TNLP Literature Review
Attention Mechanism in Neural Machine Translation

s1719048

Abstract

Neural machine translation (NMT) has obtained lots of progress recently, which translates source language to target language via a large single neural network. Due to the introduction of attention mechanism, there is more room in neural machine translation for improvements. This literature review focuses on the improvements of the attention model in neural machine translation. Several modifications based on attention model are introduced, for example, the usage of the different attention functions, the incorporation of various biases, and also the use of the recurrent neural network (RNN). All the attention models after modification obtain better performances during translation.

1 Introduction

Neural machine translation (NMT) jointly trains a large single neural network to translate. Most of the current NMT models consist of an encoder-decoder structure (Bahdanau, Cho, & Bengio, 2014). Source sentence is encoded by the encoder into a context vector, and target sentence is generated by the decoder based on the context vector and the previous predicted words. Early context vector with fixed representation (Cho et al., 2014) could hardly capture information of longer sentences, therefore, Bahdanau, Cho, and Bengio (2014) proposed an attention mechanism in NMT, which could dynamically learn the context vector by an attention model for each target word.

Following Bahdanau et al. (2014)'s work, several methods were introduced to improve the attention model. Earlier efforts were put on the different choices of attention functions (Luong, Pham, & Manning, 2015). Then, other works focused on adding biases to the original attention model, such as the *coverage* bias (Tu, Lu, Liu, Liu, & Li, 2016), the $fertility$ bias (Feng, Liu, Li, & Zhou, 2016), the *position* and $Markov$ biases (Cohn et al., 2016). Recently, recurrent neural networks(RNN) are adopted to model the attention, for example, the long short-term memory (LSTM) attention model (Yang, Hu, Deng, Dyer, & Smola, 2016), and the gated recurrent units (GRU) attention model (Zhang, Xiong, & Su, 2017). These explorations on attention model enhance the NMT system's ability to capture more useful information about the source and target languages.

This review will focus on to what extent these improvements influence the attention model, and how much effect they have on the NMT performance. To make this review specific and concise, for

each method mentioned below, the overall procedure of encoding and decoding is simplified, and the emphasis is put on the construction of the attention models. In the rest of this document, the vanilla attention model will be introduced in section 2, different types of attention models will be illustrated in section 3. Section 4 is the conclusion.

2 NMT with Attention Model

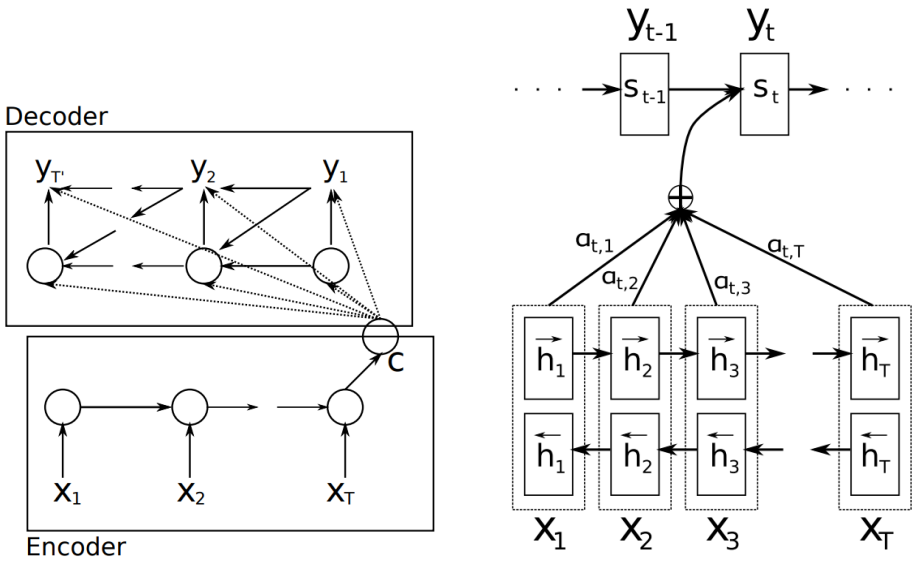
NMT models belong to the structure of encoder-decoder. The decoder generates target sentence based on the context vector that computed by the encoder. In the old version of NMT model, which is shown in Figure 1(a), the context vector is computed after encoding the whole source sentence and keeps constant when generating each target word. Therefore, all target words share the same context vector. Bahdanau et al. (2014) stated that the old format of fixed-representation context vector is not enough to compress the necessary information of longer sentences. Hence, they proposed the NMT model with attention, which is shown in Figure 1(b). The difference between these two models is how to put the attention on source words. In Figure 1(a), all source words are treated similarly when generating each target word, while in Figure 1(b), different target word owns different attention vector (or context vector). Applying attention means assigning weights to source words, the weight represents where and how much emphasis need to put on the source sentence.

