Analysis on the Recent Trends in Machine Translation

Ruimeng Song*

School of economy, Jinan University, Guangzhou, Guangdong, China

*Corresponding author e-mail: 631301060118@mails.cqjtu.edu.cn

Abstract. Machine translation is to translate one language into another language, which has undergone a great evolution. The model of machine translation has been continuously improved, aiming to make the translation effect closer to the artificial translation. This article briefly summarizes the development history of machine translation, and introduces the main models of each stage of development. The initial machine translation mode is the Rule Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). Recent mainstream translation approach enables Neural Machine Translation (NMT). It includes the input and the output, attention mechanism, and BLEU evaluation method. On this basis, there are also many expansion and innovation models, such as GPKD and other models to improve the evaluation effect. In general, machine translation can replace a part of human translation. However, it cannot completely replace human beings, because of the different human thinking and machine logic. People and machines have to cooperate with each other to improve the common efficiency.

1. Introduction

Natural Language Processing (NLP) has a wide range of applications in various fields. Precisely, NLP is the product of the development of computing technology, which can learn and understand human language. In the early years, language research focused on the automatic analysis of language structures. After years of development, various technologies in NLP are also progressing and evolving. NLP's current application directions include: language recognizer, news text mining and classification, sentiment analysis, automatic text generation, user portrait building, chatbots, knowledge graphs, and machine translation. As one of the most crucial applications of NLP, machine translation also has experienced tortuous development.

Machine translation is the translation of one language into another, as shown in Figure 1, the above sentence is marked with Source, and the following is marked with Target, and the machine translation task is to translate the sentences of the source language into the target language sentences.

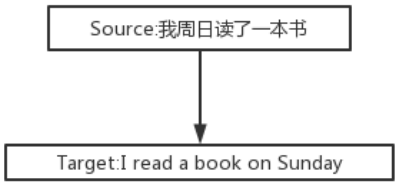


Figure 1. The goal of Machine Translation

In Figure 2, machine translation has gone through three stages of development. Before 2014, the two most important machine translation methods are RBMT and SMT. In 2014, the birth of Neural Machine Translation (NMT) had a great influence on machine translation.

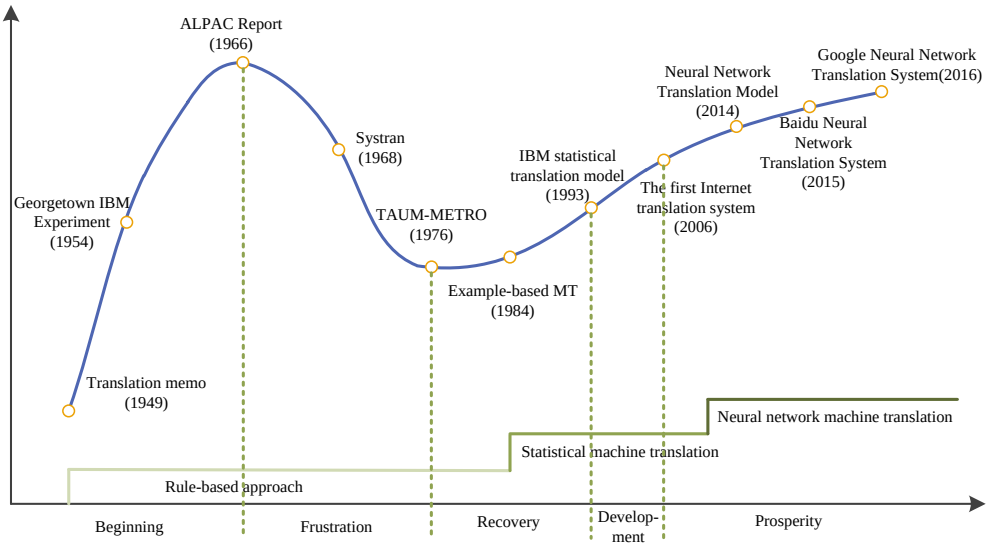


Figure 2. Changes of Machine Translation

2. Rule based machine translation (RBMT)

RMBT analyses text according to language rules created by any exports and then translates it with the help of computer programs. There are many commercial machine translation systems use RMBT due to its simple principle. As shown in Figure 3, RMBT has these three paths:

