Comparison of the effects of attention mechanism on translation tasks of different lengths of ambiguous words

Yue Hu[#]
Xi'an Jiaotong-Liverpool University
Suzhou, China
aeoluze@gmail.com

Jiahao Qin[#]
Xi'an Jiaotong-Liverpool University
Suzhou, China
jiahao.qin19@outlook.com

Zemeiqi Chen
Xi'an Jiaotong-Liverpool University
Suzhou, China
zemiqi.chen18@outlook.com

Xiaojun Zhang
Xi'an Jiaotong-Liverpool University
Suzhou, China
xiaojun.zhang01@xjtlu.edu.cn

Jingshi Zhou
Xi'an Jiaotong-Liverpool University
Suzhou, China
Jingshi.Zhou@outlook.com

Abstract

In recent years, attention mechanism has been widely used in various neural machine translation tasks based on encoder decoder. This paper focuses on the performance of encoder decoder attention mechanism in word sense disambiguation task with different text length, trying to find out the influence of context marker on attention mechanism in word sense disambiguation task. We hypothesize that attention mechanisms have similar performance when translating texts of different lengths.

Our conclusion is that the alignment effect of attention mechanism is magnified in short text translation tasks with ambiguous nouns, while the effect of attention mechanism is far less than expected in long-text tasks, which means that attention mechanism is not the main mechanism for NMT model to feed WSD to integrate context information. This may mean that attention mechanism pays more attention to ambiguous nouns than context markers. The experimental results show that with the increase of text length, the performance of NMT model using attention mechanism will gradually decline.

1 Introduction

Natural language always contains many ambiguous words. Ambiguous words mean that the same word can express many different meanings in different contexts. Since the actual meaning of ambiguous words is closely related to context information and context, it is always a challenge to machine translates sentences containing ambiguous words.

In statistical machine translation (SMT) system, we can improve the translation effect of ambiguous words by taking the context marker of ambiguous words into account. In the neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013) task, especially in the NMT (Chorowski et al., 2015) model which uses attention mechanism, the hidden state of each input contains context information. Therefore, theoretically, the attention mechanism can make the NMT model better deal with the translation task of ambiguous words. However, there are no empirical results to show that attention mechanism can obtain enough information of word sense disambiguation from hidden state. Moreover, we still don't know the deep principle of the attention mechanism to deal with ambiguous words. This paper mainly studies whether the NMT model including attention mechanism will have different performance when dealing with different states of ambiguous word text. We compared the performance of different NMT models using attention mechanism in different length texts with ambiguous words, and evaluated them with Bleu. In the following chapters, we use different NMT models with attention mechanism to test the long text and short text containing ambiguous words, and compare the influence of the length of the text containing ambiguous words on the translation effect. Our findings are summarized as follows:

- We find that the performance of NMT model on ambiguous data sets is not as good as ordinary data sets, which proves that the task of translating ambiguous nouns is more difficult than ordinary machine translation.

- We find that the performance of NMT model in short text task with ambiguous words is much better than that in long text task

Our conclusion is that the alignment effect of the attention mechanism is magnified in the short text translation task with ambiguous nouns, while the effect of the attention mechanism in the long text task is far less than expected.

#:They have the same contribution to the article.

Proceedings of the 2nd International Workshop on Discourse Processing, pages 18–21

2 Related Work

Vaswani et al. abandoned the traditional encoder decoder model which must be combined with the inherent pattern of CNN or RNN, and only used the attention mechanism, which can reduce the amount of computation and improve the parallel efficiency without damaging the final experimental results. Then, they proposed two new attention mechanisms (Vaswani et al., 2017). Tang, g et al. verified that self-attention mechanism is superior in WSD tasks (Tang et al., 2018a)(Tang et al., 2018b). All these works prove that the attention mechanism helps NMT model to achieve better performance in ambiguous word translation task. However, there is no detailed analysis on the performance of the attention mechanism in different scale texts.

This paper mainly studies the performance of attention mechanism on word sense disambiguation in different length texts. More specifically, we explore whether attention plays a better role in long text word sense disambiguation tasks as well as in short text tasks or not.

This paper mainly studies the performance of attention mechanism on word sense disambiguation in different length texts. More specifically, we explore whether the role of attention in long text translation is as good as we expected or not.

3 Evaluation

In this paper, NMT model is used to evaluate the translation of data sets with different text length. We evaluate two popular NMT models with different attention mechanisms, one is Google T5 with advanced attention mechanism, the other is Martin MT using ordinary attention mechanism (Tiedemann and Thottingal, 2020). We use two different data sets for comparative test, one is the ordinary Opus parallel corpus, and the other is the screened corpus data set containing ambiguous words.

The original text length of the data set containing ambiguous nouns ranges from three words to eighty words. In order to ensure the differentiation of NMT model in short text task and long text task, we randomly extract 2000 texts from the test data set with less than 20 words and the test data set with 40 to 80 words as the comparison data set the results were translated by NMT model and evaluated by Bleu.

3.1 Experimental Settings

We use a pre-training model based on HuggingFace's transformers (Wolf et al., 2019), which greatly reduces the experimental time. Google T5 model and Martin MT model, which use attention mechanism and are representative, are selected, and C4 dataset (Raffel et al., 2020) is used for model training.

