Target Foresight based Attention for Neural Machine Translation*

Xintong Li†, Lemao Liu‡, Zhaopeng Tu‡, Shuming Shi‡, Max Meng†

†The Chinese University of Hong Kong
{xtli, qhmeng}@ee.cuhk.edu.hk
‡Tencent AI Lab
{redmondliu, zptu, shumingshi}@tencent.com

Abstract

in neural machine translation, an attention model is used to identify the aligned source words for a target word (target foresight word) in order to select translation context, but it does not make use of any information of this target foresight word at all. previous work proposed an approach to improve the attention model by explicitly accessing this target foresight word and demonstrated the substantial gains in alignment task. however, this approach is useless in machine translation task on which the target foresight word is unavailable. in this paper, we propose a new attention model enhanced by the implicit information of target foresight word oriented to both alignment and translation tasks. empirical experiments on chinese-to-english and japanese-to-english datasets show that the proposed attention model delivers significant improvements in terms of both alignment error rate and bleu.

1 Introduction

Since neural machine translation (NMT) was proposed (Bahdanau et al., 2014), it has been attracted increasing interests in machine translation community (Luong et al., 2015b; Tu et al., 2016; Feng et al., 2016; Cohn et al., 2016). NMT not only yields impressive translation performance in practice, but also has appealing model architecture in essence. Compared with traditional statistical machine translation (Koehn et al., 2003; Chiang, 2005), one of advantages in NMT is that its architecture combines language model, translation model and alignment between source and target words in a unified manner rather than a pipeline manner, and it thereby has the potential to alleviate the issue of error propagation.

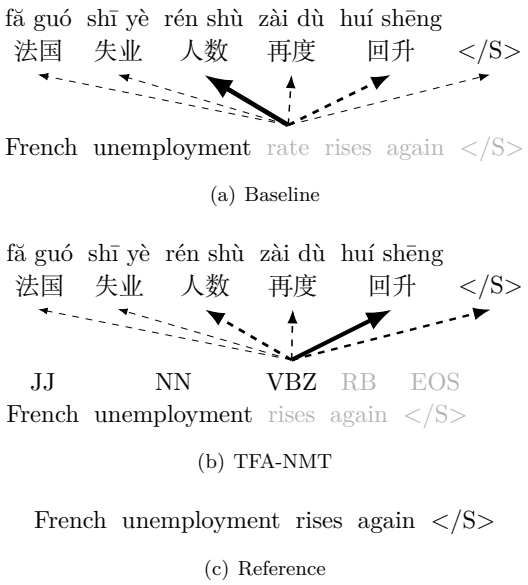


Figure 1: A running example to motivate the proposed model. (a) The baseline obtains a translation error due to the incorrect attention. (b) With the help of the target foresight information "VBZ", TFA-NMT is likely to figure out the exact translation as the reference in (c). The light font denotes the target words to be translated in future. Both dashed or solid arrowed lines denote the alignments and solid one denotes the 1-best alignment.

In NMT, the attention mechanism plays an important role. It calculates the alignments of a target word with respect to the source words for translation context selection. Although the source words are always available in inference, the target word, called target foresight word, [1]

*Work done when X. Li interning at Tencent AI Lab. L. Liu is the corresponding author.

[1]Note that the concept of foresight word in our translation task is not exactly the same as the original concept in alignment task (Peter et al., 2017). However, both of them share a common idea that foresight word should be at a later time step, and thus we respect the work in Peter et al. (2017) and maintain the same concept for easier understanding.

Proceedings of NAACL-HLT 2018, pages 1380–1390
New Orleans, Louisiana, June 1 - 6, 2018.

i.e. the first light color word in Figure 1(a), is not known but to be translated at the next time step. Therefore, this may lead to inadequate modeling for attention mechanism (Liu et al., 2016a; Peter et al., 2017). Regarding to this, Peter et al. (2017) explicitly feed this target word into the attention model, and demonstrate the significant improvements in alignment accuracy. Unfortunately, this approach relies on the premise that the target foresight word is available in advance in its alignment scenario, and thus it can not be used in the translation scenario.