The implementation of the attention model could also be illustrated in Equation (1) to (3), where s_i is the hidden state of the decoder for ith target word, h_j is the hidden state of the encoder for jth source word. f_{ij} is the output of the attention model, α_{ij} is the attention vector (or the weights), and c_i is the context vector of the ith target word. We could see that the context vector of ith target word is the weighted sum of the encoder hidden states, where the weights are derived from the attention model $a(s_{i-1}, h_j)$. The attention model is computed by a feed-forward neural network with tanh function.

$$\textbf{attention model:} \quad f_{ij} = a(s_{i-1}, h_j) = v^T \tanh(W s_{i-1} + U h_j) \tag{1}$$

$$\textbf{attention vector(weights):} \quad \alpha_{ij} = softmax(f_{ij}) \tag{2}$$

$$\textbf{context vector:} \quad c_i = \sum_{j=1}^{T} \alpha_{ij} h_j \tag{3}$$



(a) NMT model with fixed context vector

(b) NMT model with attention

Figure 1: The structures of NMT models with fixed context vector and with attention (Cho et al., 2014), (Bahdanau et al., 2014).

On an English-to-French translation task based on WMT dataset, Bahdanau et al. (2014)'s attention model helps to improve the system's BLEU score from 17.82(fixed context vector NMT model of Cho et al. (2014)) to 26.75. Meanwhile, no deterioration occurs for the performance of NMT with attention model when sentence grows longer. Therefore, attention model improves the NMT translation performance by paying different attention to the source sentence. However, the shortage here is, Bahdanau et al. (2014) only adopted a simple feed-forward neural network to train the attention model, more focuses could be put on the improvement of attention models. Next section will demonstrate several methods that explore the attention model in depth.

3 Explorations in NMT Attention Model

3.1 Global and Local Attention

As an extension to Bahdanau et al. (2014)'s work, Luong, Pham, and Manning (2015) introduced two types of attention models, the global attention model and the local attention model, which are shown in Figure 2. These two types of attention share the similar encoding-decoding steps, except the way of computing the context vector. The global attention places attention on the whole source sentence and derives the context vector, which is similar to the work of Bahdanau et al. (2014). Meanwhile, the local attention only puts emphasis on parts of the source words and computes the context vector. In both global and local attention models, three types of attention functions are used, the *dot*, the *general*, and the *concat* (Bahdanau et al. (2014) only adopted *concat*). Moreover, two types of local attentions are adopted, the *monotonic* attention and the *predictive* attention. Finally, Luong et al.

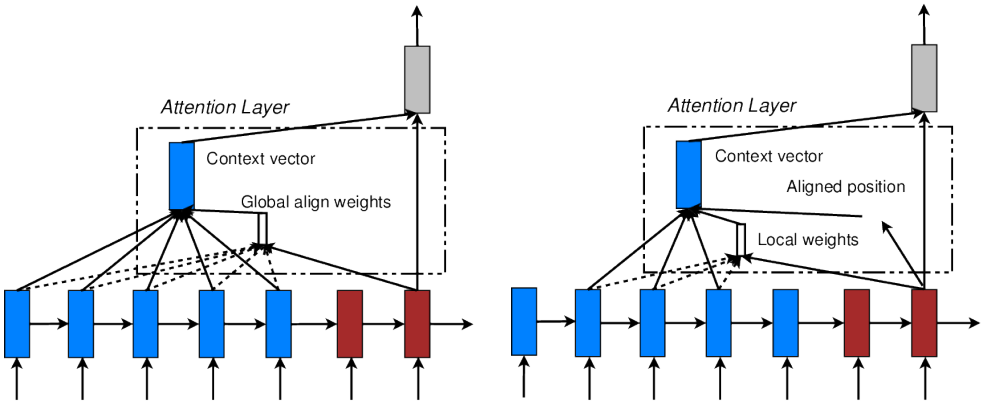
(2015) introduced an approach called $input - feeding$, where the previous context vector is also used as a bias when computing current context vector. The attention decision for each target word now also rely on the previous attention decision.

