Interactive Attention for Neural Machine Translation

Fandong Meng[1*] **Zhengdong Lu**[2] **Hang Li**[2] **Qun Liu**[3,4]

[1]AI Platform Department, Tencent Technology Co., Ltd.
fandongmeng@tencent.com
[2]Noah's Ark Lab, Huawei Technologies
{Lu.Zhengdong,HangLi.HL}@huawei.com
[3]ADAPT Centre, School of Computing, Dublin City University
[4]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS
qliu@computing.dcu.ie

Abstract

Conventional attention-based Neural Machine Translation (NMT) conducts dynamic alignment in generating the target sentence. By repeatedly reading the representation of source sentence, which keeps fixed after generated by the encoder (Bahdanau et al., 2015), the attention mechanism has greatly enhanced state-of-the-art NMT. In this paper, we propose a new attention mechanism, called INTERACTIVE ATTENTION, which models the interaction between the decoder and the representation of source sentence during translation by both reading and writing operations. INTERACTIVE ATTENTION can keep track of the interaction history and therefore improve the translation performance. Experiments on NIST Chinese-English translation task show that INTERACTIVE ATTENTION can achieve significant improvements over both the previous attention-based NMT baseline and some state-of-the-art variants of attention-based NMT (i.e., coverage models (Tu et al., 2016)). And neural machine translator with our INTERACTIVE ATTENTION can outperform the open source attention-based NMT system Groundhog by 4.22 BLEU points and the open source phrase-based system Moses by 3.94 BLEU points averagely on multiple test sets.

1 Introduction

Neural Machine Translation (NMT) has made promising progress in recent years (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015a; Jean et al., 2015; Luong et al., 2015b; Tang et al., 2016; Wang et al., 2016; Li et al., 2016; Tu et al., 2016; Shen et al., 2016; Zhou et al., 2016), in which attention model plays an increasingly important role. Attention-based NMT represents the source sentence as a sequence of vectors after a RNN or bi-directional RNN (Schuster and Paliwal, 1997), and then simultaneously conducts dynamic alignment with a gating neural network and generation of the target sentence with another RNN. Usually NMT with attention model is more efficient than its attention-free counterpart: it can achieve comparable results with far less parameters and training instances (Jean et al., 2015). This superiority in efficiency comes mainly from the mechanism of dynamic alignment, which avoids the need to represent the entire source sentence with a fixed-length vector (Sutskever et al., 2014).

However, conventional attention model is conducted on the representation of source sentence (fixed after generated) only with reading operation (Bahdanau et al., 2015; Luong et al., 2015a). This may let the decoder tend to ignore past attention information, and lead to over-translation and under-translation (Tu et al., 2016). To address this problem, Tu et al. (2016) proposed to maintain tag vec-

[*] The majority of this work was completed when the first author studied at Institute of Computing Technology, Chinese Academy of Sciences.

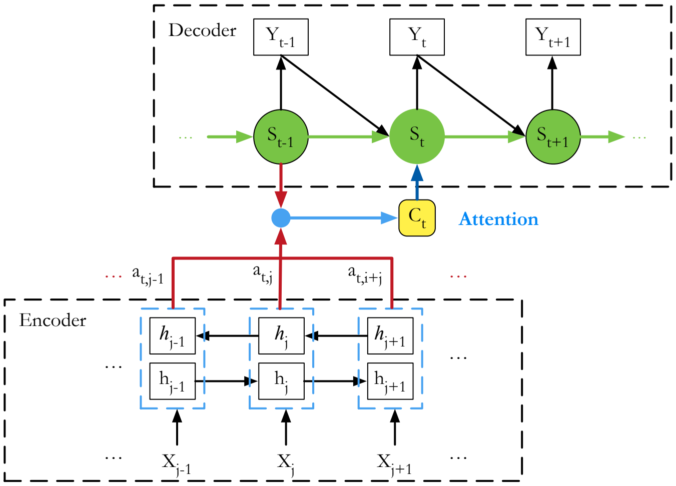

Figure 1: Illustration for attention-based NMT.

tors in source representation to keep track of the attention history, which encourages the attention-based NMT system to consider more untranslated source words. Inspired by neural turing machines (Graves et al., 2014), we propose INTERACTIVE ATTENTION model from the perspective of memory reading-writing, which provides a conceptually simpler and practically more effective mechanism for attention-based NMT. The NMT with INTERACTIVE ATTENTION is called NMT_{IA}, which can keep track of the interaction history with the representation of source sentence by both reading and writing operations during translation. This interactive mechanism may be helpful for the decoder to automatically distinguish which parts have been translated and which parts are under-translated.