To address this issue, in this paper, we propose a target foresight based attention (TFA) model oriented to both alignment and translation tasks. Its basic idea includes two steps: it firstly designs an auxiliary mechanism to predict some information for the target foresight word which is helpful for alignment; and then it feeds the predicted result into the attention model for translation. For the sake of efficiency, instead of predicting the target foresight word with large vocabulary size, we only predict its partial information, i.e. part-of-speech tag, which is proved to be helpful for word alignment (Liu et al., 2005). Figure 1(b) shows the main idea of TFA based on NMT. In order to remit the negative effects due to the prediction errors, we feed the distribution of the prediction result instead of the maximum a posteriori result into the attention model. In addition, since the target foresight words are available during the training, we jointly learn the prediction model for the target foresight words and the translation model in a supervised manner.

This paper makes the following contributions:

- It proposes a novel TFA-NMT for neural machine translation by using an auxiliary mechanism to predict the target foresight word which is subsequently used to enhance the attention model.
- It empirically shows that the proposed TFA-NMT can lead to better alignment accuracy, and achieves significant improvements on both Chinese-to-English and Japanese-to-English translation tasks.

2 Background

Given a source sentence $\mathbf{x} = \{x_1, \ldots, x_m\}$ with length m and a target sentence $\mathbf{y} = \{y_1, \ldots, y_n\}$ with length n, neural machine translation aims to model the conditional probability $P(\mathbf{y} \mid \mathbf{x})$:

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{n} P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}), \quad (1)$$

where $\mathbf{y}_{<i} = \{y_1, \ldots, y_{i-1}\}$ denotes a prefix of $\mathbf{y}$ with length $i-1$.

To achieve this, neural machine translation adopts recurrent neural network (RNN) under the encoder-decoder framework (Bahdanau et al., 2014). In encoding, an encoder reads the source sentence $\mathbf{x}$ into a sequence of representation vectors by a bidirectional recurrent neural network. Suppose h_i denotes the representation vector for x_i, and let $\mathbf{h} = \{h_1, \ldots, h_m\}$. In decoding, a decoder sequentially generates a target word according to $P(y_i \mid \mathbf{y}_{<i}, \mathbf{x})$ by using another RNN.

In Eq.(1), the distribution $P(y_i \mid \mathbf{y}_{<i}, \mathbf{x})$ is used to generate y_i as follows:

$$P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}) = \mathrm{softmax}(\phi(y_{i-1}, s_i, c_i)), \quad (2)$$

where ϕ represents a feedforward neural network, c_i is the context vector from $\mathbf{h}$ to infer y_i, and s_i denotes the hidden state at timestamp i via the decoding RNN represented by f:

$$s_i = f(s_{i-1}, y_{i-1}, c_i). \quad (3)$$

Bahdanau et al. (2014) propose an attention model to define the context c_i, inspired by the alignment model in statistical machine translation.

Given the last hidden state s_{i-1} and the encoding vectors $\mathbf{h}$, an attention model is based on a distribution consisting of α_{ij} as follows:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{ik})},$$

where e_{ij} is computed by a feedforward neural network represented by a:

$$e_{ij} = a(s_{i-1}, h_j). \quad (4)$$

The quantity α_{ij} denotes the possibility of target word y_i aligns to the source word x_j encoded by h_j. According to α_{ij}, the context

vector c_i is defined as the weighted sum of $\mathbf{h}$:

$$c_i = \sum_{j=1}^{m} \alpha_{ij} h_j. \quad (5)$$

In this way, when translating the target word y_i, the decoder will pay more attention to its aligned source words with respect to the distribution $\alpha_i = \{\alpha_{i1}, \cdots, \alpha_{im}\}$. Figure 2 shows a slice of the entire architecture for NMT at timestamp i.

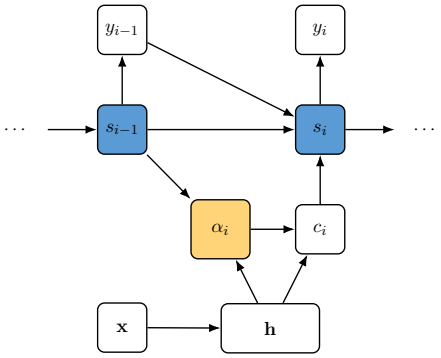


Figure 2: One slice of the architecture of Neural Machine Translation based on a generic attention.

