Neural Machine Translation with Supervised Attention

Lemao Liu, Masao Utiyama, Andrew Finch and Eiichiro Sumita
National Institute of Information and Communications Technology (NICT)
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, Japan
{lmliu,first.last}@nict.go.jp

Abstract

The attention mechanism is appealing for neural machine translation, since it is able to dynamically encode a source sentence by generating a alignment between a target word and source words. Unfortunately, it has been proved to be worse than conventional alignment models in alignment accuracy. In this paper, we analyze and explain this issue from the point view of reordering, and propose a supervised attention which is learned with guidance from conventional alignment models. Experiments on two Chinese-to-English translation tasks show that the supervised attention mechanism yields better alignments leading to substantial gains over the standard attention based NMT.

1 Introduction

Neural Machine Translation (NMT) has achieved great successes on machine translation tasks recently (Bahdanau et al., 2015; Sutskever et al., 2015). Generally, it relies on a recurrent neural network under the Encode-Decode framework: it firstly encodes a source sentence into context vectors and then generates its translation token-by-token, selecting from the target vocabulary. Among different variants of NMT, attention based NMT, which is the focus of this paper,[1] is attracting increasing interests in the community (Bahdanau et al., 2015; Luong et al., 2015). One of its advantages is that it is able to dynamically make use of the encoded context through an attention mechanism thereby allowing the use of fewer hidden layers while still maintaining high levels of translation performance.

An attention mechanism is designed to predict the alignment of a target word with respect to source words (Bahdanau et al., 2015). In order to facilitate incremental decoding, it tries to make this alignment prediction without the information about the target word itself, and thus this attention can be considered to be a form of a reordering model (see §2 for more details). In contrast, conventional alignment models are able to use the target word to infer its alignments (Och and Ney, 2000; Dyer et al., 2013; Liu and Sun, 2015), and as a result there is a substantial gap in quality between the alignments derived by this attention based NMT and conventional alignment models (54 VS 30 in terms of AER for Chinese-to-English as reported in (Cheng et al., 2016)). This discrepancy might be an indication that the potential of attention-based NMT is limited. In addition, the attention in NMT is learned in an unsupervised manner without explicit prior knowledge about alignment.[2] However, in conventional statistical machine translation (SMT), it is standard practice to learn reordering models in a supervised manner with the guidance from conventional alignment models (Xiong et al., 2006; Koehn et al., 2007; Bisazza and Federico, 2016).

Inspired by the supervised reordering in conventional SMT, in this paper, we propose a *Supervised Attention* based NMT (SA-NMT) model. Specifically, similar to conventional SMT, we first run off-the-shelf aligners (GIZA++ (Och and Ney, 2000) or fast_align (Dyer et al., 2013) etc.) to obtain the alignment of the bilingual training corpus in advance. Then, treating this alignment result as the supervision of attention, we jointly learn attention and translation, both in supervised manners. Since the

[1]Throughout this paper, without the special statement, NMT means attention-based NMT.

[2]We do agree that NMT is a supervised model with respect to translation rather than reordering.

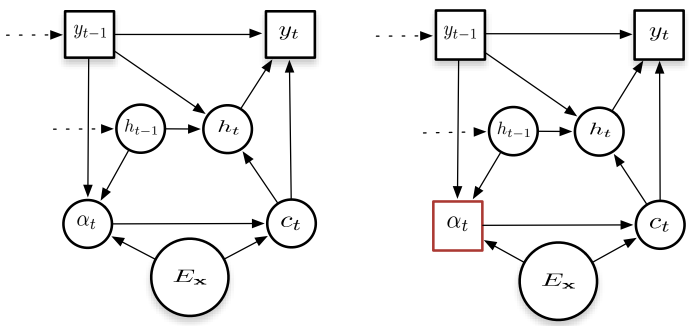
Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3093–3102, Osaka, Japan, December 11-17 2016.

conventional aligners delivers higher quality alignment, it is expected that the alignment in the supervised attention NMT will be improved leading to better end-to-end translation performance. One advantage of the proposed SA-NMT is that it implements the supervision of attention as a regularization in the joint training objective (§3.2). Furthermore, since the attention variables lies in the middle of the entire network architecture rather than the top (as the translation variables (see Figure 1(b)), it serves to mitigate the vanishing gradient problem during the back-propagation, by adding supervision into the intermediate layers in the network (Szegedy et al., 2015).