We test the efficacy of NMT_{IA} on NIST Chinese-English translation task. Experiment results show that NMT_{IA} can significantly outperform both the conventional attention-based NMT baseline (Bahdanau et al., 2015) and coverage models (Tu et al., 2016). And neural machine translator with our INTERACTIVE ATTENTION can outperform the open source attention-based NMT system Groundhog by 4.22 BLEU points and the open source phrase-based system Moses by 3.94 BLEU points.

RoadMap: In the remainder of this paper, we will start with a brief overview of attention-based neural machine translation in Section 2. Then in Section 3, we will detail the INTERACTIVE ATTENTION-based NMT (NMT_{IA}). In Section 4, we report our empirical study of NMT_{IA} on a Chinese-English translation task, followed by Section 5 and 6 for related work and conclusion.

2 Background

Our work is built upon the attention-based NMT (Bahdanau et al., 2015), which takes a sequence of vector representations of the source sentence generated by a RNN or bi-directional RNN as input, and then jointly learns to align and translate by reading the vector representations during translation with a RNN decoder. Therefore, we take an overview of the attention-based NMT in this section before detail the NMT_{IA} in next section.

2.1 Attention-based Neural Machine Translation

Figure 1 shows the framework of attention-based NMT. Formally, given an input source sequence $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$ and the previously generated target sequence $\mathbf{y_{<t}} = \{y_1, y_2, \cdots, y_{t-1}\}$, the probability of the next target word y_t is

$$p(y_t|\mathbf{y_{<t}}, \mathbf{x}) = softmax(f(\mathbf{c}_t, y_{t-1}, \mathbf{s}_t)) \tag{1}$$

where $f(\cdot)$ is a non-linear function, and $\mathbf{s}_t$ is the state of decoder RNN at time step t which is calculated as

$$\mathbf{s}_t = g(\mathbf{s}_{t-1}, y_{t-1}, \mathbf{c}_t) \tag{2}$$

where $g(\cdot)$ can be any activation function, here we adopt a more sophisticated dynamic operator as in Gated Recurrent Unit (GRU) (Cho et al., 2014). In the remainder of the paper, we will also use GRU to stand for the operator. And $\mathbf{c}_t$ is a distinct source representation for time t, calculated as a weighted sum of the source annotations:

$$\mathbf{c}_t = \sum_{j=1}^{N} a_{t,j}\mathbf{h}_j \tag{3}$$

Formally, $\mathbf{h}_j = [\overrightarrow{\mathbf{h}}_j^T, \overleftarrow{\mathbf{h}}_j^T]^T$ is the annotation of x_j, which is computed by a bi-directional RNN (Schuster and Paliwal, 1997) with GRU and contains information about the whole input sequence with a strong focus on the parts surrounding x_j. And its weight $a_{t,j}$ is computed by

$$a_{t,j} = \frac{exp(e_{t,j})}{\sum_{k=1}^{N} exp(e_{t,k})} \tag{4}$$

where $e_{t,j} = \mathbf{v}_a^T tanh(\mathbf{W}_a\mathbf{s}_{t-1} + \mathbf{U}_a\mathbf{h}_j)$ scores how well $\mathbf{s}_{t-1}$ and $\mathbf{h}_j$ match. This is called automatic alignment (Bahdanau et al., 2015) or attention model (Luong et al., 2015a), but it is essentially reading with content-based addressing defined in (Graves et al., 2014). With the attention model, it releases the need to summarize the entire sentence with a single fixed-length vector (Sutskever et al., 2014; Cho et al., 2014). Instead, it lets the decoding network focus on one particular segment in source sentence at one moment, and therefore better resolution.

2.2 Improved Attention Model

The alignment model $a_{t,j}$ scores how well the output at position t matches the inputs around position j based on $\mathbf{s}_{t-1}$ and $\mathbf{h}_j$. Intuitively, it should be beneficial to directly exploit the information of y_{t-1} when reading from the representation of source sentence, which is not implemented in the original attention-based NMT (Bahdanau et al., 2015). As illustrated in Figure 2, we add this implementation into the attention model, inspired by the latest implementation of attention-based NMT[1]. This kind of attention model can find a more effective alignment path by using both previous hidden state $\mathbf{s}_{t-1}$ and the previous context word y_{t-1}. Then, the calculation of $e(t, j)$ becomes