Unfortunately, even though the entire translation $\mathbf{y}$ is available in training, during the inference it is unknown in advance but to be generated sequentially. Specifically, when calculating α_i, one can make use of the information only from $\mathbf{x}$ and $\mathbf{y}_{<i}$ but nothing from y_i. Therefore, it is difficult to certainly specify which source words should be aligned to an unknown target word y_i. This might lead to the inadequacy of the attention model (Liu et al., 2016a; Peter et al., 2017), as explained in Figure 1(a).

3 Target Foresight Attention

In order to alleviate the issue of inadequate modeling for attention in NMT, in this section, we propose the target foresight attention for NMT, which foresees some related information of the unknown target foresight word to improve its alignments regarding to source words. The basic idea of the proposed attention model includes two steps as following:

- It firstly introduce a model to predict some information of the target foresight word. (§3.1)
- It then feeds the predicted result about the foresight target word into the attention as an additional input. (§3.2)

Therefore, as shown in Figure 1(b), when translating the third word, if the prediction model shows it to be a "VBZ", the attention model is likely to align it to the verb words such as "huí shēng" rather than "rén shù" in the source side, and then the corrected word "rises" will be translated.

3.1 Target Foresight Prediction

Ideally, it is possible to build a model to directly predict the target foresight word itself. In practice, it will be inefficient due to its large vocabulary size. As a result, we instead build a model to predict the partial information of the target foresight word, such as part-of-speech (POS) tag or word cluster, which has limited vocabulary size. In this paper, we use the POS tag as the partial information of a target foresight word because POS tag is helpful to word alignment proved by Liu et al. (2005). Furthermore, predicting a POS tag is easier than a target foresight word, so the predicted result will be more reliable for the downstream application on attention.

Suppose u_i denotes a variable indicating the POS tag of a target foresight word y_i. Our aim is to define a prediction model of u_i prior to calculate the attention probability. For simplicity, this prediction model is generally represented as $\beta_i = P(u_i \mid \mathbf{y}_{<i}, \mathbf{x})$. We consider three variant prediction models in a coarse-to-fine manner as follows.

3.1.1 Model 1

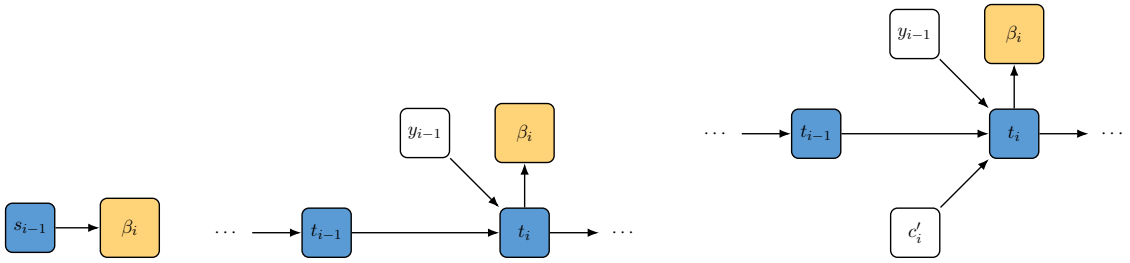
It is straightforward to define this prediction model directly based on the hidden states of the RNN in decoder by using a neural network. Formally, one can use the following equation:

$$\beta_i = P\left(u_i \mid \mathbf{y}_{<i}, \mathbf{x}\right) = \text{softmax}\left(\psi(y_{i-1}, s_{i-1})\right), \quad (6)$$

where ψ is implemented by a feedforward neural network. Note that Eq.(6) only depends on the decoding RNN hidden state s_{i-1} and it is very simple to implementation. Figure 3(a) shows its architecture.

3.1.2 Model 2

Unlike Eq.(6) relying on the same hidden s_{i-1} as the decoder, we design a specialized RNN



(a) model 1 (b) model 2 (c) model 3

Figure 3: The prediction coarse-to-fine models for target foresight information: (a) Model 1 using only the decoding hidden state s_{i-1}. (b) Model 2 using a hidden state t_i from a specialized RNN. (c) Models using a hidden state from a specialized RNN enhanced by the representation vector c'_i of $\mathbf{x}$ similar to Eq.(5).

to provide a particular hidden state for prediction of u_i. This improved prediction model is defined as follows:

$$\beta_i = P\left(u_i \mid \mathbf{y}_{<i}, \mathbf{x}\right) = \text{softmax}\left(\psi(y_{i-1}, t_i)\right), \tag{7}$$

where t_i is the hidden state of the specialized RNN defined by a GRU unit, i.e. $t_i = g(t_{i-1}, y_{i-1})$. This prediction model architecture is shown in Figure 3(b).