This paper makes the following contributions:

- It revisits the attention model from the point view of reordering (§2), and propose a supervised attention for NMT that is supervised by statistical alignment models (§3). The proposed approach is simple and easy to be implemented, and it is generally applicable to any attention-based NMT models, although in this case it is implemented on top of the model in (Bahdanau et al., 2015).
- On two Chinese-to-English translation tasks, it empirically shows that the proposed approach gives rise to improved performance (§4): on a large scale task, it outperforms three baselines including a state-of-the-art Moses, and leads to improvements of up to 2.5 BLEU points over the strongest one in this paper; on a low resource task, it even obtains about 5 BLEU points over the attention based NMT system on which is it based.

2 Revisiting Neural Machine Translation



(a) NMT (b) SA-NMT

Figure 1: One slice of the computational graphs for both (a) NMT and (b) SA-NMT. Circles denote the hidden variables; while squares denote the observable variables, which receive supervision during training. The difference (marked in red) in (b) regarding to (a) is treating α_t as an observable variable instead of a hidden variable.

Suppose $\mathbf{x} = \langle x_1, x_2, \cdots, x_m \rangle$ denotes a source sentence, $\mathbf{y} = \langle y_1, y_2, \cdots, y_n \rangle$ a target sentence. In addition, let $x_{<t} = \langle x_1, x_2, \cdots, x_{t-1} \rangle$ denote a prefix of $\mathbf{x}$. Neural Machine Translation (NMT) directly maps a source sentence into a target under an encode-decode framework. In the encoding stage, it uses two bidirectional recurrent neural networks to encode $\mathbf{x}$ into a sequence of vectors $E_{\mathbf{x}} = \langle E_{x_1}, E_{x_2}, \cdots, E_{x_m} \rangle$, with E_{x_i} representing the concatenation of two vectors for i_{th} source word from two directional RNNs. In the decoding stage, it generates the target translation from the conditional probability over the pair of sequences $\mathbf{x}$ and $\mathbf{y}$ via a recurrent neural network parametrized by θ as follows:

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \prod_{t=1}^{n} p(y_t \mid y_{<t}, E_{\mathbf{x}}) = \prod_{t=1}^{n} \mathrm{softmax}\big(g(y_{t-1}, h_t, c_t)\big)[y_t] \tag{1}$$

where h_t and c_t respectively denote an RNN hidden state (i.e. a vector) and a context vector at timestep t; g is a transformation function mapping into a vector with dimension of the target vocabulary size; and $[i]$ denotes the i_{th} component of a vector.[3] Furthermore, $h_t = f(h_{t-1}, y_{t-1}, c_t)$ is defined by an activation

[3]In that sense, y_t in Eq.(1) also denotes the index of this word in its vocabulary.

function, i.e. a Gated Recurrent Unit (Chung et al., 2014); and the context vector c_t is a dynamical source representation at timestep t, and calculated as the weighted sum of source encodings $E_{\mathbf{x}}$, i.e. $c_t = \alpha_t^\top E_{\mathbf{x}}$. Here the weight α_t implements an attention mechanism, and $\alpha_{t,i}$ is the alignment probability of y_t being aligned to x_i. α_t is derived through a feedforward neural network a as follows:

$$\alpha_t = a(y_{t-1}, h_{t-1}, E_{\mathbf{x}}) \tag{2}$$

where a consists of two layers, the top one being a softmax layer. We skip the detailed definitions of a together with $E_{\mathbf{x}}$, f and g, and refer the readers to (Bahdanau et al., 2015) instead.[4] Figure 1(a) shows one slice of computational graph for NMT definition at time step t.