$$e_{t,j} = \mathbf{v}_a^T tanh(\mathbf{W}_a\tilde{s}_{t-1} + \mathbf{U}_a\mathbf{h}_j) \tag{5}$$

where $\tilde{\mathbf{s}}_{t-1} = \mathbf{GRU}(\mathbf{s}_{t-1}, \mathbf{e}_{y_{t-1}})$ is an intermediate state tailored for reading from the representation of source sentence with the information of y_{t-1} (its word embedding being $\mathbf{e}_{y_{t-1}}$) added. And the calculation of update-state $\mathbf{s_t}$ becomes

$$\mathbf{s}_t = \mathbf{GRU}(\tilde{\mathbf{s}}_{t-1}, \mathbf{c}_t) \tag{6}$$

[1] https://github.com/nyu-dl/dl4mt-tutorial/tree/master/session2

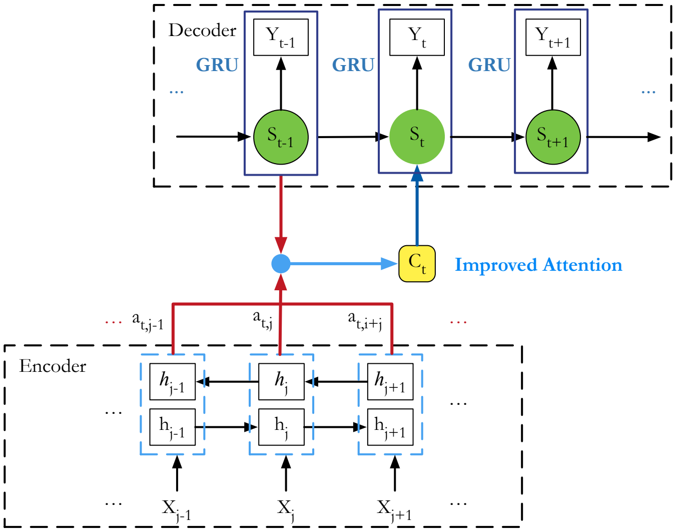

Figure 2: Illustration for improved attention model of NMT.

3 Interactive Attention

In this section, we will elaborate on the proposed INTERACTIVE ATTENTION-based NMT, called NMT$_{\mathsf{IA}}$. Figure 3 shows the framework of NMT$_{\mathsf{IA}}$ with two rounds of interactive read-write operations (indicated by the yellow and red arrows respectively), which adopts the same prediction model (Eq. 1) with improved attention-based NMT. With annotations $\mathbf{H}=\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$ of the source sentence $\mathbf{x}=\{x_1, x_2, \cdots, x_N\}$, we take $\mathbf{H}$ as a memory, which contains N cells with the jth cell being $\mathbf{h}_j$. As illustrated in Figure 3, INTERACTIVE ATTENTION in NMT$_{\mathsf{IA}}$ contains two key parts at each time step t: 1) attentive reading from $\mathbf{H}$, and 2) attentive writing to $\mathbf{H}$. Since the content in $\mathbf{H}$ changes with time, we will add time stamp on $\mathbf{H}$ (hence $\mathbf{H}^{(t)}$) and its cells (hence $\mathbf{h}_j^{(t)}$).

At time t, the state $\mathbf{s}_{t-1}$ first meets the prediction y_{t-1} to form an “intermediate” state $\tilde{\mathbf{s}}_{t-1}$, which can be calculated as follows

$$\tilde{\mathbf{s}}_{t-1} = \mathbf{GRU}(\mathbf{s}_{t-1}, \mathbf{e}_{y_{t-1}}) \tag{7}$$

where $\mathbf{e}_{y_{t-1}}$ is the word-embedding associated with the previous prediction word y_{t-1}. This “intermediate” state $\tilde{\mathbf{s}}_{t-1}$ is used to read the source memory $\mathbf{H}^{(t-1)}$

$$\mathbf{c}_t = \mathbf{Read}(\tilde{\mathbf{s}}_{t-1}, \mathbf{H}^{(t-1)}) \tag{8}$$