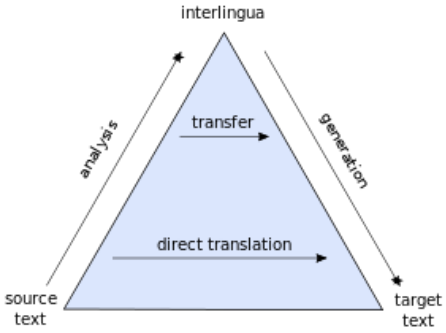


Figure 3. Paths of Rule based machine translation

2.1. Direct Translation

It is the most basic method, the process of implementing machine translation is relatively simple, which can also be comprehended to word-to-word translation.

2.2. Transfer Translation

Compared with Direct Translation, Transfer Translation refers to and takes into account the lexical, syntactic and semantic information of the original text. However, the scope of information sources is too broad, there are too many grammar rules, and there are contradictions and conflicts between them.

2.3. Interlingua translation

In fact, the scope of application of Interlingua is not very wide. Interlingua translation is only a hypothesis. For the developers, they need to learn another language which is only useful in translation progress. When achieving the decoding of language information, Interlingua translation universal form can completely separate from the language that we daily use.

To sum up, the rule-based machine translation is suitable for computer processing and conforms to the logical thinking of computer software programming, The key to the rule approach is to build a

high-quality translation rule base, however, the basic rules are all written by the professional linguists, which may takes a few years and a lot of money.

3. Statistical machine translation(SMT)

SMT generates a translation model (vocabulary, comparison, or linguistic pattern) by statistical analysis of several simultaneous corpora. The model is subsequently utilized during the actual translation process.

Statistical knowledge is generally used in this process, and the translation with the highest probability of occurrence will be selected. The probability algorithms is based on Bayes theorem:

First we name the source sentence as f, the target sentence as e. The probability of the source language can be translated into the target language $p(e|f)$, according to the Bayes theorem in Formula 1:

$$p(e|f) = \frac{p(f|e)p(e)}{p(e)} \tag{1}$$

Our purpose is to find the maximum value of $p(e|f)$. The source language is determined, and $p(f)$ is a constant value.

In SMT model, $p(e)$ is a language model, a priori probability that represents the probability that a sentence is a fluent sentence. $p(f|e)$ is a translation model, a conditional probability, that represents the probability that a sentence will be translated from e to f. Our main task is to train the model so that it knows the values of the $p(e)$ and $p(f|e)$.

Compared to the RMBT, SMT does not require manual intervention. It is completely automatic and takes less time. It also does not require a lot of money. The translation construction of machine translation is usually scanned from left to right for translation. The translation of long sentences is more difficult, and the effect is not as good as we expect. This makes statistical machine translation enable to translate most words and phrases, even though the sentence is not smooth and at the same time the missing translation may happen.

4. Neural machine translation(NMT)

4.1. Classical model

The basic modeling framework for NMT is a framework and method for transforming input sequences into output sequences, the architecture is an encoder and a decoder.

As early as the 1990s, the theory of neural machine translation had already emerged. Forcada and Ñeco used the "encoder-decoder" structure for translation conversion in 1997.

In 2003, Bengio first used neural networks to build language models, which became the beginning of neural networks in NLP.

A word list can be defined as Formula 2 [1]:

$$W = \{w_1, w_2, \dots, w_t\}, w_t \in V \tag{2}$$

Our goal is to find a good model through the function as Formula 3 [1]:

$$f(w_t, \dots, w_{t-n-1}) = \hat{P}(w_t|w_1^{t-1}) \tag{3}$$

It should satisfy the limitation as Formula 4 [1]:

$$\sum_{i=1}^{|V|} f(i, w_{t-1}, \dots, w_{t-n-1}) = 1 \tag{4}$$

The training process of neural machine translation is the process of maximizing the value of θ as Formula 5 [1]:

$$L = \frac{1}{T}\sum_t \log f(w_t, w_{t-1}, \dots, w_{t-n-1}; \theta) + R(\theta), \ \theta \text{ is the overall parameter set.} \tag{5}$$

In 2013, Recurrent Continuous Translation Models are a family of probabilistic continuous translation models that Kalchbrenner and Blunsom presented [2]. In this model, CNN first use the contents of the source text to convert them into some contiguous vectors, and then translate the state vectors into the intended language using the RNN as a decoder. Their work can be regarded as the start of NMT.