To train NMT, the following negative log-likelyhood is minimized:

$$-\sum_i \log p(\mathbf{y}^i \mid \mathbf{x}^i; \theta) \tag{3}$$

where $\langle \mathbf{x}^i, \mathbf{y}^i \rangle$ is a bilingual sentence pair from a given training corpus, $p(\mathbf{y}^i \mid \mathbf{x}^i; \theta)$ is as defined in Eq.(1). Note that even though the training is conducted in a supervised manner with respect to translation, i.e., $\mathbf{y}$ are observable in Figure 1(a), the attention is learned in a unsupervised manner, since α is hidden.

In Eq.(2), α_t is defined only on y_{t-1}, h_{t-1} and $E_{\mathbf{x}}$ but not on the target word y_t, as y_t is unknown at the current timestep $t-1$ during the testing. Therefore, at timestep $t-1$, NMT firstly tries to calculate α_t, through which NMT figures out those source words will be translated next, even though the next target word y_t is unavailable. From this point of view, the attention mechanism plays a role in reordering and thus can be considered as a reordering model. Unlike this attention model, conventional alignment models define the alignment α directly over $\mathbf{x}$ and $\mathbf{y}$ as follows:

$$p(\alpha \mid \mathbf{x}, \mathbf{y}) = \frac{\exp(F(\mathbf{x}, \mathbf{y}, \alpha))}{\sum_{\alpha'} \exp(F(\mathbf{x}, \mathbf{y}, \alpha'))}$$

where F denotes a feature function over a pair of sentences $\mathbf{x}$ and $\mathbf{y}$ together with their word alignment α, and it is either a log-probability $\log p(\mathbf{y}, \alpha \mid \mathbf{x})$ for a generative model like IBM models (Brown et al., 1993) or a well-designed feature function for discriminative models (Liu and Sun, 2015). In order to infer α_t, alignment models can readily use the entire $\mathbf{y}$, of course including y_t as well, thereby they can model the alignment between $\mathbf{x}$ and $\mathbf{y}$ more sufficiently. As a result, the attention based NMT might not deliver satisfying alignments, as reported in (Cheng et al., 2016), compared to conventional alignment models. This may be a sign that the potential of attention-based NMT is limited in end-to-end translation.

3 Supervised Attention

In this section, we introduce supervised attention to improve the alignment, which may lead to better translation performance for NMT. [5] Our basic idea is simple: similar to conventional SMT, it firstly uses a conventional aligner to obtain the alignment on the training corpus; then it employs these alignment results as supervision to train the NMT. During testing, decoding proceeds in exactly the same manner as standard NMT, since there is no alignment supervision available for unseen test sentences.

3.1 Preprocessing Alignment Supervision

As described in §2, the attention model outputs a soft alignment α, such that α_t is a normalized probability distribution. In contrast, most aligners are typically oriented to grammar induction for conventional SMT, and they usually output 'hard' alignments, such as (Och and Ney, 2000). They only indicate whether a target word is aligned to a source word or not, and this might not correspond to a distribution for each target word. For example, one target word may align to multiple source words, or no source words at all.

[4] In the original paper, α_t is not explicitly dependent on the y_{t-1} in Eq.(2), but this dependency was explicitly retained in our direct baseline NMT2.

[5] Although the alignment is loosely related to the downstream translation (Liu and Sun, 2015), substantial improvements in alignment usually leads to the improvements in translation as observed in our experiments.

Therefore, we apply the following heuristics to preprocess the hard alignment: if a target word does not align to any source words, we inherit its affiliation from the closest aligned word with preference given to the right, following (Devlin et al., 2014); if a target word is aligned to multiple source words, we assume it aligns to each one evenly. In addition, in the implementation of NMT, there are two special tokens 'eol' added to both source and target sentences. We assume they are aligned to each other. In this way, we can obtain the final supervision of attention, denoted as $\hat{\alpha}$.