After that, $\tilde{\mathbf{s}}_{t-1}$ is combined with $\mathbf{c}_t$ to update the new state

$$\mathbf{s}_t = \mathbf{GRU}(\tilde{\mathbf{s}}_{t-1}, \mathbf{c}_t) \tag{9}$$

Finally, the new state $\mathbf{s}_t$ is used to update the source memory by writing to it to finish the interaction in a round of state-update

$$\mathbf{H}^{(t)} = \mathbf{Write}(\mathbf{s}_t, \mathbf{H}^{(t-1)}) \tag{10}$$

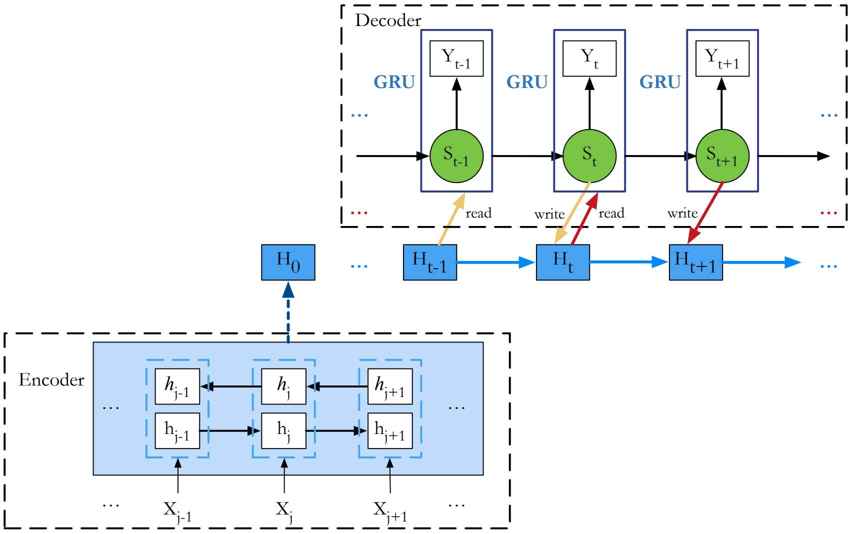

Figure 3: Illustration for the NMT_{IA}. The yellow and red arrows indicate two rounds of interactive read-write operations.

The details of **Read** and **Write** in Eq. 8 and 10 will be described later in next section.

From the whole framework of NMT_{IA}, we can see that the new attention mechanism can timely update the representation of source sentence along with the update-chain of the decoder RNN state. This may let the decoder keep track of the attention history during translation. Clearly, Interactive Attention can subsume the coverage models in (Tu et al., 2016) as special cases while conceptually simpler. Moreover, with the attentive writing, Interactive Attention potentially can modify and add more on the source representation than just history of attention, and is therefore a more powerful model for machine translation, as empirically verified in Section 4.

3.1 Read and Write of Interactive Attention

Attentive Read Formally, $\mathbf{H}^{(t')} \in \mathbb{R}^{n \times m}$ is the memory in time t' after the decoder RNN state update, where n is the number of memory cells and m is the dimension of vector in each cell. Before the state $\mathbf{s}$ update at time t, the output of reading $\mathbf{c}_t$ is given by

$$\mathbf{c}_t = \sum_{j=1}^{n} \mathbf{w}_t^R(j)\mathbf{h}_j^{(t-1)} \tag{11}$$

where $\mathbf{w}_t^R \in \mathbb{R}^n$ specifies the normalized weights assigned to the cells in $\mathbf{H}^{(t-1)}$. We can use content-based addressing to determine $\mathbf{w}_t^R$ as described in (Graves et al., 2014) or (quite similarly) use the reading mechanism such as the attention model in Section 2. In this paper, we adopt the latter one.[2]

Attentive Write Inspired by the writing operation of neural turing machines (Graves et al., 2014), we define two types of operation on writing to the memory: Forget and Update. Forget is similar

[2] Wang et al. (2016) verified the former one for the read operation on the external memory.

to the forget gate in GRU, which determines the content to be removed from memory cells. More specifically, the vector $\mathbf{F}_t \in \mathbb{R}^m$ specifies the values to be forgotten or removed on each dimension in memory cells, which is then assigned to each cell through normalized weights $\mathbf{w}_t^W$. Formally, the memory ("intermediate") after FORGET operation is given by

$$\tilde{\mathbf{h}}_i^{(t)} = \mathbf{h}_i^{(t-1)}(1 - \mathbf{w}_t^W(i) \cdot \mathbf{F}_t), \qquad i = 1, 2, \cdots, n \tag{12}$$

where

- $\mathbf{F}_t = \sigma(\mathbf{W}_F, s_t)$ is parameterized with $\mathbf{W}_F \in \mathbb{R}^{m \times m}$, and σ stands for the $Sigmoid$ activation function;
- $\mathbf{w}_t^W \in \mathbb{R}^n$ specifies the normalized weights assigned to the cells in $\mathbf{H}^{(t)}$, and $\mathbf{w}_t^W(i)$ specifies the weight associated with the ith cell in the same parametric form as $\mathbf{w}_t^R$.