3.1.3 Model 3

In model 2, the specialized RNN for u_i only cares about the target sentence $\mathbf{y}$ and ignores the information from the source sentence $\mathbf{x}$. We define a fine-grained model by taking a context vector c'_i from $\mathbf{x}$ as an additional input:

$$\beta_i = P\left(u_i \mid \mathbf{y}_{<i}, \mathbf{x}\right) = \text{softmax}\left(\psi(y_{i-1}, t_i, c'_i)\right), \tag{8}$$

where c'_i is a context vector extracted from $\mathbf{x}$ in a way similar to c_i in Eq.(5),[2] and $t_i = g(t_{i-1}, y_{i-1}, c'_i)$ is the hidden state of the specialized RNN. The architecture of this model is shown in Figure 3(c).

3.2 Feeding the Prediction Model

Suppose we have the prediction result $P(u_i \mid \mathbf{y}_{<i}, \mathbf{x})$, then we consider to feed it into the attention model. Firstly, it is natural to feed the prediction into attention by using maximum a posteriori (MAP) strategy:

$$e_{ij} = a(s_{i-1}, h_j, z_i), \tag{9}$$

where a is the function for attention similar to Eq.(4) but includes an additional input z_i, which is the MAP result of $P(u_i \mid \mathbf{y}_{<i}, \mathbf{x})$:

$$z_i = \mathbf{z}\big(\operatorname*{argmax}_{u_i} P\left(u_i \mid \mathbf{y}_{<i}, \mathbf{x}\right)\big), \tag{10}$$

where $\mathbf{z}$ denotes the embeddings of the POS tags of target foresight words, and $\mathbf{z}(u_i)$ returns the embedding of a particular POS tag u_i.

Note that in Eq.(10) the accuracy of $P(u_i \mid \mathbf{y}_{<i}, \mathbf{x})$ is important to the attention model. For example, suppose at timestamp i, the ground-truth POS tag is "NN", but one has $P(u_i = \text{NN} \mid \mathbf{y}_{<i}, \mathbf{x}) = 0.4$ and $P(u_i = \text{VV} \mid \mathbf{y}_{<i}, \mathbf{x}) = 0.41$. In this case, the prediction model selects "VV" as the POS tag of the target foresight word and ignores the ground-truth tag "NN". Then the attention model takes this error signal and may align the target foresight word to a verb word. Subsequently, this might lead to a translation error.

Therefore, we propose another method to integrate the expected embedding of u_i according to $P(u_i \mid \mathbf{y}_{<i}, \mathbf{x})$ into attention as follows:

$$z_i = \sum_{u_i} \mathbf{z}(u_i) P\left(u_i \mid \mathbf{y}_{<i}, \mathbf{x}\right). \tag{11}$$

In this way, z_i can take into account all possible POS tags u_i including the ground-truth result.

Until now, we can obtain the entire architecture of the proposed target foresight attention based NMT (TFA-NMT), as shown in Figure 4. Comparing Figure 4 with Figure 2, the only difference is the variable z_i, which is

[2] In our preliminary experiments, we tried c_i, but we found c'_i performs better.

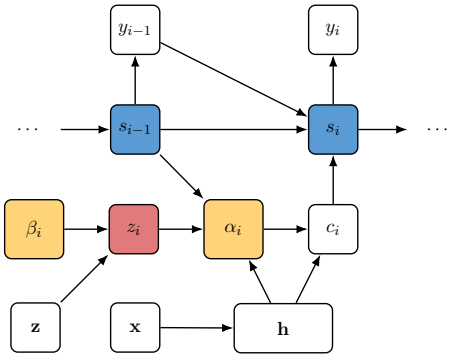


Figure 4: Neural machine translation with target Foresight attention. β_i is derived from Figure 3, z_i is from Eq.(10-11), and other nodes are similar to ones in Figure 2.

obtained from Eq.(10-11) and the prediction model as shown in Figure 3.

Note that the proposed TFA-NMT models the target foresight word, which is a future word regarding to the current time step, to conduct attention calculation. In this sense, it employs the idea of modeling future and thus resembles to the work in (Zheng et al., 2017). The main difference is that TFA-NMT models the future from the target side whereas Zheng et al. (2017) models the future from the source side. In addition, Weng et al. (2017) imposes a regularization term by using future words during training. Unlike our approach, their approach does not use future words during the inference because these words are unavailable. Anyway, it is possible to put both their approach and our approach together for further improvements.