3.2 Jointly Supervising Translation and Attention

We propose a soft constraint method to jointly supervise the translation and attention as follows:

$$-\sum_i \log p(\mathbf{y}^i \mid \mathbf{x}^i; \theta) + \lambda \times \Delta(\alpha^i, \hat{\alpha}^i; \theta) \tag{4}$$

where α^i is as defined in Eq. (1), Δ is a loss function that penalizes the disagreement between α^i and $\hat{\alpha}^i$, and $\lambda > 0$ is a hyper-parameter that balances the preference between likelihood and disagreement. In this way, we treat the attention variable α as an observable variable as shown in Figure 1(b), and this is different from the standard NMT as shown in Figure 1(a) in essence. Note that this training objective resembles to that in multi-task learning (Evgeniou and Pontil, 2004). Our supervised attention method has two further advantages: firstly, it is able to alleviate overfitting by means of the λ; and secondly it is easier to address the vanishing gradient problem by adding supervision into the intermediate layers of the entire network (Szegedy et al., 2015), because the supervision of α is more close to $E_{\mathbf{x}}$ than $\mathbf{y}$ as in Figure 1(b).

In order to quantify the disagreement between α^i and $\hat{\alpha}^i$, three different methods are investigated in our experiments:

- *Mean Squared Error* (MSE)

$$\Delta(\alpha^i, \hat{\alpha}^i; \theta) = \sum_m \sum_n \frac{1}{2}\big(\alpha(\theta)^i_{m,n} - \hat{\alpha}^i_{m,n}\big)^2$$

  MSE is widely used as a loss for regression tasks (Lehmann and Casella, 1998), and it directly encourages $\alpha(\theta)^i_{m,n}$ to be equal to $\hat{\alpha}^i_{m,n}$.

- *Multiplication* (MUL)

$$\Delta(\alpha^i, \hat{\alpha}^i; \theta) = -\log\big(\sum_m \sum_n \alpha(\theta)^i_{m,n} \times \hat{\alpha}^i_{m,n}\big)$$

  MUL is particularly designed for agreement in word alignment and it has been shown to be effective (Liang et al., 2006; Cheng et al., 2016). Note that different from those in (Cheng et al., 2016), $\hat{\alpha}$ is not a parametrized variable but a constant in this paper.

- *Cross Entropy* (CE)

$$\Delta(\alpha^i, \hat{\alpha}^i; \theta) = -\sum_m \sum_n \hat{\alpha}^i_{m,n} \times \log \alpha(\theta)^i_{m,n}$$

  Since for each t, $\alpha(\theta)_t$ is a distribution, it is natural to use CE as the metric to evaluate the disagreement (Rubinstein and Kroese, 2004).

4 Experiments

We conducted experiments on two Chinese-to-English translation tasks: one is the NIST task oriented to NEWS domain, which is a large scale task and suitable to NMT; and the other is the speech translation oriented to travel domain, which is a low resource task and thus is very challenging for NMT. We used the case-insensitive BLEU4 to evaluate translation quality and adopted the multi-bleu.perl as its implementation.

Alignment Losses	BLEU
Mean Squared Error (MSE)	39.4
Multiplication (MUL)	39.6
Cross Entropy (CE)	40.0

Table 1: Performance of SA-NMT on development set for different loss functions to supervise the attention in terms of BLEU.

Alignment Methods	BLEU
fast_align	39.6
GIZA++	40.0

Table 2: Comparision of aligners between fast_align and GIZA++ for SA-NMT in terms of BLEU on the development set.

4.1 The Large Scale Translation Task

4.1.1 Preparation

We used the data from the NIST2008 Open Machine Translation Campaign. The training data consisted of 1.8M sentence pairs, the development set was nist02 (878 sentences), and the test sets are were nist05 (1082 sentences), nist06 (1664 sentences) and nist08 (1357 sentences).

We compared the proposed approach with three strong baselines:

- Moses: a phrase-based machine translation system (Koehn et al., 2007);
- NMT1: an attention based NMT (Bahdanau et al., 2015) system at https://github.com/lisa-groundhog/GroundHog;
- NMT2: another implementation of (Bahdanau et al., 2015) at https://github.com/nyu-dl/dl4mt-tutorial.