UPDATE is similar to the update gate in GRU, deciding how much current information should be written to the memory as the added content

$$\mathbf{h}_i^{(t)} = \tilde{\mathbf{h}}_i^{(t)} + \mathbf{w}_t^W(i) \cdot \mathbf{U}_t, \qquad i = 1, 2, \cdots, n \tag{13}$$

where $\mathbf{U}_t = \sigma(\mathbf{W}_U, \mathbf{s}_t)$ is parameterized with $\mathbf{W}_U \in \mathbb{R}^{m \times m}$, and $\mathbf{U}_t \in \mathbb{R}^m$. In our experiments, the weights for reading (i.e., $\mathbf{w}_t^R$) and writing (i.e., $\mathbf{w}_t^W$) at time t are shared when conducting interaction with the source memory.

3.2 Optimization

The parameters to be optimized include the embedding of words on source and target languages, the parameters for the encoder, the decoder and other operations of NMT_{IA}. The optimization is conducted via the standard back-propagation (BP) aiming to maximize the likelihood of the target sequence. In practice, we use the standard stochastic gradient descent (SGD) and mini-batch with learning rate controlled by AdaDelta (Zeiler, 2012).

4 Experiments

We report our empirical study of NMT_{IA} on Chinese-to-English translation task in this section. The experiments are designed to answer the following questions:

- Can NMT_{IA} achieve significant improvements over the conventional attention-based NMT?
- Can NMT_{IA} outperform the attention-based NMT with coverage model (Tu et al., 2016)?

4.1 Data and Metric

Our training data consist of 1.25M sentence pairs extracted from LDC corpora[3], with 27.9M Chinese words and 34.5M English words respectively. We choose NIST 2002 (MT02) dataset as our development set, which is used to monitor the training process and decide the early stop condition. And the NIST 2003 (MT03), 2004 (MT04), 2005 (MT05), 2006 (MT06) datasets are used as our test sets. The numbers of sentences in NIST MT02, MT03, MT04, MT05 and MT06 are 878, 919, 1788, 1082, and 1664 respectively. We use the case-insensitive 4-gram NIST BLEU[4] as our evaluation metric, with statistical significance test (*sign-test* (Collins et al., 2005)) between the proposed models and the baselines.

[3] The corpora include LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

[4] `ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl`

SYSTEMS	MT03	MT04	MT05	MT06	AVERAGE
Moses	31.61	33.48	30.75	31.07	31.73
Groundhog	30.96	33.09	30.61	31.12	31.45
RNNsearch*	33.42	36.04	33.60	32.24	33.83
NMT$_{\mathsf{IA}}$	35.09*	37.73*	35.53*	34.32*	35.67

Table 1: BLEU-4 scores (%) of the phrase-based SMT system (Moses), NMT baselines: Groundhog and RNNsearch* (our implementation of improved attention model as described in Section 2.2), and our INTERACTIVE ATTENTION model (NMT$_{\mathsf{IA}}$). The "*" indicates that the results are significantly ($p<0.01$) better than those of all the baseline systems.

4.2 Training Details

In training the neural networks, we limit the source and target vocabulary to the most frequent 30K words for both Chinese and English, covering approximately 97.7% and 99.3% of two corpus respectively. All the out-of-vocabulary words are mapped to a special token UNK. We initialize the recurrent weight matrices as random orthogonal matrices. All the bias vectors are initialized to zero. For other parameters, we initialize them by sampling each element from the Gaussian distribution of mean 0 and variance 0.01^2. The parameters are updated by SGD and mini-batch (size 80) with learning rate controlled by AdaDelta (Zeiler, 2012) ($\epsilon = 1e^{-6}$ and $\rho = 0.95$). We train the NMT systems with the sentences of length up to 50 words in training data, and set the dimension of word embedding to 620 and the size of the hidden layer to 1000, following the settings in (Bahdanau et al., 2015). We also use dropout for our baseline NMT systems and NMT$_{\mathsf{IA}}$ to avoid over-fitting (Hinton et al., 2012). In our experiments, dropout was applied on the output layer with dropout rate setting to 0.5.