3.3 Learning and Inference

Suppose a set of training data is denoted by $\{\langle \mathbf{x}^k, \mathbf{y}^k, \mathbf{u}^k \rangle \mid k = 1, \cdots, K\}$. Here $\mathbf{x}^k$, $\mathbf{y}^k$ and $\mathbf{u}^k$ denotes a source sentence, a target sentence and a POS tag sequence of $\mathbf{y}^k$, respectively. Then one can jointly train both the translation model for $\mathbf{y}^k$ and the prediction model for $\mathbf{u}^k$ by minimizing the loss function:

$$\ell = -\sum_k \sum_i \big(\log P(y_i^k \mid \mathbf{y}_{<i}^k, \mathbf{x}^k) + \lambda \log P(u_i^k \mid \mathbf{y}_{<i}^k, \mathbf{x}^k) \big), \quad (12)$$

where $P(y_i^k \mid \mathbf{y}_{<i}^k, \mathbf{x}^k)$ is the translation model similar to Eq.(2) with target foresight attention, and $P(u_i^k \mid \mathbf{y}_{<i}^k, \mathbf{x}^k)$ is the target foresight prediction model as defined in Eq.(6-8), respectively. $\lambda \geq 0$ is a hyper-parameter that balances the preference between the translation model and target foresight prediction model.

According to the training objective, the proposed TFA-NMT resembles to the multi-task learning, since it jointly learns two tasks similar to (Evgeniou and Pontil, 2004; Luong et al., 2015a). The difference of our approach is obviously: in this work the prediction result of one model is integrated into the other model, while in their works, two models only share some common hidden states.

In inference, we implement two different decoding methods according two different ways to integrate the foresight prediction model into attention as described in §3.2. For the MAP feeding style, we optimize u_i according to the loss function in Eq.(12) by beam search besides optimizing y_i. However, for the expectation feeding style, we maintain the standard beam search algorithm only regarding to the translation model, i.e. by setting $\lambda = 0$.

4 Experiments

We conduct experiments on Chinese-to-English and Japanese-to-English translation tasks. The specific analyses are based on Chinese-to-English task, and the generalization ability is shown by Japanese-to-English task. Case-insensitive 4-gram BLEU is used to evaluate translation quality, and the multi-bleu.perl is adopted as its implementation.

4.1 Setup

Data The training data for Chinese-to-English task consists of 1.8M sentence pairs from NIST2008 Open Machine Campaign, with 40.1M Chinese words and 48.3M English words respectively. The development set is chosen as NIST2002 (878 sentences) and the test sets are NIST2005 (1082 sentences), NIST2006 (1664 sentences), and NIST2008 (1357 sentences).

For Japanese-to-English translation, we adopt the data sets from NTCIR-9 patent translation task (Goto et al., 2013). The training data consists of 2.0M sentence pairs with 53.4M Japanese words and 49.3M English words, the development and test sets respectively contain 2000 sentences with a single ref-

Model	# Para.	Speed		Performance	
		Train	*Decode*	*BLEU*	*FPA*
NEMATUS	105M	2858.8	86.6	38.65	–
+2-LAYER	+6M	2522.5	84.1	38.57	–
+MODEL1	+2M	1844.9	72.0	38.83	69.03
+MODEL2	+12M	1666.1	70.1	39.26	69.95
+MODEL3	+27M	1485.2	59.1	**40.63**	**71.91**

Table 1: Speeds and performances of the proposed models. "Speed" is measured in words/second for both training and decoding, and performances are measured in terms of BLEU scores ("BLEU") and foresight prediction accuracy ("FPA") on the development set. Higher BLEU and FPA scores denote better performance.

erence, following (Goto et al., 2013; Liu et al., 2016b) for further comparison.

Implementation We compare the proposed models with two strong baselines from SMT and NMT:

- MOSES (Koehn et al., 2007): an open source phrased based translation system with default configuration.
- NEMATUS (Sennrich et al., 2017): an generic attention based NMT.