We developed the proposed approach based on NMT2, and denoted it as **SA-NMT**.

We followed the standard pipeline to run Moses. GIZA++ with grow-diag-final-and was used to build the translation model. We trained a 5-gram target language model on the Gigaword corpus, and used a lexicalized distortion model. All experiments were run with the default settings.

To train NMT1, NMT2 and SA-NMT, we employed the same settings for fair comparison. Specifically, except the stopping iteration which was selected using development data, we used the default settings set out in (Bahdanau et al., 2015) for all NMT-based systems: the dimension of word embedding was 620, the dimension of hidden units was 1000, the batch size was 80, the source and target side vocabulary sizes were 30000, the maximum sequence length was 50, [6] the beam size for decoding was 12, and the optimization was done by Adadelta with all hyper-parameters suggested by (Zeiler, 2012). Particularly for SA-NMT, we employed a conventional word aligner to obtain the word alignment on the training data before training SA-NMT. In this paper, we used two different aligners, which are fast_align and GIZA++. We tuned the hyper-parameter λ to be 0.3 on the development set, to balance the preference between the translation and alignment. Training was conducted on a single Tesla K40 GPU machine. Each update took about 3.0 seconds for both NMT2 and SA-NMT, and 2.4 seconds for NMT1. Roughly, it took about 10 days to NMT2 to finish 300000 updates.

4.1.2 Settings on External Alignments

We implemented three different losses to supervise the attention as described in §3.2. To explore their behaviors on the development set, we employed the GIZA++ to generate the alignment on the training set prior to the training SA-NMT. In Table 1, we can see that MUL is better than MSE. Furthermore, CE performs best among all losses, and thus we adopt it for the following experiments.

[6] This excludes all the sentences longer than 50 words in either source or target side only for NMT systems, but for Moses we use the entire training data.

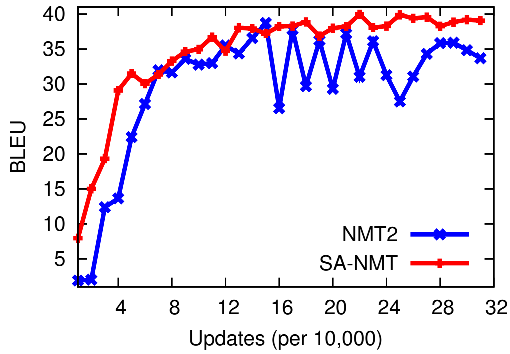


Figure 2: Learning curves of NMT2 and SA-NMT on the development set.

Systems	nist02	nist05	nist06	nist08
Moses	37.1	35.1	33.4	25.9
NMT1	37.8	34.1	34.7	27.4
NMT2	38.7	35.3	36.0	27.8
SA-NMT	40.0*	37.8*	37.6*	29.9*

Table 3: BLEU comparison for large scale translation task. The development set is nist02, and the test sets are nist05,nist06 and nist08. '*' denotes that SA-NMT is significantly better than Moses, NMT1 and NMT2 with $p < 0.01$. Note that Moses is trained with more bilingual sentences and an additional monolingual corpus.

In addition, we also run fast_align to generate alignments as the supervision for SA-NMT and the results were reported in Table 2. We can see that GIZA++ performs slightly better than fast_align and thus we fix the external aligner as GIZA++ in the following experiments.