Inspired by the effort on easing the training of very deep architectures (Hinton and Salakhutdinov, 2006), we use a simple pre-training strategy to train our NMT$_{\mathsf{IA}}$. First we train a regular attention-based NMT model (Bahdanau et al., 2015). Then we use the trained NMT model to initialize the parameters of NMT$_{\mathsf{IA}}$ except for those related to the operations of INTERACTIVE ATTENTION. After that, we fine-tune all the parameters of NMT$_{\mathsf{IA}}$.

4.3 Comparison Systems

We compare our NMT$_{\mathsf{IA}}$ with four systems:

- **Moses** (Koehn et al., 2007): an open source phrase-based translation system[5] with default configuration. The word alignments are obtained with GIZA++ (Och and Ney, 2003) on the training corpora in both directions, using the "grow-diag-final-and" balance strategy (Koehn et al., 2003). The 4-gram language model with modified Kneser-Ney smoothing is trained on the target portion of training data with the SRILM toolkit (Stolcke and others, 2002),

- **Groundhog**: an open source NMT system[6] implemented with the conventional attention model (Bahdanau et al., 2015).

- **RNNsearch***: our in-house implementation of NMT system with the improved conventional attention model as described in Section 2.2.

- **Coverage Model**: state-of-the-art variants of attention-based NMT model (Tu et al., 2016) which improve the attention mechanism through modeling a soft coverage on the source representation

[5] http://www.statmt.org/moses/

[6] https://github.com/lisa-groundhog/GroundHog

Systems	MT03	MT04	MT05	MT06	Average
RNNsearch*-80	33.34	37.10	33.38	33.70	34.38
NN-Cover-80	33.69	38.05	35.01	34.83	35.40
NMT$_{\text{IA}}$-80	35.69*+	39.24*+	35.74*+	35.10*	36.44

Table 2: BLEU-4 scores (%) of the conventional attention-based model (RNNsearch*-80), the neural network based coverage model (NN-Cover-80) (Tu et al., 2016) and our Interactive Attention model (NMT$_{\text{IA}}$-80). "-80" means the models are trained with the sentences of length up to 80 words, which is consistent with the setting in (Tu et al., 2016). The "*" and "+" denote that the results are significantly ($p<0.01$) better than those of RNNsearch*-80 and NN-Cover-80 respectively.

by maintain a coverage vector to keep track of the attention history during translation.

4.4 Main Results

The main results of different models are given in Table 1. Before proceeding to more detailed comparisons, we first observe that

- RNNsearch* outperforms Groundhog, which is implemented with the conventional attention model as described in Section 2.1, by 2.38 BLEU points averagely on four test sets;
- RNNsearch* only exploit sentences of length up to 50 words with 30K vocabulary, but can achieve averagely 2.10 BLEU points higher than the open source phrase-based system Moses, which is trained with full training data.

Clearly from Table 1, NMT$_{\text{IA}}$ can achieve significant improvements over RNNsearch* by 1.84 BLEU points averagely on four test sets. We conjecture it is because our Interactive Attention mechanism can keep track of the interaction history between the decoder and the representation of source sentence during translation, which may be helpful for the decoder to automatically distinguish which parts have been translated and which parts are under-translated.

4.5 Interactive Attention Vs. Coverage Model

Tu et al. (2016) proposed two coverage models to let the NMT system to consider more about untranslated source words. Basically, they maintain a coverage vector for each hidden state for source to keep track of the attention history and feed the coverage vector to the attention model to help adjust future attention. Although we do not maintain a coverage vector, our Interactive Attention can potentially do similar things, therefore subsuming coverage models as special cases. We hence compare our Interactive Attention model with the coverage model in (Tu et al., 2016). There are two coverage models proposed in (Tu et al., 2016), including linguistic coverage model and neural network based coverage model (NN-Cover). Since the neural network based coverage model generally yields better results, we mainly compare with the neural network based coverage model. Although the coverage models are originally implemented on Groundhog in (Tu et al., 2016), they can be easily adapted to the "RNNsearch*". Following the setting in (Tu et al., 2016), we conduct the comparison with the training sentences of length up to 80 words. Clearly from Table 2, our NMT$_{\text{IA}}$-80 outperforms the NN-Cover-80 by +1.04 BLEU scores averagely on four test sets.

