Improved Neural Machine Translation with Source Syntax

Shuangzhi Wu[†*] **, Ming Zhou**[‡] **, Dongdong Zhang**[‡]
[†]Harbin Institute of Technology, Harbin, China
[‡]Microsoft Research
{v-shuawu, mingzhou, dozhang}@microsoft.com

Abstract

Neural Machine Translation (NMT) based on the encoder-decoder architecture has recently achieved the state-of-the-art performance. Researchers have proven that extending word level attention to phrase level attention by incorporating source-side phrase structure can enhance the attention model and achieve promising improvement. However, word dependencies that can be crucial to correctly understand a source sentence are not always in a consecutive fashion (i.e. phrase structure), sometimes they can be in long distance. Phrase structures are not the best way to explicitly model long distance dependencies. In this paper we propose a simple but effective method to incorporate source-side long distance dependencies into NMT. Our method based on dependency trees enriches each source state with global dependency structures, which can better capture the inherent syntactic structure of source sentences. Experiments on Chinese-English and English-Japanese translation tasks show that our proposed method outperforms state-of-the-art SMT and NMT baselines.

1 Introduction

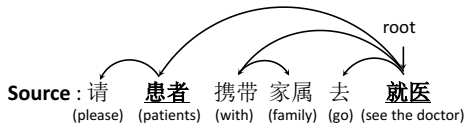
Recently, Neural Machine Translation (NMT) with the attention-based encoder-decoder framework [Bahdanau *et al.*, 2015] has achieved significant improvements in translation quality of many language pairs such as English-German, English-French and Chinese-English [Bahdanau *et al.*, 2015; Luong *et al.*, 2015a; Wu *et al.*, 2016]. In a conventional NMT model, an encoder maps the source sentence of various lengths into sequences of intermediate hidden vector representations. Then these hidden vectors are combined, weighted by attention mechanism, and used by the decoder to generate translations. In most cases, both encoder and decoder are implemented as recurrent neural networks (RNNs).

Currently, many methods have been proposed to improve the sequence-to-sequence NMT model since it was first proposed by [Bahdanau *et al.*, 2015; Sutskever *et al.*, 2014]. Previous work mostly focuses on addressing the problem of out-of-vocabulary words [Jean *et al.*, 2015], designing attention mechanism [Luong *et al.*, 2015a], and more efficient parameter learning [Shen *et al.*, 2016]. These methods regard sentences as sequences of words where the syntactic structures inherent in languages are neglected.

Recently, inspired by the successful application of source-side syntactic information in statistic machine translation (SMT) [Liu *et al.*, 2006], [Eriguchi *et al.*, 2016b] propose a new attentional NMT model which takes advantage of the source-side syntactic information based on the Head-driven Phrase Structure Grammar [Sag *et al.*, 1999]. They align each target word with both source words and source phrases. This kind of extension is effective to handle cases that one target word may correspond to a fragment of consecutive source words. However, the long distance syntactic dependencies of the source-side, which can be crucial to correctly understand a sentence, are not explicitly concerned in all previous work. Although, in theory, the encoder RNN is able to remember sufficiently long history, we can still observe substantial incorrect translations which are both fluent and grammatical but violate the meaning of source sentences. Figure 1 shows an incorrect translation example which relates to the source syntactic structure. Though the translation is well formed and grammatical, its meaning is inconsistent with the given source sentence. The NMT model can not well capture the dependency between word "患者(patients)" (subject) and "就医(see the doctor)" (predicate). Even for the phrase attention based model, this kind of relations still can not be explicitly modeled as the words are inconsecutive and in long distance. This demonstrates that it still remains a challenge for NMT encoder to capture such subtle long-range word dependencies for correctly understanding source sentences. Actually, syntactic dependency trees can well address and model such long-distance word correspondence. In Figure 1, if the dependency between the root word *"就医(see the doctor)"* and its subject word "患者(patients)" denoted by a link can be encoded by the NMT encoder, the NMT model is more likely to generate a correct translation.

In this paper, we address the above problem and propose to improve NMT by leveraging the source-side dependency tree to explicitly incorporate source word dependencies into NMT framework. Based on source dependency trees, we enrich each encoder state from both child-to-head and head-to-child with global knowledge from the dependency structure.

[*] Contribution during internship at Microsoft Research.