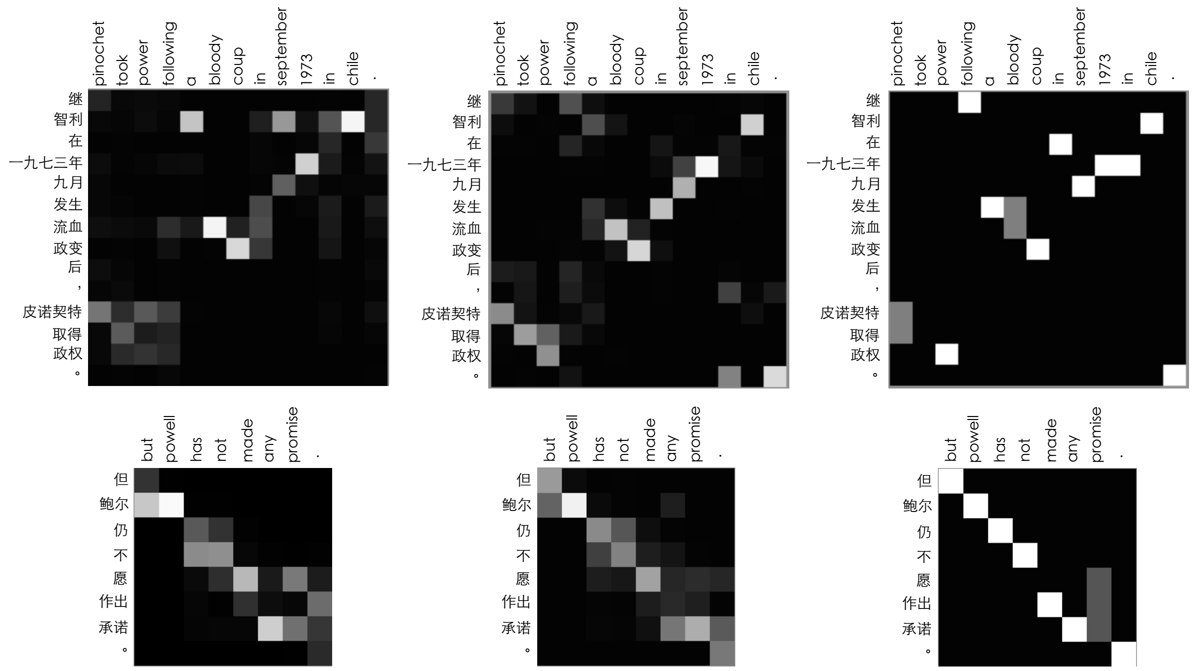
4.1.3 Results on Large Scale Translation Task

Figure 2 shows the learning curves of NMT2 and SA-NMT on the development set. We can see that NMT2 generally obtains higher BLEU as the increasing of updates before peaking at update of 150000, while it is unstable from then on. On the other hand, SA-NMT delivers much better BLEU for the beginning updates and performs more steadily along with the updates, although it takes more updates to reach the peaking point.

Table 3 reports the main end-to-end translation results for the large scale task. We find that both standard NMT generally outperforms Moses except NMT1 on nist05. The proposed SA-NMT achieves significant and consistent improvements over all three baseline systems, and it obtains the averaged gains of 2.2 BLEU points on test sets over its direct baseline NMT2. It is clear from these results that our supervised attention mechanism is highly effective in practice.

4.1.4 Results and Analysis on Alignment

As explained in §2, standard NMT can not use the target word information to predict its aligned source words, and thus might fail to predict the correct source words for some target words. For example, for the sentence in the training set in Figure 3 (a), NMT2 aligned 'following' to '皮诺契特 (gloss: pinochet)' rather than '继 (gloss: follow)', and worse still it aligned the word '.' to '在 (gloss: in)' rather than '。' even though this word is relatively easy to align correctly. In contrast, with the help of information from the target word itself, GIZA++ successfully aligned both 'following' and '.' to the expected source words (see Figure3(c)). With the alignment results from GIZA++ as supervision, we can see that our SA-NMT can imitate GIZA++ and thus align both words correctly. More importantly, for sentences in the unseen test set, like GIZA++, SA-NMT confidently aligned 'but' and '.' to their correct source words respectively as in Figure3(b), where NMT2 failed. It seems that SA-NMT can learn its alignment behavior from GIZA++, and subsequently apply the alignment abilities it has learned to unseen test sentences.



(a) NMT2 (b) SA-NMT (c) GIZA++

Figure 3: Example (soft) alignments of (a) NMT2 (i.e., standard NMT with unsupervised attention), (b) SA-NMT (i.e. NMT with supervised attention), and (c) GIZA++ on two Chinese-English sentence pairs. The soft alignments of (c) is converted from hard alignment as in §3.1. The first row shows the alignments of the sentence pair from the training set while the second row shows the alignments from test sets.

