ISSN: 1813-4890

Research on Machine Translation based on Neural Network

Bo He[a], Jiaoqiu Shi[b]

School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

[a]hebo@cqut.edu.cn, [b]jq88@stu.cqut.edu.cn

Abstract

Machine translation, is the process of using electronic computers to perform automatic translation from one language to another. Due to the rapid development of next-generation artificial intelligence and neural network technology, neural network-based machine translation technology is changing rapidly and its performance significantly exceeds that of traditional machine translation methods. Neural network-based machine translation can model data sequences with a distributed continuous space representation model that can capture more hidden information, and the process of translation does not rely on the design of any artificial features, and the learning of features is obtained entirely by neural network computation. This paper briefly describes the classical machine translation methods, makes a description of the basic ideas of neural machine translation, and gives a specific introduction to the research on neural network-based machine translation, and finally is a summary of the literature and an outlook on neural machine translation.

Keywords

Machine Translation; Neural Network; Neural Machine Translation.

1. Introduction

Machine translation refers to the process of automatic translation by computer, which is an important research hotspot in the field of artificial intelligence and natural language processing. The realization of machine translation often requires the integration of knowledge from multiple disciplines, such as mathematics, linguistics, computer science, psychology, etc. The lack of any aspect of effort cannot achieve breakthrough results, so machine translation is also a challenging task in natural language processing. The research exploration of machine translation can be traced back to the birth stage of electronic computers in the 1940s, and Warren Weaver [1], who is known as the pioneer of machine translation, put forward the first influential proposal of machine translation in 1949, marking the formal introduction of the idea of machine translation. Since the creation of machine translation tasks in the 1940s, machine translation has gone through two phases: rule-based machine translation and statistical machine translation. Since 2014, with the development of machine learning techniques, deep learning-based Neural Machine Translation (NMT) has been gradually developed, and it has already achieved significant advantages on most tasks in just a few years. Neural machine translation is the modelling of the entire translation process using neuronal networks. This approach enables end-to-end learning without making any implicit structural assumptions about the text and without relying on artificially defined features; all translation models are implemented by training under an end-to-end model, and the entire translation decoding process is the act of forward operations or inference on the neural network. The current research capability of neural machine translation is significantly higher than previous means of machine translation research, and it is not only the preferred research method for machine

ISSN: 1813-4890

translation research, but also has become the core of major commercial machine translation research techniques.

2. Classical Machine Translation Methods

There are two classical approaches to machine translation: rule-based machine translation and statistical machine translation. The first generation of machine translation technology mainly uses rule-based machine translation methods, whose main idea is to introduce linguistic knowledge in the source and target languages through rules defined by formal grammars. The rule-based machine translation method is highly dependent on language laws, and although it has a certain degree of generality, the cost of obtaining rules is high, the quality of language rules is too dependent on the experience and knowledge of linguists, the maintenance of rules, and the compatibility of old and new rules are all bottlenecks that are difficult to break through. In response to the problems of the rule-based approach, instance-based machine translation was proposed in the mid-1980s [2]. The basic idea of this method is to find an example in the double utterance library that is similar to the sentence to be translated, and then modify the translation of the example, such as replacing, adding, deleting and a series of operations on the translation, so as to get the final translation. However, this method requires very high precision in translating instances, and an error in one instance may result in not even one sentence type being translated correctly. Instance maintenance is more difficult, the construction of instance libraries usually requires word-level aligned annotations, and ensuring the quality of word alignment is a very difficult task, which makes it significantly more difficult to maintain instance libraries.

Statistical machine translation started to become the dominant approach to machine translation in the 1990s [3,4]. The statistical-based machine translation method translates by using linguistic knowledge learned from the corpus without human hand-written linguistic rules, and the translation quality depends mainly on the size of the corpus, which is mainly divided into three types based on words, phrases and syntax. Specifically, the translation process is modeled by learning language patterns from monolingual corpus and translation patterns from bilingual multilingual corpus, and by implementing these statistical patterns. Whether words or phrases, even syntactic structures, can be learned automatically by means of statistical machine translation models. The human is more interested in defining the features required for translation and the form of the basic translation units, while the translation knowledge is stored in the parameters of the model.