We implement the proposed models on top of NEMATUS. We use Stanford Log-linear Part-Of-Speech Tagger (Toutanova et al., 2003) to produce POS tags for the English side. For both Chinese-to-English and Japanese-to-English tasks, we limit the vocabularies to the most frequent 30K words for both sides. All the out-of-vocabulary words are mapped to a spacial token "UNK". Only the sentences of length up to 50 words are used in training, with 80 sentences in a batch. The dimension of word embedding is 620. The dimensions of both feed forward NN and RNN hidden layer are 1000. The beam size for decoding is 12, and the cost function is optimized by Adadelta with hyper-parameters suggested by Zeiler (2012). Particularly for TFA-NMT, the foresight embedding is also 620, and the hyper-parameter λ is 1.

4.2 Impact of Components

We conduct analyses on Chinese-to-English translation task, to investigate the impact of the added components and to figure out their best configuration for further testing in the next subsection.

4.2.1 Model Architectures

Table 1 lists the speeds and performances of the proposed models. Clearly the proposed approach improves the translation quality in all cases, although there are still considerable differences among the proposed variants.

Model Complexity The proposed models introduce a few parameters to the NMT baseline system NEMATUS, which has 105M parameters. The most complex model (i.e., MODEL3) introduces 27M new parameters, which are small compared with the baseline model. As seen, the proposed models significantly slows down the training speed, which we attribute to the new softmax operation over the foresight tags and more gradient operations associated with the new training objective, i.e., Eq.(12). For decoding, the most complex model reduces speed by around 30%, which is the cost of the proposed approach for improving translation quality.

Performance We measure the performance with BLEU and the result is shown in Table 1. MODEL1 marginally improves performance by guiding the decoder states to embed information for predicting foresight tags. MODEL2 achieves further improvement by introducing a new specific hidden layer to explicitly separate the predict function from the decoder states. MODEL3 achieves the best performance by adopting an independent attention model to attend corresponding source parts for foresight prediction, which may not be the same as the attended source parts for translation. We conduct the significant test using Kevin Gimpel's toolkit (Clark et al., 2011). We found that MODEL1 is not signif-

Type	Perc.	FPA	AER	
		OURS	BASE	OURS
Noun	30.13%	77.49	28.97	26.50
Verb	12.39%	71.94	37.06	33.93
Adj.	9.43%	55.99	34.67	31.86
Prep.	14.66%	79.40	**84.04**	**76.95**
Dete.	10.08%	72.06	80.15	76.51
Punc.	8.01%	74.89	**91.74**	**66.51**
Others	15.30%	81.22	53.64	39.11
All	100%	74.87	49.67	42.56

Table 2: Performances on syntactic categories. "BASE" denotes "NEMATUS", and OURS denotes the proposed model.

icantly better than baseline, but MODEL2 is significantly better with p<0.05 and MODEL3 is significantly better with p<0.01. Given that simply introducing an additional layer ("+2-LAYER") does not produce any improvement on this data, we believe the gain of our model is not only from the more introduced parameters. Besides, we augment the word embedding by concatenating the POS tag embedding, proposed by (Sennrich and Haddow, 2016), the BLEU is 38.96, which indicating the improvement of our model is not only from the POS tagging. In order to further validate the improvements of variant proposed models, we evaluate the foresight prediction accuracy (FPA) for three proposed prediction models. We found that the fine-grained MODEL3 achieves the best FPA, indicating a good estimated foresight is very important to obtain the gains in terms of BLEU.

4.2.2 Analysis on Syntactic Categories

In this experiment, we investigate which category of generated words benefit most from the proposed approach in terms of alignments measured by alignment error rate (AER) (Och, 2003). We carry out experiments on the evaluation dataset from (Liu and Sun, 2015), which contains 900 manually aligned Chinese-English sentence pairs. Following (Luong et al., 2015b), we force-decode both the bilingual sentences including source and reference sentences to obtain the attention matrices, and then we extract one-to-one alignments by picking up the source word with the highest alignment confidence as the hard alignment. As shown in Table 2, the AER improvements are modest for content words such as Noun, Verb, and adjective ("Adj.") words; but there are substantial improvements for function words such as preposition words ("Prep.") and punctuations ("Punc.").

Train (λ)	Decode	BLEU	$\bigtriangledown$
1	EXP	40.63	–
0	EXP	39.36	-1.27
1	MAP	40.34	-0.29

Table 3: Effect of foresight supervision signal in training (i.e., λ) and foresight representations in decoding: EXP for expectation and MAP for maximum a posteriori.

