, there are two types of intermediate language system designed: one is knowledge-based system (such as Margarete Master's machine translation system, KBMT of
Yale University, KBMT of Colgate University), the other is non-knowledge-based intermediate language system (such as CETA, DLT and Roaetta).

Among the three machine translation systems above, the conversion system demonstrates more efficiency which accounts for the fact that most of the highly recognized rule-based MT systems are conversion systems. For example, the MT system of SailLabs belongs to this type.

The advantages of rule-based machine translation: detailed rules can accurately and intuitively describe the grammatical structure of a language, and contain multiple language aspects and levels. As long as the rules can be recognized by the computer in the process of translation, the results will be guaranteed.

Disadvantages: the establishment of rule-based demands large amount of manual labor and material resources. Even though resources are largely consumed, the completeness of rules cannot be ensured and linguistic phenomena cannot be all included.

2.3. Corpus-based machine translation system

There are two kinds of corpus-based MT systems: statistical MT and case-based MT, both of which use corpus as the source of translating knowledge, consequently both of them can be called corpus-based MT. The differences between the two approaches are as follows:

In statistical machine translation, statistical data replace corpus as the representation of knowledge; before translation, the acquisition of knowledge has been completed, and the corpus is no longer used in the whole process.

In case-based machine translation, bilingual case base itself is a kind of expressing form of translation knowledge (not necessarily the only one), which does not need to be exerted as a representation of knowledge; before translation, the acquisition of translation knowledge has not been completed thoroughly, and the corpus still needs to be queried and utilized in the process of translation.

2.3.1. Machine translation system based on statistics

In the mid-1950s, Weaver proposed in his memorandum the strategy of "decoding" (a statistical method) for machine translation, but it was not implemented due to the limitations of the relevant study at that time. On the basis of weaver’s thought, Peter Brown of IBM and other scholars proposed a mathematical model based on statistical machine translation, namely, the noise channel model.

The model can be illustrated as that a language S changes after a noisy channel, which is displayed by another language T at the output end of the channel. The essence of translation is how to restore the source language S closest to the original meaning according to the observed language T. Language T is the output in the channel model as well as the source language in the translation process. On the contrary, language S is the input in the channel model, but the target language in the translation process. From this point of view, any sentence in one language may be the translation of several sentences in another language, but the possibility of occurrence of these sentences may vary. The purpose of searching in the translation process is to find out the best possible alternative, which is to calculate the most probable one for all possible target language S as the translation of source language T. Expressed by the formula:

$$T=\arg \max P(T|S)=\arg \max P(T)P(S|T)$$

Fig 1: calculation of the best possible alternative from S to T

In the formula above, S stands for the source language, T for the optimal translation sought out, P(T) for the language model of the target language, P(S / T) for the translation model, and arg max represents the process of selecting the optimal translation.

![Diagram](image-url)
At present, the most well-known statistical MT is the Candide system of IBM. Although statistical machine translation has made some achievement in the field of speech recognition, it still needs further advancement in several aspects including a large amount of bilingual case base, data provision, aligned knowledge base, and a better search algorithm to find the optimal translation.

2.3.2. Case-based machine translation system

In traditional MT, knowledge is translated into code manually with the guidance of certain principles, which is highly time-consuming and laborious. In order to overcome this difficulty, Nagao, a Japanese machine translation expert, put forward a new method in the 1980s, using the existing translation examples (bilingual texts) as the knowledge source. This new strategy is called Analogy-based MT, which is the well-recognized Example-based MT (EBMT) afterwards.

In case-based machine translation system, bilingual corpus is the main source of knowledge base. There are two major fields in the knowledge base. One is to save the sentences in the source language, while the other is to save the corresponding translation. When a user enters a sentence, the translation system will compare the target sentence with the existing source language sentence field in the case base to seek the most similar source language sentence with the target sentence, and adjust it by referring to the corresponding translation of the source language sentence, and finally output the translation. Its basic structure, proposed by Federica, and components are illustrated in the following figure.

Fig 3: case-based MT structure

The scale and coverage of the corpus determine the quality of the translation of EBMT system to a great extent. Therefore, many scholars are committed to the research of bilingual corpus, such as Gale and Kay. In addition, Satoshi also proposes a machine translation method that combines rules with examples, and achieves satisfactory translation results.

The advantages of case-based machine translation system: it can mark the translation results with the reliability factor; the speed of translation processing can be improved by parallel processing; the robust performance is relatively better by utilizing the best matching reasoning; it can be shared with other systems which contributes to its ductility.

Disadvantages: in the process of bilingual automatic alignment, if accurately finding similar translations is needed, not only alignment at sentence level is needed at some occasion, but also alignment at phrase level or even vocabulary level, which is difficult to achieve. Moreover, there are many ambiguities in phrase-leveled alignment, which decrease the accuracy and quality of translation. When the sentences sought out for translation are used, the analysis of the structure of the source language is not sufficient, resulting in the loss of part of the original information.

3. An analysis of the frontier trend of machine translation

In recent years, with the rapid development of artificial intelligence, neural machine translation (NMT) research has made large pace of advancement. The performance of neural network machine translation
in many languages has been greatly developed, far surpassing the traditional statistical machine translation (SMT). At present, NMT research is a hit frontier of natural language processing research.

3.1. The basic idea of neural machine translation
The idea of applying neural network to machine translation has been advocated for a long time in academic field. In the 1990s, Castano accomplished the translation based on neural machine system by utilizing small-scale parallel corpus. However, due to the limitation of parallel corpus scale and hardware computing competence, his group failed to achieve transcendental effect. With the booming of deep learning, neural network is generally employed with statistical machine translation in conducting word alignment, dependency parsing, rule extraction and other tasks.