Since there are no excessive restrictions on the translation process, statistical machine translation has a very flexible way of generating translations, so the system can handle more diverse sentences. However, this approach also poses some problems: first, although there is no need to manually define translation rules or templates, statistical machine translation systems still require manually defined translation features, and improving translation quality often requires extensive feature engineering, which leads to a decisive impact of good or bad manual feature design on the system; Secondly, statistical machine translation has more modules and the system development is more complicated; again, as the training data increases, the model of statistical machine translation (e.g., phrase translation table) will be significantly larger, which consumes more system storage resources.

3. Neural Machine Translation Methods

3.1. Basic Ideas of Neural Machine Translation

In recent years, end-to-end representation learning-based approaches are changing the way we process natural language due to the booming development of deep learning techniques and

ISSN: 1813-4890

their widespread use in various industries, and neural machine translation has emerged from this trend. On the one hand, neural machine translation still continues the idea of statistical modeling and data-driven based, and thus is consistent with previous studies in the definition of the basic problem; on the other hand, neural machine translation departs from the assumption of the implicit translation structure in statistical machine translation, while using distributed representation to model text sequences, which makes it possible to view the translation problem from a completely new perspective. Nowadays, neural machine translation has become a hot spot for machine translation research and application, and the quality of translation has been greatly improved.

Compared to statistical machine translation methods, neural machine translation is based on a continuous space representation model, a distributed continuous space representation model that captures more hidden information. It is an end-to-end model that does not rely on any implicit structural assumptions, and end-to-end learning models the problem more directly; it does not rely on any artificial feature design, or its features are implicit in the distributed representation, these "features" are automatically learned, so the neural machine translation is not limited by artificial thinking, the learned features are more comprehensive description of the problem. The models of neural machine translation are represented by neural networks, and most of the model parameters are real matrices, so the consumption of storage resources is small, and the neural networks can be developed and debugged as a whole, and the cost of system building is relatively low.

3.2. Encoder-decoder Framework

Machine translation is often seen as a transformation of one sequence into another. Neural machine translation implements sequence-to-sequence conversion using an encoder-decoder framework. The role of the encoder is to encode the input text sequence, extract the information from the source sequence for distributed representation, and then the decoder reconverts this information into the output text sequence. As shown in the encoder-decoder architecture in Figure 1, given a Chinese sentence "我对你感到满意", the encoder will encode this sentence to generate a vector representation, i.e., the vector (0.2, -1.6, 5, 0.7, -2) in the figure, and then send this vector to the decoder as input, and the decoder will decode this vector from left to word by word into the target language translation.

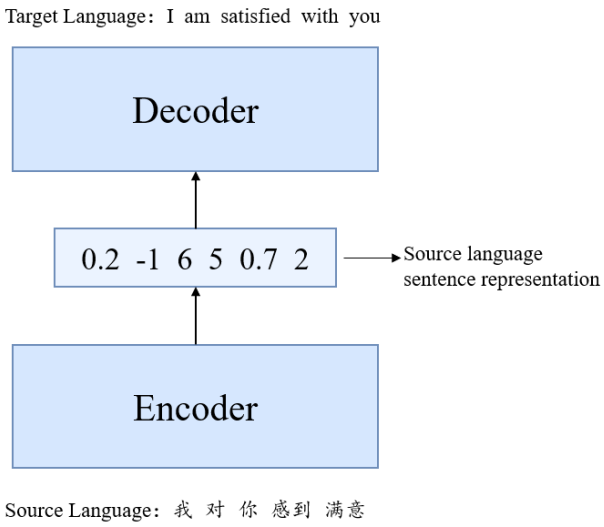

Figure 1. Encoder-decoder architecture diagram

After the representation of the source language sentences is determined, the corresponding encoder and decoder structures need to be designed. In today's mainstream neural machine