In 2014, a unique neural network model with two RNNs dubbed RNN Encoder-Decoder was introduced by Cho et al [3].The learning method proposed by Sutskever et al. is end-to-end sequences, which first assume that the sequence structure is minimal. This approach is based on LSTM, where both encoder and decoder are RNN structures. This architecture has also been launched into Google Translate, which still has a certain impact. Another major step is forward in the field of NMT, and the quality of translation of this translation method can be somewhat beyond that of humans. This is an improvement over the translation model and process based on the previous one.

Based on the above research, Bahdanau et al. introduced attention mechanisms that significantly improved the translation performance of NMT models. Better correction of the logarithmic probability of translation requires joint training of all parts of the translation system, including alignment mechanisms [4]. More and more people have paid attention to the simple structure and remarkable performance of attention mechanism, and researchers around the world have developed more models based on this.

In 2017, the birth of the paper "Attention is All You Need" greatly promoted the development of machine translation. Attention mechanism is a new simple network architecture. Different from traditional CNN and RNN, the entire network structure is attention mechanism, which has nothing to do with other models. Transformer's key benefits are its multi-head and self-conservation, as a result, it may operate in parallel, substantially accelerating the training process.

In 2018, BERT, a fresh model for language representation was proposed, which is innovative in combining and modulating all text content at all layers and pre-training deep bidirectional representation in unmarked text.

4.2. Encoder-Decoder

The encoder-decoder framework is a typical representation-based model. "Represents" is the information entered into text. The role of the encoder is to transform the input sequence of words into a new form through some kind of transformation, which can be stored in the machine and understood by the machine. We take this form as text feature, and then the decoder converts this "representation" back into a sequence of words in the output.

The two networks are first combined and trained to derive the target sequence of a given source sequence in the conditional probability maximization case. The text feature representation for tasks involving natural language processing is the embedding word embedding. Text feature representation is the most core and basic part of the natural language understanding process.

In practical applications, the encoding and decoding process is shown in the Figure 4. <eos>(End of Sequence) indicates the termination of the sequence, and <sos>(Start of Sequence) indicates the beginning of the sequence.

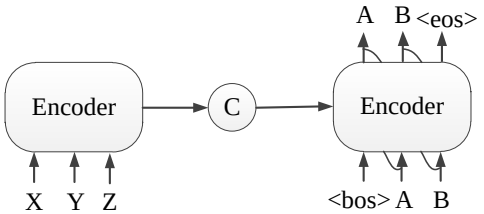


Figure 4. Process of encoding and decoding

Traditional text features represent the use of one-hot encoding as the earliest word vector. When word appears, the word is labeled as 1 and the other words are marked as 0, and its sparsity brings about a "dimensional disaster". Therefore, some researchers have proposed word vector models based

on the distributed representations. Based on distributed assumptions, such as CBOW, Skip-Gram, Word2vec, and GloVe, the words have similar meanings.

4.3. Attention

Before the attention mechanism was developed, as the sequence of the source language became longer, encoding the information of the entire sentence into a vector of a fixed dimension could cause the loss of sentence information, which would reduce the accuracy of the translation and not get the results we wanted. Moreover, the importance of the information in different parts of a sentence is also different, and some of the information in the sentence will be strongly linked. For example, we can easily associate "drinking water" from "thirsty", however, it is difficult for us to associate "Sunday" and the other weak information.

The attention mechanism is discovered and developed because of these problems, and its core is to generate the different text feature vectors for different target language words. We can understand the attention mechanism as a weighted processing of text feature information. While we give more information more attention, we give less weight to less contributed information, weakening its impact on the outcome. This simplifies the translation process to some extent, also increases efficiency and makes it easier for us to notice the important information.