The reason can be explained as follows. The content words are easy to align with AER under 38 as shown in Table 2, and thus it is more difficult to gain over the BASE. On the other hand, as depicted in Table 2, function words are inherently more difficult than content words. These findings satisfy the linguistic intuition: content words tend to be less involved in multiple potential correspondences than function words, and function words tend to be attached to content words, as pointed out by Pianta and Bentivogli (2004). Fortunately, TFA-NMT can predict the POS tag for target foresight word with high confidence and thus it can improve the alignment quality by using of POS tags, which is useful for alignment task (Liu et al., 2005).

It is surprising that the AER for "Prep.", "Det." and "Punc." is relatively low especially for BASE. The main reason can be explained from the quantities y_{i-1}, s_i, and c_i in Eq.(2) as follows. These highly frequent function words are usually easy to be translated by using the history information from y_{i-1} and s_i even if c_i is not confident enough. For example, it is relatively easy to guess the "comma" by using the history words in language model task, where there are no bilingual information at all. Therefore, during the training, the model tries to adjust the parameters for highly frequent words from y_{i-1} and s_i while neglecting the attention model.

4.2.3 Foresight Strategies

Table 3 shows the performances of different foresight strategies in both training and de-

coding. Without an explicit objective to guide the training of foresight prediction model (i.e., $\lambda = 0$), the performance decreases by 1.27 BLEU points. When feeding the best foresight predicted result to the attention model (i.e., MAP), the performance decreases by 0.29 BLEU points. We attribute this to the propagation of prediction errors, which can be alleviated by using a weighted representation of all predicted results (i.e., EXP).

In the following experiments, we use "$\lambda = 1$ and EXP" as the default setting for the final system TFA-NMT.

4.3 Main Results

Chinese-to-English Task Table 4 shows the translation performances for the Chinese-to-English translation task. As seen, the proposed approach significantly outperforms the baseline system (i.e., NEMATUS) in all cases, demonstrating the effectiveness and university of our model.

Japanese-to-English Task Table 5 shows the translation quality of the NMT baseline and our TFA-NMT on Japanese-to-English task. From the table, we can see that our model still achieves a significant improvement of 1.22 and 1.31 BLEU points on the development and test set, respectively. This shows that the proposed approach works well across different language pairs.

5 Related Work

Attention model becomes a standard component for many applications due to its ability of dynamically selecting the informative context from sequential representations. For example, Xu et al. (2015) propose an attention based neural network for image caption task and advance the state-of-the-art results; Yin et al. (2015) put the attention structure between a pair of convolution networks for answer selection, paraphrase identification and textual entailment tasks. In the context of machine translation, the idea of attention based neural networks has been pioneered by Bahdanau et al. (2014); Luong et al. (2015b) and achieved impressive results over the traditional statistical machine translation. Since then many research works have been devoted to improve the neural machine translation by enhancing attention models.

Tu et al. (2016) design a coverage vector for the translation history and then integrates it into the attention model. Similarly, Meng et al. (2016) maintain a tag vector to keep track of the attention history and Sankaran et al. (2016) memorize historical alignments and accumulate them as temporal memory to improve the attention model. In addition, Zhang et al. (2017) improve the attention with a gated operator for encoding states and a decoding state, and and Dutil et al. (2017) enhance attention through a planning mechanism. Furthermore, Feng et al. (2016) adopt a recurrent structure for attention to take long-term dependencies into account, Zhou et al. (2017) propose a look-ahead attention by additionally modeling the translation history, and Cohn et al. (2016) incorporate structural biases into attention models. Recently Chen et al. (2017) introduce the syntactic knowledge into attention models. These works are essentially similar to the propose approach, since we introduce auxiliary information from a target foresight word into the attention model. However, there is a significant difference between our approach and their approaches. Our auxiliary information biases to the word to be translated at next timestep while theirs biases to the information available so far at the current timestep, and thereby our approach is orthogonal to theirs.

The works mentioned above improve the attention models by access auxiliary information, and thus they modify the structure of attention models in both inference and learning. In contrast, Mi et al. (2016); Liu et al. (2016b); Chen et al. (2016) maintain the structure of the attention models in inference but utilize some external signals to supervise the outputs of attention models during the learning. They improve the generalization abilities of attention models by use of the external aligners as the signals, which typically yield alignment results accurate enough to guide the learning of attention.