Reference : Patients should go to see the doctor with their family

NMT : Patients should take their families to see the doctor

Figure 1: Example of incorrect translation from conventional NMT system. The arrows refer to the dependency link in the dependency tree.

Two extra sequences are extracted with structural knowledge and encoded by another two RNNs which are used to improve the encoder states. With the enriched source states, the decoder generates target translation via attention mechanism in the same way as in most NMT models. We will describe our method in detail in Section 3. We evaluate our method on publicly available data sets with Chinese-English and English-Japanese translation tasks. Experimental results on Chinese-English task show that our model significantly improves translation accuracy over the conventional NMT and SMT baseline systems. Experiments on English-Japanese task also show that our method can achieve better performance than the state-of-the-art tree-based NMT model in [Eriguchi *et al.*, 2016b].

The major differences between our work and previous tree-based method [Eriguchi *et al.*, 2016b] are in two folds:

(1) We model source word relations that are important for understanding source sentences, however, they focus on the mismatch problem that one target word may attend to a source phrase (multiple consecutive words).

(2) Our model enhances the NMT by enriching each encoder state with global source dependency structure, however, they improve NMT model by proposing a phrase level attention.

2 Background

Different from SMT consisting of multiple sub-models, NMT is an end-to-end paradigm [Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015] directly modeling the conditional translation probability $p(Y|X)$ of the source sentence X = x_1,x_2,x_3,...,x_n and the target Y = y_1,y_2,y_3,...,y_m with the RNN encoder and the RNN decoder. The RNN encoder bidirectionally encodes the source sentence into a sequence of context vectors H = h_1,h_2,h_3,...,h_n, where $h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$, $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$ are calculated by two RNNs from left-to-right and right-to-left respectively as follows,

$$\overrightarrow{h}_i = f_{RNN}(x_i, \overrightarrow{h}_{i-1})$$
$$\overleftarrow{h}_i = f_{RNN}(x_i, \overleftarrow{h}_{i+1})$$

where f_{RNN} can be a Gated Recurrent Unit (GRU) [Cho *et al.*,] or a Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] in practice. In this paper, we use GRU for all RNNs.

Based on target history and the source context, the RNN decoder computes the target translation in sequence by

$$p(Y|X) = \prod_{j=1}^{m} p(y_j|y_{<j}, H) \quad (1)$$

Typically, for the jth target word, the probability $p(y_j|y_{<j}, H)$ is computed by

$$p(y_j|y_{<j}, H) = g(s_j, y_{j-1}, c_j) \quad (2)$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_i, s_j is the j-th hidden state of decoder RNN, computed by

$$s_j = f_{RNN}(y_{j-1}, s_{j-1}, c_j)$$

c_j is the source context which is calculated by the attention mechanism. The attention mechanism is proposed to softly align each decoder state with the encoder states, where the attention score a_{jk} is computed to explicitly quantify how much each source word contributes to the target word at each time step

$$a_{jk} = \frac{\exp(e_{jk})}{\sum_{d=1}^{n} \exp(e_{jd})} \quad (3)$$

The calculation for e_{jk} can be in several ways [Luong *et al.*, 2015b], in this paper we compute e_{jk} by

$$e_{jk} = v_a^T \tanh(W_a s_{j-1} + U_a h_k) \quad (4)$$

where v_a, W_a, U_a are the weight matrix. The final source context c_j is the weighted sum of all encoder states

$$c_j = \sum_{k=1}^{n} a_{jk} h_k \quad (5)$$

The overview of the attention-based NMT is shown in Figure 2. Although the attention mechanism is effective to model the correspondences between source and target, the long distance syntactic dependencies in the source-side still remain a challenged for a conventional NMT model.

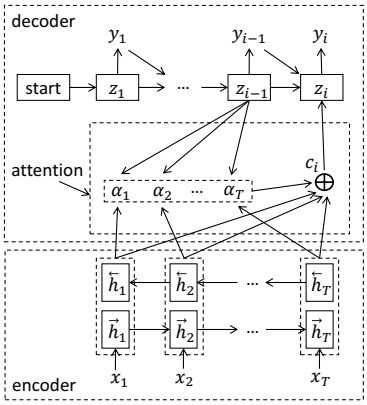


Figure 2: Overview of NMT framework with attention.