translation systems, the encoder consists of a word embedding layer and an intermediate network layer. When a sequence of words is input, the word embedding layer maps each word to a multidimensional real representation space, a process also known as word embedding. The intermediate layer then performs a deeper abstraction of the word embedding vector to obtain an intermediate representation of the input word sequence. There are many ways to implement the middle layer, for example: recurrent neural networks, convolutional neural networks, and self-attentive mechanisms are all common structures used in the model. The structure of the decoder is almost the same as that of the encoder. In the recurrent neural network-based translation model, the decoder only has one more output layer than the encoder, which is used to output the probability of word generation at each target language location, while in the self-attentive mechanism-based translation model, in addition to the output layer, the decoder has one more attention layer than the encoder, which is used to help the model better utilize the source language information.

3.3. Neural Machine Translation Research

Neural machine translation technology first originated from the probabilistic language model of neural networks proposed by Bengio et al. in 2003. Discrete characters are then represented as continuous dense distributed vectors through the use of neural networks, and such distributed vectors effectively alleviate the data sparsity problem [5]. In 2013, the first machine translation model built entirely from neural networks was proposed by Nal Kalchbrenner and Phil Blunsom at the University of Oxford, which used CNNs and RNNs to form an "Encoder-Decoder" structure [6]. The encoder is composed of a convolutional neural network CNN, which can obtain historical information and process variable-length strings, and the decoder is composed of a recurrent neural network RNN, which can directly model the translation probability. While previous studies have used deep neural networks only as an auxiliary method for language modeling, their study consists entirely of deep neural networks, marking the independent use of deep learning methods in the subject of machine translation. However, the implementation of this work is more complex and the method suffers from problems such as gradient disappearance/explosion. In 2014, Sutskever et al. of Google team proposed sequence-to-sequence (seq2seq) learning while introducing long-short term memory structure (LSTM) to neural machine translation, an approach that alleviates the problem of gradient disappearance/explosion and allows the network to selectively remember information through the design of forgetting gates, alleviating the problem of distance dependence in long sequences [7]. In the same year, Cho et al. proposed gated recurrent units (GRU) instead of LSTM for machine translation tasks [8], and GRU is actually an optimization of LSTM with simplified internal structure, reduced training parameters, and improved training efficiency. However, the model represents the source language sentences of different lengths as a fixed-length vector in the process of encoding, and the longer the sentence, the more information is lost, while the model cannot represent the alignment relationship between the input and output sequences, so it does not guarantee the translation quality effectively. In 2015, Bahdanau et al. proposed the attention mechanism [9], which effectively solves this problem, while encoding using a bidirectional recurrent neural network (Bi-RNN), which further enhances the encoder's ability to characterize information. In 2016 Google released a GNMT system based on a multilayer recurrent neural network approach [10]. The system integrated the neural machine translation technology of the time with many improvements. In less than a year afterwards, Facebook worked on a new neural machine translation system using a convolutional neural network CNN [11], which achieved a higher level of translation than the recurrent neural network RNN-based machine translation system, and the speed of machine translation was greatly improved. In 2017, a new translation structure, Transformer, was proposed by Vaswani et al. It does not use recurrent neural networks and convolutional neural networks at all, but only uses a multi-headed attention mechanism and feedforward neural networks to model, demonstrates