A more detailed comparison between conventional attention model (RNNsearch*-80), neural network based coverage model (NN-Cover-80) (Tu et al., 2016) and NMT$_{\text{IA}}$-80 suggests that our NMT$_{\text{IA}}$-80 is quite consistent on outperforming the conventional attention model and the coverage model. Figure 4 shows the BLEU scores of generated translations on the test sets with respect to the length of the

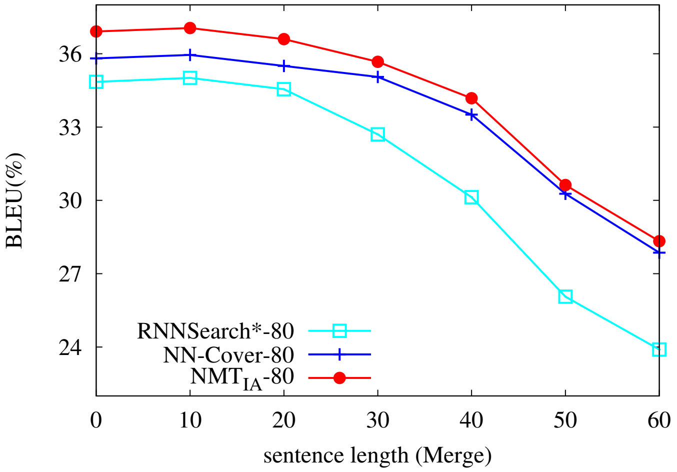

Figure 4: The BLEU-4 scores (%) of generated translations on the merged four test sets with respect to the lengths of source sentences. The numbers on X-axis of the figure stand for sentences *longer than* the corresponding length, e.g., 40 for source sentences with > 40 words.

source sentences. In particular, we test the BLEU scores on sentences longer than $\{$0, 10, 20, 30, 40, 50, 60$\}$ in the merged test set of MT03, MT04, MT05 and MT06. Clearly, on sentences with different length, NMT$_{\text{IA}}$-80 always yields consistently higher BLEU scores than the conventional attention-based NMT and the enhanced version with the neural network based coverage model. We conjecture that with the attentive writing (described in Section 3.1), INTERACTIVE ATTENTION potentially can modify and add more on the source representation than just history of attention, and is therefore a more powerful model for machine translation.

We also provide some actual translation examples (see Appendix) to show that our INTERACTIVE ATTENTION can get better performance then baselines, especially on solving under-translation problem. We think the interactive mechanism of NMT$_{\text{IA}}$ is helpful for the decoder to automatically distinguish which parts have been translated and which parts are under-translated.

5 Related Work

Our work is related to recent works that focus on improving attention models (Luong et al., 2015a; Cohn et al., 2016; Feng et al., 2016). Luong et al. (2015a) proposed to use global and local attention models to improve translation performance. They use a global one to attend to all source words and a local one to look at a subset of source words at a time. Cohn et al. (2016) extended the attention-based NMT to include structural biases from word-based alignment models, which achieved improvements across several language pairs. Feng et al. (2016) added implicit distortion and fertility models to attention-based NMT to achieve translation improvements. These works are different with our INTERACTIVE ATTENTION approach, as we use a rather generic attentive reading while at the same time performing attentive writing.

Our work is inspired by recent efforts on attaching an external memory to neural networks, such as neural turing machines (Graves et al., 2014), memory networks (Weston et al., 2014; Meng et al., 2015) and exploiting an external memory (Tang et al., 2016; Wang et al., 2016) during translation. Tang et al. (2016) exploited a phrase memory for NMT, which stores phrase pairs in symbolic form. They let the decoder utilize a mixture of word-generating and phrase-generating component, to generate a sequence of multiple words all at once. Wang et al. (2016) extended the NMT decoder by maintaining an external memory, which is operated by reading and writing operations of neural turing machines (Graves et al., 2014), while keeping a read-only copy of the original source annotations along side the "read-write" memory. These powerful extensions have been verified on Chinese-English translation tasks. Our INTERACTIVE ATTENTION is different from previous works. We take the annotations of source sentence as a memory instead of using an external memory, and we design a mechanism to directly read from and write to it during translation. Therefore, the original source annotations are not accessible in later steps. More specially, our model inherited the notation and some simple operations for writing from (Graves et al., 2014), while NMT_{IA} extends it to "unbounded" memory for representing the source. In addition, although the read-write operations in INTERACTIVE ATTENTION are not exactly the same with those in (Graves et al., 2014; Wang et al., 2016), our model can also achieve good performance.

6 Conclusion

We propose a simple yet effective INTERACTIVE ATTENTION approach, which models the interaction between the decoder and the representation of source sentence during translation by using reading and writing operations. Our empirical study on Chinese-English translation shows that INTERACTIVE ATTENTION can significantly improve the performance of NMT.

