A Neural Machine Translation Model Based on Sequence to Dependency

Ting Yang1, Shinan Zhao1, He Chen2 and Ting Liu*  
1 School of Foreign Language Studies, Shaanxi University of Chinese Medicine, Xianyang, Shaanxi, 712046, China  
2 School of Food and Biologlical Engineer, Shaanxi University of Science & Technology, Xi’an, Shaanxi, 710021, China  
3 School of Physics& Electronic Engineering, Xianyang Normal University, Xianyang, Shaanxi,712000, China  
*Corresponding author’s e-mail: xuexiaqing0601@dingtalk.com

Abstract. A supervised self-attention network is proposed to introduce the dependency structure into the Transformer model. This scheme is mainly proposed based on the characteristics of the self-attention mechanism in Transformer, which converts the dependency syntax tree into two equivalent adjacency matrices, and then uses the adjacency matrix to supervise the self-attention network in the Transformer's encoder. So that Transformer learns how to model the dependency structure of the source language. Although this scheme is simple, the effect is remarkable. In particular, there is no need to modify the network structure, as long as two additional loss functions are added in the process of training Transformer. In the decoding process, there is no need to use an external syntactic analyzer to perform syntactic analysis on the source language, but it can automatically construct the dependent syntactic structure of the source language. Experiments demonstrate that the quality of the constructed dependency syntactic structure is also good.

1. Introduction

In recent years, with the rapid development of machine translation technology, machine translation systems have gradually been applied in various fields to help people complete a series of cross-language tasks. Compared with the traditional rule-based machine translation method, the statistical machine translation method has the advantages of stronger scalability and robustness, which can obtain better translation performance from a larger corpus[1]. But it needs high-quality parallel corpus as support, and high-quality parallel corpus is very costly. Deep learning has brought great changes to machine translation, making it no longer necessary to face the problem of feature design in statistical machine translation[2]. Neural machine translation usually uses sequence-sequence neural network models to model translation tasks, and its translation results far exceed traditional statistical machine translation methods in terms of fluency and fidelity, bringing machine translation technology to a new level. Normally, the neural machine translation model uses an encoder to encode source language sentences in the form of sequences, and then uses a decoder to generate target language sentences in the same sequence[3]. However, the composition of sentences in a language is not a simple list of words. In fact, every sentence is constrained by the grammatical structure of the language. For example, an English sentence can be composed of a subject, a predicate, and an
object[4,5]. Each English word can play one or more roles. They need to be combined into a correct sentence under the constraints of grammar. The grammatical structure is a very important part of the language. The neural machine translation model often ignores the grammatical structure in the process of translation, leading that the translation still have some errors that violate the grammatical constraints. In the field of natural language processing, grammar is usually represented by a syntactic structure. In recent years, neural machine translation based on syntax has become a hot spot in neural machine translation research.

This paper proposes a new method of applying source language syntactic structure in neural translation model, and the syntax adopts dependent syntactic structure. This method first obtains different traversal sequences by traversing the source language dependency tree in different ways, and then uses the obtained traversal sequence together with the source language sentence as the input of the neural translation model. In fact, the different traversal methods of the tree structure imply the knowledge of the tree structure to varying degrees, and the tree structure can be reconstructed through the results of these traversals. Taking these sequences containing structured information as input can encode the syntactic structure to a certain extent. This method can be universally used in the Transformer model of the neural translation model based on RNN.

2. Principle

In order to combine the encoder of neural machine translation with the dependency structure of the source language, this paper constructs two different sequences on the basis of the dependency syntax tree of the source language. These two sequences are descriptions from different perspectives of the source language dependency syntax tree, which are called Child Enriched Sequence (CES) and Head Enriched Sequence (HES) respectively. In order to encode two structured sequences of CES and HES in the neural machine translation model, this article uses two additional encoders, named CES-Encoder and HES-Encoder respectively, to encode this sequence. Take the RNN-based neural machine translation model framework as an example. The two additional encoders are named CES-RNN and HES-RNN, combined with the original two-way RNN encoding structure, each word $x_j$ in the source language sentence will be encoded by the encoder, and four corresponding hidden layer vectors will be generated. The hidden layer vectors obtained by the bidirectional RNN encoder of the original neural machine translation are $\tilde{h}_j$ and $\tilde{h}_j$. The hidden layer vector generated by CES-RNN is $\tilde{h}^l_j$, and the hidden layer vector generated by HES-RNN is $\tilde{h}^h_j$. According to the above four vectors, the final hidden layer $h_j$ of word $x_j$ can be calculated, and $h_j$ can be expressed as

$$h_j = \tanh\left( W_h \tilde{h}_j + U_h \tilde{h}_j + V_h \tilde{h}^l_j + F_h \tilde{h}^h_j \right)$$  \hspace{1cm} (1)$$

Where $W_h, U_h, V_h$ and $F_h$ are weight matrix.