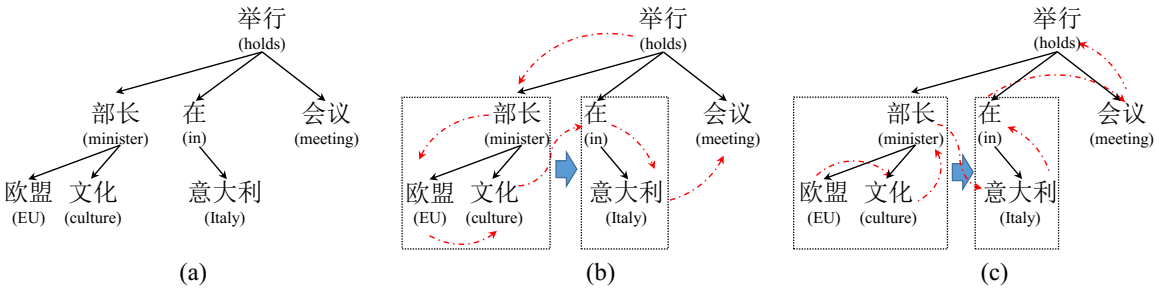


Figure 3: (a) The dependency tree for the Chinese sentence "欧盟*(EU)* 文化*(culture)* 部长*(minister)* 在*(in)* 意大利*(Italy)* 举行*(holds)* 会议*(meeting)*". (b) Child Enriched Structure. (c) Head Enriched Structure. The dashed arrows denote the construction of the two syntactic structures.

3 Our Method

To incorporate syntactic word relations into NMT, we propose to take advantage of the dependency tree to explicitly model source word dependencies for NMT encoder. Dependency tree is always used to characterize dependency relationships between words. Each word in the tree has a parent word which it depends on, except for the *root* word. There are no constituent labels in a dependency tree, the tree directly models word dependencies and syntactic structures of arbitrary distance. Figure 3 (a) gives an example of a dependency tree. Arrows point from the head node to its child nodes. Given a source sentence $X = x_1, x_2, .., x_j, .., x_n$, where n is the sentence length, and its corresponding dependency tree T, we denote w^h as a possible head node in T, w_l^h as the leftmost child node (or a leftmost subtree) of w^h, w_r^h as the rightmost child node (or a rightmost subtree) of w^h, and $w_1^h, ..., w_j^h$ as the rest child nodes (or subtrees) of w^h. All w belongs to X. Based on the dependency tree T, two kinds of dependency structures are extracted which are Child Enriched Structure (CES) and Head Enriched Structure (HES). In this section, we will introduce CES and HES respectively and how to incorporate them into NMT model.

3.1 Child Enriched Structure

In most previous work which attempts to leverage syntactic structure in neural networks such as [Tai *et al.*, 2015], a bottom-up fashion is used to construct representations for syntactic trees. That means leaf nodes are used as inputs for the construction of head nodes. This kind of encoding is good enough to make representations for the whole trees, which can facilitate tasks like sentiment classification, predicting the semantic relatedness of two sentences and so on. However, in the sequence to sequence generation tasks, especially neural machine translation, the generation of each target word may depend on arbitrary source word. Thus each source hidden state is expected to contain sufficient source information which can contribute to a better decoding. In contrast to the bottom-up fashion which leverages leaves to enrich heads, we propose a child enriched structure (CES) to enrich child nodes with global syntactic structures based on the dependency tree. Two kinds of context are defined in this structure

(1) w^h is a direct context for w_l^h.

(2) For a head node w^h, its former child nodes (or subtrees) are contexts for its latter child nodes (or subtrees). For example w_l^h is a direct context for w_1^h, w_1^h is a direct context for w_2^h.

Figure 3 (b) gives a brief introduction of the two kinds of context. For the head node "部长(minister)", it is a context for "欧盟(EU)" based on (1), and " 欧盟(EU)" is a context for "文化(culture)" denoted by the dashed arrow based on (2). When the child node is a sub-tree rather than a leaf node, for example, "部长(minister)" and "在(in)" are children of "举行(holds)", in this case the whole left-side subtree (left box) should be a context for building the right-side subtree (right box). To encode this kind of structure in NMT, we use another sequence generated by the pre-order traversal from the dependency tree. We find that this kind of traversal perfectly caters to both (1) and (2) as illustrated by the path of dashed arrows in Figure 3 (b).