There are three types of attention mechanisms, self-attention, soft attention and hard attention. The self-attention mechanism of the NLP field is used to do text processing. Soft attention is more focused on an area or channel, which is a definitive attention. Strong attention is more of a focus, while at the same time a random prediction process, with more emphasis on dynamic changes.

4.4. Evaluation mechanism

At first, machine translation has a way to manually evaluate the effect of translation. Manual evaluation is generally scored anonymously by multiple reviewers, and then the evaluation results of all evaluators were integrated for the final score. Manual evaluation can accurately reflect the translation quality of sentences, and is the most authoritative and credible evaluation method, however, it consumes a lot of time and money, which is an irreparable disadvantage.

4.4.1 Bilingual Evaluation Understudy (BLEU)

BLEU is an indicator used to assess the difference between the "candidate" generated by the model and the actual sentence. In the translation process, if the "candidate" and "reference" are the same, then the BLEU value is 1, which is our best expectation. If they are completely different, then the BLEU value is 0, which is not what we want to see, indicating that there are some problems in the translation process.

The BLEU value was proposed in 2002 by IBM scientist Kishore Papineni in his paper. The Automatic Evaluation of Machine Translation approach is known as BLEU. BLEU is fast, inexpensive, and language-independent, highly correlated with manual evaluation, with little marginal value per run [5].

Almost all papers use this indicator to evaluate the effectiveness of translation. It has almost become an official indicator, and some people even have a high BLEU value as the only requirement.

A simple n-gram evaluation is prone to some problems and sentences with high precision, however, the completely incorrect meaning can occur.

The modified n-gram precision is proposed in this paper:

It is possible to count the instances of the current word in each translation used as a reference. We can get Formula 6:

$$C = \min(T, S) \tag{6}$$

C represents the number of successful matching for the current word.T represents the maximum number of occurrences per translation. The candidate sentence's occurrence count is shown by S.

Finally, dividing the sum of the number of successful matches for each word by the length of the candidate sentence is the modified 1-gram evaluation.

When we generalize 1-gram to the n-gram direction, as n-grams getting larger, the decrease in accuracy is not linear but exponential.

Brevity Penalty (BP) can be computed in Formula 5 [5]:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (7)$$

Where c and r both stand for length, c stands for candidate translation, and r stands for valid reference corpus.

BLEU can be calculated based on the BP value [5]:

$$BLEU: BP\exp(\sum_{n=1}^{N} w_n \log p_n) \quad (8)$$

4.4.2 METEOR standard

The METEOR standard was proposed in 2004 after Lavir discovered the significance of recall in the evaluation indicators. Meteor Universal automatically extracts language resources from dual text trained on machine translation systems, and it can learn manual judgment results in multiple language directions and turn them into a generic set of parameters to assess translation quality. It also can provide the evaluation for the previously unsupported target languages [6]. Measuring accuracy and recall based on the complete corpus, METEOR, unlike BLEU, produces a measure. It may be more accurate than BLEU in some respects, however, in terms of synthesis, BLEU is still the most influential and widely used assessment

5. Current Model

There are many new ways to improve the previous translation model in different ways.

In Synchronous interactive multilingual neural machine translation (SimNMT) [7], each target uses historical and future information from all target languages to predict the production of the language. The experiment concluded in tests that the method achieved significant improvements compared to the highly competitive bilingual NMT model and the multilingual NMT model.

There are mainly two functions for Partial Vector Quantization (P-VQ), one is to speed up the prediction of the word probability in the softmax layer, and another is to compress the word embedding matrix. It is revealed in experiments of Chinese to-English translation that the suggested approach can minimize 74.35% of the word embedding parameters and 74.42% of the softmax layer FLOPs. On the WMT test, the average BLEU score only drops by 0.04 concurrently [8].

The quantity of sub-layer in the decoder of Compressed Attention Network is only one but not three. It is revealed that the speed of the model is promoted by 1.42 times through the widely experiments on tasks of 14 WMT machine translation. As a result, the performance of the model plays at the nearly same level as a strong baseline. At the same time, it is already two times swifter than the standard baseline which is extensively applied but no loss in performance [9].