This method is also applicable to Transformer, the most mainstream neural machine translation model. On the basis of the original Transformer model, CES-Encoder and HES-Encoder are additionally introduced to encode CES and HES. The encoding result also uses the above formula to generate the final hidden layer vector for each source language word. Figure 1 shows the improved Transformer encoder structure without the decoder part, and the encoder consists of three identical structures, the difference is that each encoder has its own input.
Figure 1. Source dependency based encoder of Transformer.

Figure 1 shows the improved Transformer encoder structure which is not require the decoder part, and the encoder consists of three identical structures, the difference is that each encoder has its own input. It is worth to be noted that no matter in CES or HES, the positional encoding is given sequentially, which means that the same word may have different positional encodings in different sequences.

3. Experiment and result

We have experimentally verified the translation model through Chinese-English translation. The training data contains 2 million parallel corpora, which is a subset of the LDC dataset, and the data belongs to the field of news. In the experiment, the word vector dimension is set to 512, and the hidden layer representation dimension of the RNN is 1024. The size of the vocabulary and the processing of unregistered words depend on different translation tasks. All network parameters are randomly initialized by Gaussian distribution. Set the learning rate to 1, and the batch size to 128. In addition, this paper uses the Adadelta algorithm to automatically adjust the learning rate and update the network parameters. During the decoding process, the bar search size is set to 12.

Figure 2. The BLEU changes in terms of parsing accuracy

Figure 2 shows the difference between the BLEU of SSE-NMT or SD-NMT and RNNsearch. It can be seen from the trend of the curve in the figure that when the syntactic accuracy is 50%, the performance is very poor. Using such a syntactic analyzer to process bilingual data will bring many
errors. A large number of syntactic errors caused the translation ability of the neural translation model not to be enhanced, but also brought negative effects. When the syntactic accuracy is increased to 60%, the performance of SD-NMT and SSE-NMT is enhanced, but there is no significant difference compared with the baseline model RNNsearch. When the syntactic accuracy was further improved to more than 70%, the syntax began to play a positive role in promoting the neural translation model. Therefore, it can be seen that the accuracy of syntax has a relatively large impact on the translation model based on syntax.

For all Transformer-based neural translation models, we set the network according to the basic model parameters proposed by Vaswani. Table 1 shows the experimental results on all test sets. The numbers in bold indicate a significant increase compared to the benchmark system.

Table 1 Evaluation results on the NIST Chinese-English task

| Models              | NIST2005 | NIST2006 | NIST2008 | NIST2012 | AVERAGE |
|---------------------|----------|----------|----------|----------|---------|
| HPMST               | 35.34    | 33.72    | 25.64    | 28.26    | 30.74   |
| RNNSearch           | 38.07    | 38.87    | 31.13    | 28.57    | 34.16   |
| SSE-MMT+CES         | 38.59    | 40.03    | 31.89    | 30.66    | 35.29   |
| SSE-MMT+HES         | 38.76    | 40.12    | 33.03    | 31.28    | 35.80   |
| SSE-MMT+CES+HES     | 39.02    | 40.39    | 34.24    | 31.89    | 36.39   |

The experimental results in Table 1 show a significant increase compared with the benchmark system. It can also be seen from Table 1 that the evaluation results of all neural network-based translation models are higher than HPSMT, and all models based on source language syntax information are due to the basic RNNSearch model. Based on this, it can be seen that in this task, the syntactic results of the source language are indeed helpful to the neural machine translation model. It can be proved that the syntactic results of the source language are indeed helpful to the neural machine translation model.

4. Conclusion
This paper proposes a method of applying source language dependent syntactic structure in neural machine translation model. This method converts the syntax tree into a corresponding structured sequence through different traversal methods of the source language dependent syntax tree, and then uses an additional encoder for encoding. The modeling process is simple and effective, and does not rely on complex graph model networks. At the same time, this method is commonly used in RNN-based translation models and Transformers. Through a large number of experiments and analysis, it is proved that the method proposed in this chapter can effectively improve the translation ability of neural machine translation. Although the syntactic structure is very helpful to neural machine translation, there are very few languages with syntactic structure annotation at this stage. Therefore, in the future, we can consider migrating languages with syntactic structure to rare languages in the translation process to obtain their syntactic structure, and at the same time use the syntactic structure to assist machine translation.

Acknowledgments
This work was supported in part by Shaanxi Provincial Key Research and Development Program 2021NY-213, Xianyang Normal College Youth Key Program XSYGG201603, Special Scientific Research Program of Shaanxi Provincial Department of Education 19JK0928, Shaanxi Provincial Social Science Foundation Program 2020K014.
References

[1] Kitaev N, Klein D. (2018) Constituency Parsing with A Self-Attentive Encoder. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia. pp. 165-178.

[2] Zhang X, Lapata M, Wei F, et al. (2018) Neural Latent Extractive Document Summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. pp. 779-784.

[3] Wang H, Wang X, Xiong W, et al. (2019) Self-Supervised Learning for Contextualized Extractive Summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. pp. 2221-2227.

[4] Ren P, Chen Z, Ren Z, et al. (2017) Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA. pp. 95-104.

[5] Xu J, Durrett G. (2019) Neural Extractive Text Summarization with Syntactic Compression. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China. pp. 3290-3301.