6 Conclusion

It has been argued that the traditional attention model in neural machine translation suf-

System	Model	Dev	MT05	MT06	MT08	Ave.
(Liu et al., 2016b)	MOSES	–	35.4	33.7	25.0	31.37
	NMT-J	–	36.8	36.9	28.5	34.07
(Liu et al., 2016a)	SA-NMT	40.0	**37.8**	37.6	29.9	35.10
This work	NEMATUS	38.65	36.32	36.10	28.24	33.55
	TFA-NMT	**40.63**	37.70	**38.01**	**30.12**	**35.28**

Table 4: Evaluation of translation performance on Chinese-to-English task.

System	Model	Dev	Test
(Liu et al., 2016b)	MOSES	28.6	30.2
	NMT-J	33.0	34.1
This work	NEMATUS	33.92	35.01
	TFA-NMT	35.14	36.32

Table 5: Evaluation of translation performance on Japanese-to-English task.

fers from model inadequacy due to the lack of information from the target foresight word (Peter et al., 2017; Liu et al., 2016a). To address this issue, this paper proposes a new attention model, which can serve for both alignment and translation tasks, by implicitly making use of the target foresight word. Empirical experiments on Chinese-to-English and Japanese-to-English tasks demonstrate that the proposed attention based NMT delivers substantial gains in terms of both BLEU and AER scores.

In future work, it is promising to exploit other target foresight information such as word cluster besides the POS tags in this paper, and it is also interesting to apply this idea on top of other attention models such as the local attention in Luong et al. (2015b).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2017. Syntax-directed attention for neural machine translation. *arXiv preprint arXiv:1711.04231* .

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. *arXiv preprint arXiv:1607.01628* .

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 263–270.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 176–181.

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. *arXiv preprint arXiv:1601.01085* .

Francis Dutil, Caglar Gulcehre, Adam Trischler, and Yoshua Bengio. 2017. Plan, attend, generate: Planning for sequence-to-sequence models. *arXiv preprint arXiv:1711.10462* .

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*. ACM, pages 109–117.

Shi Feng, Shujie Liu, Nan Yang, Mu Li, Ming Zhou, and Kenny Q Zhu. 2016. Improving attention modeling with implicit distortion and fertility for machine translation. In *COLING*. pages 3082–3092.

Isao Goto, Ka-Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K Tsou. 2013. Overview of the patent machine translation task at the ntcir-10 workshop. In *NTCIR*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting*

of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics, pages 177–180.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 48–54.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016a. Neural machine translation with supervised attention. *arXiv preprint arXiv:1609.04186* .

Lemao Liu, Masao Utiyama, Andrew M Finch, and Eiichiro Sumita. 2016b. Agreement on target-bidirectional neural machine translation. In *HLT-NAACL*. pages 411–416.

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 459–466.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *AAAI*. pages 2295–2301.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015a. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114* .

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015b. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

Fandong Meng, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Interactive attention for neural machine translation. *arXiv preprint arXiv:1610.05011* .

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. *arXiv preprint arXiv:1608.00112* .

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 160–167.

Jan-Thorsten Peter, Arne Nix, and Hermann Ney. 2017. Generating alignments using target foresight in attention-based neural machine translation. *The Prague Bulletin of Mathematical Linguistics* 108(1):27–36.

Emanuele Pianta and Luisa Bentivogli. 2004. Knowledge intensive word alignment with knowa. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 1086.

Baskaran Sankaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. 2016. Temporal attention model for neural machine translation. *arXiv preprint arXiv:1608.02927* .

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357* .

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892* .

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 173–180.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811* .

Rongxiang Weng, Shujian Huang, Zaixiang Zheng, XIN-YU DAI, and Jiajun CHEN. 2017. Neural machine translation with word predictions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 136–145.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. pages 2048–2057.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193* .

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

Biao Zhang, Deyi Xiong, and Jinsong Su. 2017. A gru-gated attention model for neural machine translation. *arXiv preprint arXiv:1704.08430* .

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lili Mou, Xinyu Dai, Jiajun Chen, and Zhaopeng Tu. 2017. Modeling past and future for neural machine translation. *arXiv preprint arXiv:1711.09502* .

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2017. Look-ahead attention for generation in neural machine translation. *arXiv preprint arXiv:1708.09217* .