More specifically, the differences among basic attention model, the global attention model, and the local attention model could be entailed by the change in the Equation (1), the new equations are shown in Equation (4) to (6). The difference among these equations could be summarized as, first, Equation (1) only uses tanh function while Equation (4) adopts three kinds of attention functions. Second, local attention vector has an extra item after $softmax$ (Equation (6)), it decides which subset of source words need to be focused on.

$$\textbf{global and local attention model:} \quad \mathbf{f} = a(\mathbf{s}, \mathbf{h}) = \begin{cases} \mathbf{sh} & dot \\ \mathbf{s}^T W \mathbf{h} & general \\ \mathbf{v}^T \tanh(W\mathbf{s} + U\mathbf{h}) & concat \end{cases} \tag{4}$$

$$\textbf{global attention vector:} \quad \alpha_{ij} = softmax(f_{ij}) \tag{5}$$

$$\textbf{local attention vector:} \quad \alpha_{ij} = softmax(f_{ij}) \exp(-\frac{(s - p_t)^2}{2\sigma^2}) \tag{6}$$



(a) global attention (b) local attention.

Figure 2: The structure of global attention and local attention. (Luong et al., 2015)

Luong et al. (2015) conducted experiments on the English-German bidirectional translation task based on the WMT dataset. Comparisons were made among several existing machine translation systems. Specifically, compared with the basic attention model proposed by Bahdanau et al. (2014), the global attention model improved the BLEU score by 2.8. And the local attention model outperforms the global model, which improved the BLEU score by 0.9.

In summary, Luong et al. (2015) indeed explore the attention model in depth. The global model adds more attention functions based on the basic attention model of Bahdanau et al. (2014), which improves the translation performance. Meanwhile, the local model reduces the computation cost by only looking at a small window in the source sentence, which even outperforms the global attention model. However, there are still some aspects could be improved in this work. First, more language pairs

could be used to train the evaluate the NMT model. Because of the morphology diversity of different languages, this attention model might not suitable for all languages. Second, the translation between two languages has lots of properties, for example, the word order always is the same. These properties could be used as an extra variant when training the attention model. Except the $input - feeding$ bias, more biases could be incorporated into attention. Next section will focus on different biases used in attention model.

3.2 Biases Used in Attention Model

In mature statistical machine translation (SMT), different variants, such as the monotonic bias (Zens & Ney, 2004) and the fertility bias (Lopez, 2008), could be adopted to improve SMT system's performance. And alignment-based SMT has reached a certain height (Matusov, Zens, & Ney, 2004). Similarly, these variants(biases) could also be mapped to NMT attention model. To supplement the previous $input - feeding$ bias used in (Luong et al., 2015), other types of bias are proposed recently, which could be incorporated into the attention model in NMT and improve the system's performance.

Tu et al. (2016) introduced the *coverage* bias into attention model, which could summarize the previous attention history and reuse it to derive context vector. Meanwhile, Feng, Liu, Li, and Zhou (2016) tried the *distortion* bias to address the word re-order problem, and the $fertility$ bias to deal with the alignment between the amount of translated words and the number of source words. Moreover, Cohn et al. (2016) proposed four biases, the *position* bias, the $Markov\ condition$ bias, the *local* $fertility$ bias, and the *global* $fertility$ bias. Each bias represents a property of the word-based machine translation model. *position* bias means the word orders in the source and target sentences are similar. $fertility$ bias indicates each source word type corresponds to a fixed number of target words. $Markov\ condition$ bias reflects the relation between source and target is monotonic. All the biases mentioned above could be added to the attention model. As shown in Equation (7), the biases are added inside the *tanh* function and used to address properties of the relationship between two languages.

$$\textbf{biases attention model:} \quad f_{ij} = v_a^T \tanh(\mathbf{W_a}[\mathbf{t}; \mathbf{s}] + \textbf{biases}) \tag{7}$$

Tu et al. (2016) carried out experiments on language pair of Chinese-English, then evaluated their model based on the translation and alignment quality. Their results showed that *coverage* bias helps to avoid over-translation and under-translation. Feng et al. (2016)'model is also evaluated on Chinese-English language pair, and their *distortion* model outperforms Bahdanau et al. (2014)'s work by 2 in BLEU score, their $fertility$ model outperforms Bahdanau et al. (2014)'s work by 1 in BLEU score. Moreover, Cohn et al. (2016) conducted experiments on four language pairs, English-Romanian, English-Chinese, English-Russian, and English-Estonian. They compared their model with both the NMT model without attention (Cho et al., 2014), and the basic NMT model with attention (Bahdanau et al., 2014). The scores on perplexity show that the NMT model with biases attention outperforms other two models on all four language pairs. The results are shown in Table 1.

Table 1: The perplexity among four language pairs of three models in Cohn et al. (2016)'s work.