Acknowledgements

Liu is partially supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre at Dublin City University. We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*, pages 1724–1734.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California, June.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*, pages 531–540.

Shi Feng, Shujie Liu, Mu Li, and Ming Zhou. 2016. Implicit distortion and fertility models for attention-based encoder-decoder NMT model. *CoRR*, abs/1601.03317.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, July.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL on interactive poster and demonstration sessions*, pages 177–180, Prague, Czech Republic, June.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *Proceedings of IJCAI*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July.

Fandong Meng, Zhengdong Lu, Zhaopeng Tu, Hang Li, and Qun Liu. 2015. Neural transformation machine: A new architecture for sequence-to-sequence learning. *CoRR*, abs/1506.06442.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of ACL*, pages 1683–1692, Berlin, Germany, August.

Andreas Stolcke et al. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pages 901–904.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Yaohua Tang, Fandong Meng, Zhengdong Lu, Hang Li, and Philip L. H. Yu. 2016. Neural machine translation with external phrase memory. *CoRR*, abs/1606.01792.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of ACL*, pages 76–85, Berlin, Germany, August. Association for Computational Linguistics.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Memory-enhanced decoder for neural machine translation. In *Proceedings of EMNLP*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *CoRR*, abs/1410.3916.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *CoRR*, abs/1606.04199.

APPENDIX: Actual Translation Examples

In appendix we give some example translations from RNNsearch*-80, NN-Cover-80 and NMT$_{IA}$-80, and compare them against the reference. We highlight some correct translation segments (or under-translated by baseline systems) in blue color and wrong ones in red color.

Example Translations

src	北韩 指称 这 项 核子 僵局 系 仅 涉及 美国 的 双边 议题 ， 其他 国家 进行 干预 只 会 使 问题 复杂化 。
ref	North Korea said the nuclear stalemate is a bilateral topic of discussion with United States only. The interference of other countries will only complicate the issue.
RNNsearch*-80	north korea claimed that the nuclear stalemate was only involved in bilateral issues in the united states , and other countries will find it more complicated .
NN-Cover-80	the north korea said that it had only involved bilateral talks in the united states and other countries would interfere with the issue .
NMT$_{IA}$ -80	north korea claimed that this nuclear stalemate was only related to the us bilateral agenda , and interference in other countries could only complicate the problem .

src	平壤 采取 上述 行动 之后 四 天 ， 联合国 安全 理事会 的 五 个 常任 理事国 都 为 此 一 危机 采取 预防性 外交 行动 。
ref	Four days after Pyongyang made the above move, five permanent members of the UN Security Council have all taken preventive diplomatic actions on this crisis.
RNNsearch*-80	pyongyang has taken these actions four days ago , and the five permanent members of the un security council have taken precautions against this crisis .
NN-Cover-80	in a four - day operation , the five permanent members of the un security council have taken preventive diplomatic actions for the crisis .
NMT$_{IA}$ -80	in the four days after pyongyang took the above action , the five permanent members of the un security council have taken preventive diplomatic actions for this crisis .

src	菲律宾 政府 原本 计划 于 本 月 稍后 就 恢复 正式 和谈 一 事 与 菲共 举行 初步 磋谈 。
ref	The Philippine government originally planned to hold preliminary discussions with the Philippine communists on the resumption of formal peace talks later this month.
RNNsearch*-80	the philippine government originally planned to hold a preliminary meeting with <UNK> on friday .
NN-Cover-80	the philippine government plans to resume formal peace talks with <UNK> later this month .
NMT$_{IA}$ -80	the philippine government originally planned to hold a preliminary discussion on the resumption of formal peace talks later this month .

src	他 表示 ： 「 显然 的 ， 我们 需要 做 的 第一 件 事情 是 将 伊拉克 安全 部队 国际化 ， 另 一 件 事情 是 将 过渡 政府 交给 联合国 。 」
ref	He said: "Obviously, the first thing we need to do is to internationalize the security force in Iraq. The other thing is to turn the transitional government over to the United Nations."
RNNsearch*-80	he said : " obviously , the first thing we need to do is to <UNK> iraqi security forces to the united nations . "
NN-Cover-80	he said : " obviously , we need the first thing to internationalize the security forces in iraq , and another thing is to hand over the transitional government to the united nations . "
NMT$_{IA}$ -80	he said : " obviously , the first thing we need to do is to internationalize the security forces in iraq , and the other is to send the transitional government to the united nations . "