3.2 Head Enriched Structure

In addition to child enriched structure, we also enrich the head nodes with its child nodes in the head enriched structure (HES). For this structure, another two kinds of context are defined,

(1) The first one is the same with the second one in CES.

(2) w_r^h is a direct context for w^h.

Figure 3 (c) gives a brief introduction of HES. For the construction of sub-tree in the left box, "欧盟(EU)" is first regarded as context for its neighbor "文化(culture)", then "文化(culture)" is used to enrich "部长(minister)". In addition, the former sub-tree in the left box is context for its neighbor sub-tree in the right box. To encode this kind of structure for NMT, we use another sequence generated by post-order traversal from the dependency tree which perfectly caters to the above description of HES as illustrated by the path of dashed arrows in Figure 3 (c).

3.3 The Computation in Encoder

We use two extra RNNs named CES-RNN and HES-RNN to encode the two structural sequences in addition to the bi-directional RNNs (bi-RNN). Thus for each source word x_j,

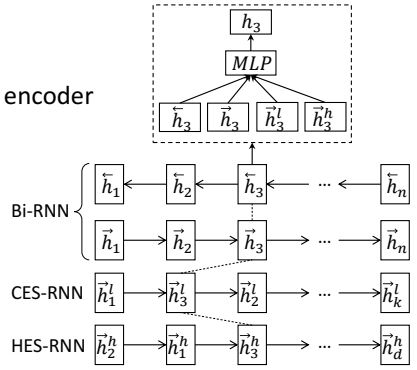


Figure 4: Overview of our encoder. The bottom two sequences are two possible sequences constructed according to CES and HES. We omit the dependency tree of the source sentence X.

we have four hidden state vectors generated by the encoder. We denote the two hidden vectors for word x_j from the bidirectional RNNs as $\vec{h}_j$ and $\overleftarrow{h}_j$, and denote $\vec{h}^l_j$ as the hidden vector from CES-RNN, $\vec{h}^h_j$ as the hidden vector from HES-RNN. The final hidden vector h_j used in the decoder is calculated by the four vectors. We do not directly concatenate them, because this may have the problem that the concatenated vector contains much more information than is necessary for decoding. We apply a MLP function with a smaller hidden size to the four recurrent states before the attention model, as below

$$h_j = \tanh(W_h\vec{h}_j + U_h\overleftarrow{h}_j + V_h\vec{h}^l_j + F_h\vec{h}^h_j) \qquad (6)$$

where W_h, U_h, V_h and F_h are weight matrices. This allows the model to combine the hidden vectors and filter out redundant information. The decoder and attention mechanism of our model remain the same with the conventional NMT model shown in Figure 2. We give an overview of the encoder in Figure 4. Due to space limitation, the detailed structure of our encoder is only illustrated at timestamp 3. The index k is the last word in the CES sequence which may not be the n-th word of X so as to d. Specially, the original sequence of the source sentence X is the in-order sequence of the dependency tree which indeed contains structural information from a linguistic perspective. In the following sections, we refer to our method as **S**ource **S**yntax-aware **NMT** (**SSNMT**) model.

4 Experiments

We conduct experiments on the Chinese-English translation task as well as the English-Japanese translation task where the same data set from WAT 2016 ASPEC corpus [Nakazawa *et al.*, 2016] [1] is used for a fair comparison with other work. In addition, we also evaluate the performance of our model in terms of source sentence length.

[1]http://orchid.kuee.kyoto-u.ac.jp/ASPEC/

4.1 Setup

In the Chinese-English translation task, the bilingual training data consists of a set of LDC datasets [2], which has 1M sentence pairs with around 24.5M Chinese words and 28.3M English words. The development data set is NIST2003, and the testing data are NIST2005, NIST2006, NIST2008 and NIST2012 evaluation sets. All the English words are lowercased in training and testing.

In the English-Japanese translation task, we use top 1.5M sentence pairs from ASPEC English-Japanese corpus. The development data contains 1,790 sentences, and the test data contains 1,812 sentences with single reference per source sentence. For the Japanese side, we employ KyTea [Neubig *et al.*,] as the segmentation method.

For the source-side dependency structures of both tasks, we use two in-house developed arc-eager dependency parsers based on work in [Zhang and Nivre, 2011] which are trained on Penn Treebank and Chinese Treebank data respectively to process the source data.