The method of Group-Permutation based Knowledge Distillation (GPKD) transforms the deep Transformer model into a shallower one by compressing it. Moreover, it performs substantially better than the method of SKD. The approach of Skipping Sub-Layer which is proposed suppresses the problem of overfitting through the means of ignoring sub-layers randomly during in the training stage for the system of highly deep encoder [10].

As a powerful many-to-English NMT model, with only six source languages used for training, SixT+ is accessible in 100 source languages. With XLM-R large, the decoder embedding and full encoder are initialized by SixT+, the process of training of the encoder and decoder layers consists of strategy with two stages[11].

6. Conclusion

Optimistically, under many specific conditions, the translation results of machine translation are great, and they are even close to the results of artificial translation. No matter how machine translation

evolves and how many good models are created, it cannot completely replace human translation, because the goal of translation is determined by people. Also, as technology advances, so will people's needs for machine translation, In addition to communicating the text's meaning, it can also express feelings. For human translation, we need Faithfulness, Expressiveness, and Elegance. Faithfulness means saying exactly what the original means, without bias. Expressiveness means fluency in translation and the ability to paraphrase the text in an incomplete way. Elegance means that the translation should be catchy, with beautiful sentences.

One can achieve these three goals by translating after thinking, so the accuracy rate of human translation can be close to 100%.However, it also depends on the level of translators, the level of original expression, the industry field, the delivery time and other factors. The accuracy of machine translation depends on factors such as language, industry domain, quality of the original text, training corpus, training model, and etc.

Machine translation liberates experts from some mechanical, so that they can engage in more creative work, such as bilingual writing, etc.. In some cases, machine translation is far less good than one would expect. For example, in machine translation and high-precision simultaneous interpretation tasks, machine translation still needs more polish. There is no doubt that machine translation can help humans, and it can even one day replace some low-end human translation work.

Machine translation systems focus on the single goal of translation, rather than the complex tasks we mentioned above. Moreover, people and computers operate substantially and differently. The biological abilities we use in our translation process are completely incomprehensible to machines, and it is unlikely that they will be used by machines in the future. Machines use knowledge of their own understanding, such as a statistical word representation. This knowledge does not need to be understood by people, which doesn't seem to be thought in the process of our translation. From the perspective of system development, computers do not need to understand how people think, because the machine only needs to follow the program settings to complete the language conversion task step by step.

Human translation and machine translation have their own advantages and disadvantages, and they are not antagonistic. Human and machine working together is the common development trend and direction of human translation and machine translation in the future. Machine translation is fast and efficient, while human translation is accurate and elegant. These two ways may help each other, which is also the only way to improve translation quality and efficiency.

References

[1] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. A neural probabilistic language model. Journal of machine learning research, 3 (Feb): 1137-1155, 2003. Google Scholar Google Scholar Digital Library Digital Library.

[2] Kalchbrenner, N., & Blunsom, P. (2013, October). Recurrent continuous translation models. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1700-1709).

[3] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[4] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[5] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

[6] Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380).

[7] He, H., Wang, Q., Yu, Z., Zhao, Y., Zhang, J., & Zong, C. (2021, May). Synchronous interactive decoding for multilingual neural machine translation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 14, pp. 12981-12988).

[8] Zhang, F., Tu, M., & Yan, J. (2021, May). Accelerating Neural Machine Translation with Partial Word Embedding Compression. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 16, pp. 14356-14364).

[9] Li, Y., Lin, Y., Xiao, T., & Zhu, J. (2021, May). An efficient transformer decoder with compressed sub-layers. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 15, pp. 13315-13323).

[10] Li, B., Wang, Z., Liu, H., Du, Q., Xiao, T., Zhang, C., & Zhu, J. (2021, May). Learning light-weight translation models from deep transformer. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 15, pp. 13217-13225).

[11] Chen, G., Ma, S., Chen, Y., Zhang, D., Pan, J., Wang, W., & Wei, F. (2022, May). Towards Making the Most of Cross-Lingual Transfer for Zero-Shot Neural Machine Translation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 142-157).