	Zh-En	En-Zh	Ru-En	EN-Ru	Et-En	En-Et	Ro-En	En-Ro
(Cho et al., 2014)	5.35	8.60	61.9	67.3	18.2	31.4	10.3	11.5
(Bahdanau et al., 2014)	4.77	7.49	41.7	43.0	12.8	19.4	6.62	7.30
(Cohn et al., 2016)	4.31	6.24	39.9	40.6	11.8	17.0	5.89	6.35

Each model mentioned above introduces useful variants to attention model. Moreover, the *coverage* bias of Tu et al. (2016) also reduce the alignment errors of attention model. Feng et al. (2016) adopted recurrent neural network to build the bias in attention model, which is a new attempt based on the simple feed-forward neural network in Bahdanau et al. (2014)'s work. Cohn et al. (2016) tried different language pairs, which makes the results more convincing.

In conclusion, the combination of biases and attention model improves the whole model's performance, which entails that modeling biases on attention could capture important characteristics of both the source and target languages. These biases enhance the representative ability of the attention model. However, there are several places that might be improved in the future. First, considering the morphology of different languages, some of the properties might not be suitable for some specific languages. For example, the translation between English and Japanese does not obey the 'similar word order' rule. Therefore, more attentions should be paid to the diversity of language morphology. Second, the biases mentioned above could be combined together to form a new attention model. The combination of biases might improve the translation performance.

3.3 RNN Used inside NMT Attention Model

Recurrent neural networks (RNN) are used in both the encoder and decoder of the NMT model. In the encoder side, the context vector is derived from the hidden states in encoder RNN, and in the decoder side, the target words are generated based on the hidden states of the decoder RNN. Different kinds of RNN could be adopted in the encoder-decoder, for example, the long short-term memory units (LSTM) and the gated recurrent units (GRU). Because both of the source and target sentence in NMT is a sequence of words, which could be treated as the sequence-to-sequence task and addressed by RNN (Sutskever, Vinyals, & Le, 2014). Similarly, the attention vectors also appear in sequence (each target word has a unique attention vector), thus, RNN could also be used in attention model. This section will focus on how RNN could be adopted in the attention model.

Yang, Hu, Deng, Dyer, and Smola (2016) chose to use LSTM to model the attention before and after each source word, which means they adopted RNN to the Uh_j item in Equation (1). The new attention model could be interpreted by Equation (8), (9) and (10). As mentioned before, j is the index of the jth source word. The *dynamic memory* (Equation (9)) adopts LSTM, which considers a window ($\pm k$) of attention vectors α before and after the jth source word. We could also interpret this *dynamic memory* as adding information about attention history to the encoder hidden states vector used in attention model. Therefore, in Yang et al. (2016)'s work, LSTM is used to add more information about attention history to attention model.

$$\textbf{attention map:} \quad \tilde{\alpha_{ij}} = [\alpha_{i,j-k}...\alpha_{i,j+k}] \tag{8}$$

$$\textbf{dynamic memory:} \quad d_{ij} = LSTM(d_{i-1,j}, \tilde{\alpha_{ij}}) \tag{9}$$

$$\textbf{attention model:} \quad f_{ij} = v^T \tanh(Ws_{i-1} + U[h_j, d_{ij}]) \tag{10}$$

Yang et al. (2016) conducted experiments on WMT English-German task and NIST Chinese-English task. Compared to the model of Bahdanau et al. (2014), new model with window size 11 increases the BLEU score by 0.5 on English-German task. Meanwhile, the improvement on Chinese-English task is even obvious, where the BLEU score is improved by 1.5. This better performance indicates that adopting RNN inside the attention model could improve the translation performance. It is a new attempt to modify attention model by supplementing information instead of adding biases after the attention model.

Following Yang et al. (2016)'s work, Zhang, Xiong, and Su (2017) proposed a GRU attention model in NMT. They combined the encoder states and decoder states together and fed the combination to GRU units. New representation was passed through the attention model to derive the attention vector. To explain this model in a unified form, we could change the Uh_j item in Equation (1) to a new item Uh_{gru}, where h_{gru} is the output of the GRU unit. Then, the GRU attention model could be described by Equation (11). Then the difference between LSTM-based attention and GRU-based attention is how to modify the encoder hidden states. But the core idea of both methods is same, which aims at providing more information to the attention model.

$$\textbf{attention model:} \quad f_{ij} = v^T \tanh(Ws_{i-1} + Uh_{gru}) \tag{11}$$

Zhang et al. (2017) conducted experiments on Chinese-English translation task. Their NMT with GRU-based attention model outperforms the basic attention NMT model (Bahdanau et al. (2014)) by 1.66 in BLEU score. This improvement indicates that GRU could also help to enhance the attention model.