In the neural network training, we limit the vocabulary size to 30K high frequency words for both source and target languages. All low frequency words are normalized into a special token `unk` and post-processed by following the work in [Luong *et al.*, 2015b]. The size of word embeddings is set to 512 for both tasks. The dimensions of hidden states for all RNNs are set to 1024. All model parameters are initialized randomly with Gaussian distribution and trained on a NVIDIA Tesla K40 GPU. The stochastic gradient descent (SGD) algorithm is used to tune parameters with a learning rate of 1.0 and a batch size of 128. In the update procedure, Adadelta [Zeiler, 2012] algorithm is used to automatically adapt the learning rate. We use the beam search strategy for decoding with a beam size of 12.

Two baselines are used in our experiments which are a phrasal system and a neural translation system, denoted by HPSMT and RNNsearch respectively. HPSMT is an in-house implementation of the hierarchical phrase-based model [Chiang, 2005], where a 4-gram language model is trained using the modified Kneser-Ney smoothing [Kneser and Ney,] algorism over the English Gigaword corpus (LDC2009T13) plus the target data from the bilingual corpus. RNNsearch is an in-house implementation of the attention-based neural machine translation model [Bahdanau *et al.*, 2015] using the same parameter settings as illustrated before including word embedding size, hidden vector dimension, beam size, as well as the same mechanism for OOV word processing.

The evaluation results are reported with the case-insensitive IBM BLEU-4 [Papineni *et al.*, 2002]. A statistical significance test is performed using the bootstrap resampling method proposed by [Koehn, 2004] with a 95% confidence level. For English-Japanese task, we use the official evaluation procedure provided by WAT 2016.[3], where both BLEU and RIBES [Isozaki *et al.*, 2010] are used for evaluation. We also compare our method with previous tree-to-sequence model proposed by [Eriguchi *et al.*, 2016b].

[2]LDC2002E18, LDC2003E07, LDC2003E14, LDC2006E34, LDC2006E85, LDC2006E92, , LDC2004T07, LDC2004T08

[3]http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index .html

Settings	NIST 2005	NIST 2006	NIST 2008	NIST 2012	Average
HPSMT	35.53	33.16	26.21	27.32	30.56
RNNsearch	37.01	38.64	30.05	28.74	33.61
SSNMT\HES	37.46	39.03	30.78	29.12	34.10
SSNMT\CES	37.62	39.10	30.50	29.51	34.18
SSNMT	**38.01**	**39.85**	**31.01**	**29.82**	34.67

Table 1: Evaluation results on Chinese-English translation task with BLEU% metric. The "Average" column is the averaged result of all test sets. SSNMT\CES denotes SSNMT excluding CES and SSNMT\HES denotes SSNMT excluding HES.The numbers in bold indicate statistically significant difference ($p < 0.05$) from baselines.

	BLEU	RIBES	System Description
SMT Hiero	32.56	74.70	Moses' Hierarchical Phrase-based SMT
SMT Phrase	29.80	69.19	Moses' Phrase-based SMT
SMT T2S	33.44	75.80	Moses' Tree-to-String Syntax-based SMT
[Eriguchi *et al.*, 2016a]	31.52	79.39	Character-based decoder
[Luong *et al.*, 2015a]	34.64	81.60	Single model and single layer
[Eriguchi *et al.*, 2016a]($d = 512$)	34.91	81.66	Tree-to-string model
RNNsearch	34.83	80.92	Single model and single layer
SSNMT	35.85	81.64	Single model and single layer

Table 2: Evaluation results on English-Japanese translation task. All NMT methods are single models with single layer for both encoder and decoder. Results from ensemble models are not listed.

4.2 Evaluation on Chinese-English Translation

We first evaluate our method on the Chinese-English translation task. The evaluation results over all NIST test sets against baselines are listed in Table 1. SSNMT\HES denotes SSNMT excluding head enriched structure and only use CES, and SSNMT\CES refers to excluding head enriched structure. Generally, all NMT models outperform HPSMT on the average BLEU showing that NMT models usually achieve better results than SMT model. Compared with RNNsearch, our SSNMT with the two structural sequences performs better with a gain of about 1 BLEU point on average, which shows that the structural context provided by the two sequences bring a positive effect on a conventional NMT.