Methods	AER
GIZA++	30.6*
NMT2	50.6
SA-NMT	43.3*

Table 4: Results on word alignment task for the large scale data. The evaluation metric is Alignment Error Rate (AER). '*' denotes that the corresponding result is significanly better than NMT2 with $p < 0.01$.

Table 4 shows the overall alignment results on word alignment task in terms of the metric, alignment error rate. We used the manually-aligned dataset as in (Liu and Sun, 2015) as the test set. Following (Luong and Manning, 2015), we force-decode both the bilingual sentences including source and reference sentences to obtain the alignment matrices, and then for each target word we extract one-to-one alignments by picking up the source word with the highest alignment confidence as the hard alignment. From Table 4, we can see clearly that standard NMT (NMT2) is far behind GIZA++ in alignment quality. This shows that it is possible and promising to supervise the attention with GIZA++. With the help from GIZA++, our supervised attention based NMT (SA-NMT) significantly reduces the AER, compared with the unsupervised counterpart (NMT2). This shows that the proposed approach is able to realize our intuition: the alignment is improved, leading to better translation performance.

Note that there is still a gap between SA-NMT and GIZA++ as indicated in Table 4. Since SA-NMT was trained for machine translation instead of word alignment, it is possible to reduce its AER if we aim to the word alignment task only. For example, we can enlarge λ in Eq.(4) to bias the training objective towards word alignment task, or we can change the architecture slightly to add the target information crucial for alignment as in (Yang et al., 2013; Tamura et al., 2014).

Systems	CSTAR03	IWSLT04
Moses	44.1	45.1
NMT1	33.4	33.0
NMT2	36.5	35.9
SA-NMT	39.8*	40.7*

Table 5: BLEU comparison for low-resource translation task. CSTAR03 is the development set while IWSLT04 is the test set. '*' denotes that SA-NMT is significantly better than both NMT1 and NMT2 with $p < 0.01$.

4.2 Results on the Low Resource Translation Task

For the low resource translation task, we used the BTEC corpus as the training data, which consists of 30k sentence pairs with 0.27M Chinese words and 0.33M English words. As development and test sets, we used the CSTAR03 and IWSLT04 held out sets, respectively. We trained a 4-gram language model on the target side of training corpus for running Moses. For training all NMT systems, we employed the same settings as those in the large scale task, except that vocabulary size is 6000, batch size is 16, and the hyper-parameter $\lambda = 1$ for SA-NMT.

Table 5 reports the final results. Firstly, we can see that both standard neural machine translation systems NMT1 and NMT2 are much worse than Moses with a substantial gap. This result is not difficult to understand: neural network systems typically require sufficient data to boost their performance, and thus low resource translation tasks are very challenging for them. Secondly, the proposed SA-NMT gains much over NMT2 similar to the case in the large scale task, and the gap towards Moses is narrowed substantially.

While our SA-NMT does not advance the state-of-the-art Moses as in large scale translation, this is a strong result if we consider that previous works on low resource translation tasks: Arthur et al. (2016) gained over Moses on the Japanese-to-English BTEC corpus, but they resorted to a corpus consisting of 464k sentence pairs; Luong and Manning (2015) revealed the comparable performance to Moses on English-to-Vietnamese with 133k sentences pairs, which is more than 4 times of our corprus size. Our method is possible to advance Moses by using reranking as in (Neubig et al., 2015; Cohn et al., 2016), but it is beyond the scope of this paper and instead we remain it as future work.

5 Related Work

Many recent works have led to notable improvements in the attention mechanism for neural machine translation. Tu et al. (2016) introduced an explicit coverage vector into the attention mechanism to address the over-translation and under-translation inherent in NMT. Feng et al. (2016) proposed an additional recurrent structure for attention to capture long-term dependencies. Cheng et al. (2016) proposed an agreement-based bidirectional NMT model for symmetrizing alignment. Cohn et al. (2016) incorporated multiple structural alignment biases into attention learning for better alignment. All of them improved the attention models that were learned in an unsupervised manner. While we do not modify the attention model itself, we learn it in a supervised manner, therefore our approach is orthogonal to theirs.