ISSN: 1813-4890

powerful performance without using a sequence-aligned recurrent framework, and cleverly solves the long-distance dependency problem in translation [12]. Transformer is the first model built entirely based on the attention mechanism, which is faster to train and obtains better results on translation tasks, leaping to become the most mainstream neural machine translation framework today. Since then, a lot of work has been done on machine translation, and Chen et al. fused the self-attentive mechanism with the "RNN-RNN" model and achieved significant improvement in translation results [13]. Indurthi et al. proposed a highly attention-based neural machine translation model that further enhances the representational power of attentional mechanisms [14]. Zhou et al. proposed a bi-directional simultaneous decoding approach, where the model dynamically decides the decoding direction of each word [15]. Zhou Xiaoqing et al [16] gave an algorithm that can integrate the output data between the bottom layer and the hierarchy to greatly reduce the time complexity of modeling for the phenomenon of information degradation that can occur in multilayer network structures. Yuqin Ming et al [17] proposed a method for adversarial learning optimization using GAN, which nicely improved the robustness of the NMT model and also obtained better translation performance. Wang et al [18] proposed a new self-attentive mechanism that reduces the overall self-attentive complexity of the model in time and space, resulting in a significant improvement in memory and time efficiency. Beltagy et al [19] introduced a sliding, expansion and fusion window mechanism approach to improve the translation performance of the model in response to the inability of the transformer-based model to handle long sequences. Zaheer et al [20] applied windowing and global attention mechanismsin solving the problem of the limitation of model time complexity on sequence length, which enabled the model to handle longer contexts and greatly improved the performance of translation. Kitaev et al [21] used locally sensitive hash attention to replace dot product attention and used reversible residual networks to replace standard residual networks, which are more efficient in memory usage and faster in speed on long sequences. Choromanski et al [22] simplified the model complexity by introducing a generic attention mechanism approach.

Research based on end-to-end machine translation has much room for advancement in other aspects besides the improvement of its model, and can also combine linguistic knowledge to effectively improve the quality of the translation. Wu et al [23] were the first to introduce dependent syntactic knowledge in RNN-based translation models and proposed a method with syntactic knowledge fusion, which has three encoders and two decoders needed to provide dependent syntactic information of the target language at the same time. The model fuses the dependent syntactic information of the target language at the decoding end and obtains the output at the decoding end by the guidance of dependent syntactic knowledge, but the method is not targeted for neural machine translation under low resource conditions. Zhang et al [24] integrated the source language side of the grammar by cascading the intermediate representation of the dependent parser with the word embedding. The method consists of a parsing model and a neural machine translation model, and the implicit state generated by the encoder of the parsing model is used as the input of the translation model, and the result of the dependent parsing of the source language sentence can be obtained while translating, but the method does not allow learning word units at the source language end. Saunders et al [25] used grammatical representations to interweave words and proposed a derivation-based representation that can replicate the original tree directly from the sequence, thus maintaining structural information, but this leads to the appearance of longer sequences and requires the use of gradient accumulation for effective training. Choshen et al [26] proposed a general method for Transformer-based tree and graph decoding based on generating transformation sequences, which was experimentally shown to outperform the standard Transformer decoder. An et al [27] segmented English long sentences using syntactic information from the source language and demonstrated the effectiveness of machine translation based on long sentence

segmentation. Wang et al [28] fused source language syntactic parse trees into convolutional neural networks and achieved good results in Chinese-Vietnamese translation.

4. Summary and Outlook

In summary, the main research approach in the field of machine translation is neural machine translation, and neural machine translation techniques are also playing an increasingly important role in various NLP tasks. NMT represents a new machine translation model that has fully surpassed statistical machine translation in terms of translation performance and translation quality. However, neural machine translation cannot fully meet the human requirements for translation quality, and it has many problems to be solved. First, neural machine translation requires the support of large-scale floating-point operations, and the inference speed of the model is low. In order to obtain high-quality translation results, the support of a large number of GPU devices is often required, and the cost of computational resources is high. Secondly, due to the lack of human a priori knowledge to guide the translation process, the operation process of neural machine translation lacks interpretability and the system is less intervenable; In addition, neural machine translation still requires manual design of the network structure and a lot of manual involvement in the setting of various hyperparameters of the model and the selection of training strategies. Although the method still has some shortcomings, such as the model architecture still needs to be optimized, the training algorithm needs to be strengthened and improved, and the interpretability of the model in the training process needs to be improved, but neural machine translation will definitely become the future development direction of machine translation.

Acknowledgments

This research is supported by the postgraduate innovation fund project of Chongqing University of Technology (No. gzlcx20223198).

References

[1] Weaver W. Translation. Memorandum. Reprinted in WN Locke and AD Booth, eds[J]. Machine Translation of Languages: Fourteen Essays, 1949.