We used the data set from contrawsd (Rios Gonzales et al., 2017) with ambiguous words as a comparison with the ordinary opus data set. All Bleu scores were evaluated by SacreBleu (Rios Gonzales et al., 2017). After filtering ambiguous nouns, 5000 sentences of test data set containing ambiguous nouns are left in contrawsd data set. At the same time, the same amount of data is randomly selected from opus data set, and these test data are divided into different groups according to the text length. The data with significant difference in text length will be evaluated by NMT model with attention mechanism.

4 Results

Table 1 shows the performance of two NMT models for long text and short text processing on two datasets. No matter T5 or Martin MT, there is no obvious difference in the processing of long text and short text in ordinary Opus dataset. However, on the contrawsd dataset with ambiguous words, the performance of both NMT models in short text tasks is significantly higher than that in long text tasks.

Model-Dataset	Short-Text	Long-Text	Mean
T5-ContraWSD	26.078	7.877	16.978
T5-OPUS	25.992	22.817	24.404
Martin-ContraWSD	25.323	12.479	18.901
Martin-OPUS	26.805	25.067	25.936

Table 1: BLEU score in different task

Therefore, compared with the long text task, the NMT model based on the attention mechanism performs better in short text translation tasks with ambiguous nouns. In order to further verify the relationship between text length and NMT model with attention mechanism in translation performance, we further split the data set and get the trend shown in Figure 1.

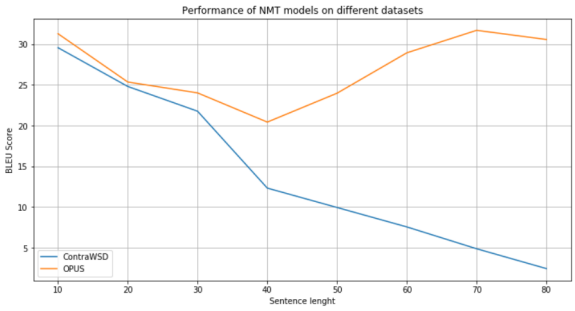

Figure 1: Examples of cohesion

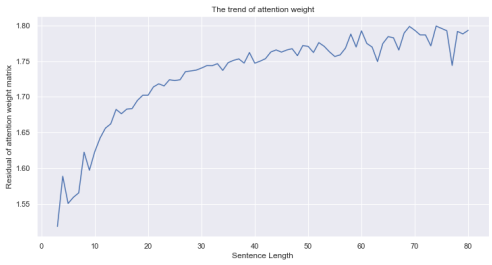

Figure 2: Examples of cohesion

5 Analysis

Through the data, it can be found that the average performance score of two NMT models with attention mechanism on ambiguous data sets is not as good as that of ordinary data sets, which proves that the text with ambiguous words does affect the performance of machine translation tasks. That is to say, the analysis of the performance of the NMT model with attention mechanism on ambiguous text is meaningful.

By comparing the performance of NMT model on different length texts, we can get the following information. First of all, the different text length on ordinary opus dataset has no obvious effect on NMT model translation performance. On the contrawsd dataset with ambiguous words, it is obvious that the performance of NMT model in the translation task of ambiguous text tends to decline with the increase of text length.

Next, we explore the details of the attention weight when dealing with ambiguous words. Due to different text scales, we use a residual coefficient to express the attention weight matrix. The specific calculation method is as follows:

$$R = \frac{\sum(|E_n - w_i|)}{n}$$

Where n is the length of the text, e is the diagonal matrix of size n, and wi is the attention weight matrix of layer I. The larger the residual coefficient r is, the less attention weight is. Then we test the variation trend of R under different text lengths on the contrawsd dataset containing ambiguous word text, and get the results shown in Figure 2.

The results show that with the increase of text length, the performance of the attention weight gradually decreases, which just confirms that the performance of the NMT model with attention mechanism decreases with the increase of the length of the text.

6 Conclusion

This paper analyzes the performance of the attention mechanism in NMT for different length texts with ambiguous words. We use Opus and contrawsd as test sets to evaluate the translation of NMT model under different length texts. The results showed that the Bleu scores of T5 model on short text and long text were 26.078 and 7.877, while those of Martin MT model were 25.323 and 12.479, respectively. The sparsity of ambiguous words in the training set may be the main problem leading to incorrect translation. However, the attention mechanism does perform better on short text data sets than on long text data sets. This is probably because the fact that the attention mechanism tends to pay more attention to the ambiguous words themselves than to the context markers.

It is significant to understand the performance differences of attention mechanism on different scale text data sets, which may indicate the potential optimization direction of attention mechanism in such tasks. We hope that our future work can continue to improve our understanding of the deep-seated principle of attention mechanism in NMT model, analyze the details of attention weight in NMT model when dealing with ambiguous word text tasks, and explore how to improve the translation effect of NMT model in different scale text sense disambiguation tasks.

Acknowledgments

This work is supported by the XJTLU KSF project (Grant Number: KSF-E-24) and GDUFS open project (Grant Number: CTS201501). The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-Based Models for Speech Recognition. *Advances in Neural Information Processing Systems*, 28:577–585.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*. ArXiv: 1910.10683.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018a. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. In *Tang, Gongbo; Müller, Mathias; Rios, Annette; Sennrich, Rico (2018). Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, 2 November 2018 - 4 November 2018.*, Brussels. ACL.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018b. An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation. *arXiv e-prints*, 1810:arXiv:1810.07595.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv e-prints*, 1910:arXiv:1910.03771.