The fundamental conception of neural machine translation bear much resemblance with that of statistical machine translation, which is probability maximization. In the modeling of translation, only neural network is employed to conduct the transformation from source language to target language. Different from the discrete representation of statistical machine translation, neural machine translation applies continuous space representation to convey words, phrases and sentences. In terms of translation modeling, the processing steps of statistical machine translation, such as word alignment, phrase extraction and phrase probability calculation, are not demanded, but neural network is employed to complete the mapping from source language to target language. In neural machine translation, encoder-decoder framework is normally used to construct the transformation from source sequence to target sequence. Read the source language and input "X1", "x2", "X3", "X4", and output the fixed dimension semantic encoding vector C; the decoder reads the vector, decodes and generates the target language word sequence "Y1", "Y2", "Y3", as shown in the following figure:

3.2. Convolutional neural network
Convolutional neural networks (CNN) has achieved great success in image classification. When processing images, pixels are practically used as input. Unlike image tasks, input is basically sentences or documents in matrix form. Each row of a matrix is a vector, which is essentially a serialized input. Convolutional neural networks are usually utilized in text classification. For instance, in emotion classification, information recognition and topic classification, the convolution operation will lose the information of location of some words, therefore, the effect is not desirable when implementing the mission of serialization input. However, in recent years, a lot of research has retained the parallel ability of convolution neural network by integrating the information of human location, meanwhile enhancing the processing of serialized input.

4. Conclusion and Prospect
At present, neural machine translation has achieved tremendous success, and its impact on large number of language pairs has surpassed statistical machine translation. Since 2014, a great number of scientific research achievements and practical products have been created. Nevertheless, problems still exist in this field of study and the following orientations of research might be further pursued.
4.1. Improvement of the interpretability of translation framework

Neural machine translation based on encoder-decoder structure realizes the direct translation from source language to target language, but compared with statistical machine translation, neural machine translation process bears much resemblance to running in black box, which is difficult to explain the translation process from the linguistic perspective. It has been proved that it is an important research direction of neural machine translation in the future to analyze translation process from the perspectives of visualization and extraction of implicit syntactic structure information, so as to correct translation errors.

4.2. Multilingual Machine Translation

Translation system learning is a universal representation, in which sentences with the same meaning in different languages are expressed in similar ways, which provides a good foundation for the study of transfer learning between multiple languages. Based on multilingual parallel corpora, or multilingual comparable corpora, the study of multilingual machine translation based on neural networks is of academic and practical value to the translation of low-resourced languages and even data-rich languages. The translation of data-rich languages has both academic and practical value which is an important direction of natural language processing.

4.3. Multimodal translation

In the traditional neural machine translation process, the process of text translation is independent of the scenarios of translation. Therefore, the result of neural machine translation is unable to adaptively generate text translation results suitable for translation scenarios. However, image and text information in the same scene belong to heterogeneous information, and there is a huge semantic gap between them. Therefore, integrating the aligned multimodal features into neural machine translation network to achieve multimodal neural machine translation is a direction worthy of exploration to improve the translation effect and even realize intelligent translation.

Neural machine translation represents a brand-new machine translation model. At present, its performance in the mainstream language pairs has surpassed that of statistical machine translation and become the mainstream technology. Neural machine translation can learn features directly from parallel corpora, with low application difficulty, and can effectively deal with long-distance dependence through long-term memory and attention mechanisms. Although it has some disadvantages in the aspects of resource dependence, training algorithm, interpretability etc., this approach will achieve great development in the future.

Acknowledgments

This paper is the result of the research of Tianjin Higher Vocational and Technical Education Association: Research on the Strategy of Implementation of "Moral Education" in International Cruise Talents Cultivation. Project No. 2020-3019.

Reference

[1] Alkhouli T, Ney H. (2017) Biasing attention-based recurrent neural networks using external alignment information. Proc of the 2nd WMT. Copenhagen: ACL, 1: 108-117.
[2] Chen H., Huang S., Chiang D., et al. (2018) Combining character and word information in neural machine translation using a multi-level attention. Proceedings of the North American Chapter of the Association for Computational Linguistics: 1284-1293.
[3] Cho K., Van Merrienboer B., Bahdanau D., et al. (2014) On the properties of neural machine translation: encoder - decoder approaches. Proceedings of SSST - 8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar. pp. 103 – 111.
[4] Chung J., Gulcehre C., Cho K., et al. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.
[5] Ding Y., Liu Y., Luan H. et al. (2017) Visualizing and understanding neural machine translation.
In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1: 1150-1159.

[6] Domhan T., (2018) How much attention do you need? A granular analysis of neural machine translation architectures. In: Proc of the 56th ACL. Melbourne. ACL, pp. 1799-1808.

[7] Federica, Ricardo, Paolo, (2002) Searching Similar Sentences for EMBT, SEBD’02, Italia.

[8] Fishel M., Georgakopoulou Y., Penkale S., Petukhova V., Rojc M., Volk M., Way A., (2012) From subtitles to parallel corpora. In: The 16th Annual Conference of the European Association for Machine Translation. Italy. 28 May 2012 - 30 May 2012: 3-6.

[9] Gehring J., Aull M., Grangier D., et al. (2017) A convolutional encoder model for neural machine translation. ar-Xiv:1611.02344v3.

[10] Gu J., Hassan H., Devlin J., et al. (2018) Universal neural machine translation for extremely low resource languages. ar-Xiv:1802.05368v2.

[11] Katsumata S., Matsumura Y., Yamagishi H., et al. (2018) Graph-based filtering of out-of-vocabulary words for encoder-decoder models. In: Proc of the 56th ACL. Melbourne. ACL, pp. 112-119.