It has always been standard practice to learn reordering models from alignments for conventional SMT either at the phrase level or word level. At the phrase level, Koehn et al. (2007) proposed a lexicalized MSD model for phrasal reordering; Xiong et al. (2006) proposed a feature-rich model to learn phrase reordering for BTG; and Li et al. (2014) proposed a neural network method to learn a BTG reordering model. At the word level, Bisazza and Federico (2016) surveyed many word reordering models learned from alignment models for SMT, and there are some neural network based reordering models, such as (Zhang et al., 2016). Our work is inspired by these works in spirit, and it can be considered to be a recurrent neural network based word-level reordering model. The main difference is that in our approach the reordering model and translation model are trained jointly rather than separately as theirs.

Supervising the attention variables for attention-based neural networks is pioneered by Liu et al.

(2016). On image caption task, Liu et al. (2016) supervise the attention with external guidances in either a strong or a weak supervision manner. Their method requires the training data to be associated with direct annotation or indirect annotation. In parallel to our work, particularly on machine translation, Mi et al. (2016) and Chen et al. (2016) guide the attention for NMT from conventional word alignment models as teachers without any annotation on machine translation task. The differences of our work lie in that: we consider the attention as a form of a reordering model, which is thereby straightforward to be learned from conventional word alignment models; and we also provide a theoretical explanation why the attention leads to the worse alignment accuracy than the conventional word alignment models, standing upon the point view of reordering.

6 Conclusion

It has been shown that attention mechanism in NMT is worse than conventional word alignment models in its alignment accuracy. This paper firstly provides an explanation for this by viewing the attention mechanism from the point view of reordering. Then it proposes a supervised attention for NMT with guidance from external conventional alignment models, inspired by the supervised reordering models in conventional SMT. Experiments on two Chinese-to-English translation tasks show that the proposed approach achieves better alignment results leading to significant gains relative to standard attention based NMT.

Acknowledgements

We would like to thank Xugang Lu for invaluable discussions and three anonymous reviewers for many valuable comments and helpful suggestions on this work.

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. *CoRR*, abs/1606.02006.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Arianna Bisazza and Marcello Federico. 2016. A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Computational Linguistics*, 42.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. In *Proceedings of AMTA*.

Yong Cheng, Shiqi Shen, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Agreement-based joint training for bidirectional attention-based neural machine translation. In *Proceedings of IJCAI*.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

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 NAACL-HLT*.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of ACL*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *In Proc. NAACL*.

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04.

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.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL: Demonstrations*.

E.L. Lehmann and G. Casella. 1998. *Theory of Point Estimation*. Springer Verlag.

Peng Li, Yang Liu, Maosong Sun, Tatsuya Izuha, and Dakun Zhang. 2014. A neural reordering model for phrase-based translation. In *Proceedings of COLING*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan L. Yuille. 2016. Attention correctness in neural image captioning. *CoRR*, abs/1605.09553.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of IWSLT*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. In *Proceedings of EMNLP*.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural reranking improves subjective quality of machine translation: NAIST at WAT2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL*, pages 440–447.

Reuven Y. Rubinstein and Dirk P. Kroese. 2004. *The Cross Entropy Method: A Unified Approach To Combinatorial Optimization, Monte-carlo Simulation (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2015. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2014. Recurrent neural networks for word alignment model. In *Proceedings of ACL*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of ACL*.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of ACL*.

Nan Yang, Shujie Liu, Mu Li, Ming Zhou, and Nenghai Yu. 2013. Word alignment modeling with context dependent deep neural network. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, August.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*.

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Hai Zhao, Graham Neubig, and Satoshi Nakamura. 2016. Learning local word reorderings for hierarchical phrase-based statistical machine translation. *Machine Translation*.