[2] Makoto Nagao. “A framework of a mechanical translation between Japanese and English by analogy principle”. In: Artificial and human intelligence, 1984, pages 351–354.

[3] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Freder-ick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. “A Statistical Approach to Machine Translation”. In: volume 16. 2. Computational Linguistics,1990, pages 79–85.

[4] William A. Gale and Kenneth W. Church. “A program for aligning sentences in bilingual corpora”. In: volume 19. 1. Computational Linguistics, 1993, pages 75–102.

[5] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[6] Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1700-1709.

[7] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 2(7): 3104-3112.

[8] Cho K, Merrienboer B V, Gulçehre C, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation[C]. Conference on Empirical Methods in Natural Language Processing. 2014: 1724-1734.

[9] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]. 3rd International Conference on Learning Representations. 2015: 408-422.

[10] Yonghui, W., et al. "Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

[11] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. “Convolutional Sequence to Sequence Learning”. In: volume 70. International Conference on Machine Learning, 2017, pages 1243–1252.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is All You Need”. In: International Conference on Neural Information Processing, 2017, pages 5998–6008.

[13] Chen M X, Firat O, Bapna A, et al. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 76-86.

[14] Indurthi S R, Chung I, Kim S. Look harder: A neural machine translation model with hard attention[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3037-3043.

[15] Zhou L, Zhang J, Zong C. Synchronous Bidirectional Neural Machine Translation[J]. Transactions of the Association for Computational Linguistics, 2019, 7(5): 91-105.

[16] Xiaoqing Zhou, Xiangyu Duan, Hongfei Yu, et al. Multi-layer information fusion for neural machine translation [J]. Journal of Xiamen University, 2019, 58 (02): 149-157.

[17] Yuqin Ming, Tian Xia, Yanbing Peng. Neural machine translation based on GAN model optimization [J]. Chinese Journal of Informatics, 2020, 34 (04): 47-54.

[18] Wang S, Li B Z, Khabsa M, et al. Linformer: Self-attention with linear complexity[J]. arXiv preprint arXiv:2006.04768, 2020.

[19] Beltagy I, Peters M E, Cohan A. Longformer: The long-document transformer[J]. arXiv preprint arXiv:2004.05150, 2020.

[20] Zaheer M, Guruganesh G, Dubey K A, et al. Big Bird: Transformers for Longer Sequences[C]. Proceedings of the 34th Conference on Neural Information Processing Systems. 2020: 17283-17297.

[21] Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient transformer [J]. arXiv preprint arXiv: 2001. 04451, 2020.

[22] Choromanski K, Likhosherstov V, Dohan D, et al. Rethinking attention with performers [J]. arXiv preprint arXiv: 2009. 14794, 2020.

[23] Wu S, Zhang D, Zhang Z, et al. Dependency-to-Dependency Neural Machine Translation [J]. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 2018, 26(11):2132-2141.

[24] Zhang M, Li Z, Fu G, et al. Syntax-Enhanced Neural Machine Translation with Syntax-Aware Word Representations[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, United States, 2019: 1151-1161.

[25] Saunders D, Stahlberg F, De Gispert A, et al. Multi-representation ensembles and delayed SGD updates improve syntax-based NMT[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 319-325.

[26] Choshen L, Abend O. Transition based Graph Decoder for Neural Machine Translation [EB/OL]. [2021-05-10]. https://arxiv.org/pdf/2101.12640.

[27] Zhijin Li, Hua Lai, Yonghua Wen, Shengxiang Gao. Neural machine translation incorporating a bidirectional dependent self-attentive mechanism [J/OL]. Computer Applications:1-8[2022-07-11]. https://kns.cnki.net/kcms/detail/51.1307. TP. 20220426.1335.002.html.

[28] Wang Z H, He J Y L, Yu Z T, et al. Chinese-Vietnamese Convolutional Neural Machine Translation with Incorporating Syntactic Parsing Tree [J]. Journal of Software, 2020, 31(12):3797-3807.