In summary, both the LSTM-based attention model (Yang et al., 2016) and the GRU-based attention model (Zhang et al., 2017) enhance the previous attention model and improve the NMT system's performance. They add more information (history of attention vectors) to the attention model by implementing different kinds of RNN. However, there are several places that could be focused on in the future work. First, RNN could also be used to cooperate with the decoder hidden states to supplement information related to target words. Second, other types of neural networks might be experienced, for example, the convolutional neural network.

4 Conclusion

This literature review introduced the attention model used in NMT, then summarized several improvements on the attention model. Bahdanau et al. (2014) addressed the representation limitation of a fixed-length context vector and proposed the attention model. Then, Luong et al. (2015) added different attention functions based on Bahdanau et al. (2014)'s model, and also tried the local attention. After that, Tu et al. (2016), Feng et al. (2016), and Cohn et al. (2016) incorporated different biases

with the attention model, which helps grasp properties of language pairs. Moreover, LSTM and GRU are respectively adopted by Yang et al. (2016) and Zhang et al. (2017), which supplement information to the attention model, and makes a further improvement on the NMT system.

All the models mentioned in this literature review outperform the basic attention-based NMT model on different language pairs. In order to make the comparisons more clear, formulas of different NMT attention models are illustrated in Table 2, from which we could tell where the differences are. As shown in this table, except the attention functions of Luong et al. (2015), other attention models vary based on the item inside the 'tanh'. Different constructions could capture different characteristics of both languages pairs and attention models. Most of the biases represent various properties of the language pairs, while the RNN term is the reflection of the attention history.

Table 2: Different attention models.

	attention model
(Bahdanau et al., 2014)	$f_{ij} = a(s_{i-1}, h_j) = v^T \tanh(W s_{i-1} + U h_j)$
(Luong et al., 2015)	$\mathbf{f} = a(\mathbf{s}, \mathbf{h}) = \begin{cases} \mathbf{sh} & dot \\ \mathbf{s}^T W \mathbf{h} & general \\ \mathbf{v}^T \tanh(W\mathbf{s} + U\mathbf{h}) & concat \end{cases}$
(Tu et al., 2016)	$f_{ij} = v_a^T \tanh(\mathbf{W_a}[\mathbf{t}; \mathbf{s}] + \mathbf{biases})$
(Feng et al., 2016)	
(Cohn et al., 2016)	
(Yang et al., 2016)	$f_{ij} = v^T \tanh(W s_{i-1} + U[h_j, d_{ij}])$
(Zhang et al., 2017)	$f_{ij} = v^T \tanh(W s_{i-1} + U h_{gru})$

In summary, all these modifications on the attention model have better performances. However, considering the advanced techniques used in SMT, the increase of larger dataset, and also the development of RNN, more efforts could be put on the existing NMT attention model. From my limited view, there are several aspects might be focused on in the future:

- RNN could also be adopted to supplement the decoder hidden states, which means in the Equations in Table 2, RNN could incorporate with s_{i-1}. Because when deriving the context vector by attention model, the predicted target words also need to be considered.
- Models could be trained on new dataset, a larger dataset might improve the system's performance.
- Morphological diversity could be considered during the model construction. Morphology-rich language(Arabic) and morphology-poor language(English) will have different influences on the attention model. Considering translation between different morphology languages is mature in SMT (Avramidis & Koehn, 2008), we could map the techniques used in SMT to the NMT attention model.

References

Avramidis, E., & Koehn, P. (2008). Enriching morphologically poor languages for statistical machine translation. *Proceedings of ACL-08: HLT*, 763–770.

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

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*.

Cohn, T., Hoang, C. D. V., Vymolova, E., Yao, K., Dyer, C., & Haffari, G. (2016). Incorporating structural alignment biases into an attentional neural translation model. *arXiv preprint arXiv:1601.01085*.

Feng, S., Liu, S., Li, M., & Zhou, M. (2016). Implicit distortion and fertility models for attention-based encoder-decoder nmt model. *arXiv preprint arXiv:1601.03317*.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)*, *40*(3), 8.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Matusov, E., Zens, R., & Ney, H. (2004). Symmetric word alignments for statistical machine translation. In *Proceedings of the 20th international conference on computational linguistics* (p. 219).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*.

Yang, Z., Hu, Z., Deng, Y., Dyer, C., & Smola, A. (2016). Neural machine translation with recurrent attention modeling. *arXiv preprint arXiv:1607.05108*.

Zens, R., & Ney, H. (2004). Improvements in phrase-based statistical machine translation. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*.

Zhang, B., Xiong, D., & Su, J. (2017). A gru-gated attention model for neural machine translation. *arXiv preprint arXiv:1704.08430*.