In addition, we also investigate the effects of the two sequences separately. According to Table 1, "SSNMT\HES" and "SSNMT\CES" can improve the performance of RNNsearch by about 0.49 and 0.57 BLEU point on average respectively. This demonstrates that the two sequences can bring positive effect on NMT from different perspectives where CES encodes structure information to each child node and HES in contrast enriches head nodes with its children.

4.3 Evaluation on English-Japanese Translation

In this section, we report results on the English-Japanese translation task. To have a fair comparison in the experiments, we use the same training data and follow the pre-processing steps recommended in WAT 2016[4] as well as the official evaluation procedure provided by WAT 2016.[5] Table 2 shows the comparison results from 8 systems with the evaluation metrics of BLEU and RIBES. The results in the first 3 rows are produced by SMT systems taken from the official WAT 2016. The remaining results are from NMT systems, among which the bottom two rows are taken from our in-house NMT systems and others refer to the results in [Eriguchi *et al.*, 2016a]. It notes that the English-Japanese translation result of [Luong *et al.*, 2015a] is also taken from [Eriguchi *et al.*, 2016a]. All the results are from single models without ensemble. According to Table 2, NMT results still outperform SMT results, which is similar to our Chinese-English evaluation results. Our RNNsearch can achieve comparable results with the NMT model in [Luong *et al.*, 2015a]. When adding both CES and HES, SSNMT outperforms all the other NMT models where 0.94 more BLEU point is achieved compared with the previous tree-to-sequence model in [Eriguchi *et al.*, 2016b]. This demonstrates that the source-side long distance dependencies captured by our method indeed have a positive effect on the translation performance.

4.4 Effect on Long Sentences

In this Section, we make a further comparison between our SSNMT and the RNNsearch baseline. As our method can take advantage of source long distance dependencies, it is more likely to generate complete translations and consistent meaning with source sentences even though the source length becomes longer. We evaluate the BLEU performance on the test sets of the two tasks with respect to the length of source sentences. Five groups of sentences are collected on the Japanese test set and the merged Chinese test set of NIST 2005, NIST 2006, NIST 2008 and NIST 2012, where source length ranges are {20-, 20-30, 30-40, 40-50, 50+}. The statistic of the five groups is shown in Table 3.

Figure 5 shows the evaluation results on the Chinese-English translation task. Clearly, our method always yields consistently higher BLEU scores than the RNNsearch baseline in terms of different lengths. When the length comes to "50+", our method outperforms the baseline most. This is because our method can encode source-side syntactic structures and provide more global knowledge for each source state

[4]http://lotus.kuee.kyoto-u.ac.jp/WAT/baseline/dataPreparationJE.html

[5]http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Length	CH-EN	EN-JE
20-	2,707	689
20-30	1,620	629
30-40	1,015	302
40-50	496	139
50+	381	53

Table 3: Data statistic of the five groups. 20- refers to lengths shorter than 20 and 50+ means lengths are longer than 50

which contributes to a better result. The BLEU results on the English-Japanese test set is shown in Figure 6. There is the same tendency for BLEU scores with the results on Chinese-English task. Notably, on the "50+" group, our method still outperforms the baseline most by a margin of 4.21 BLEU points. This again shows the effective of our method on long source sentences.

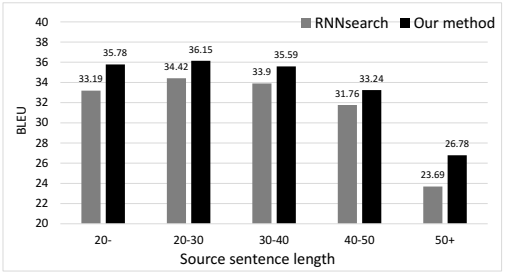


Figure 5: BLEU evaluation on the Chinese-English test set with respect to lengths of source sentence.

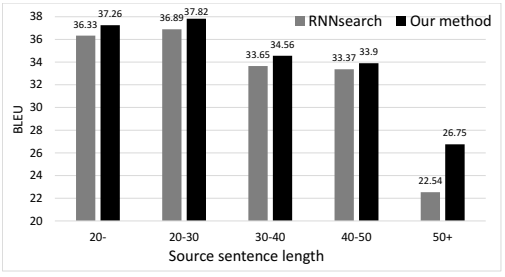
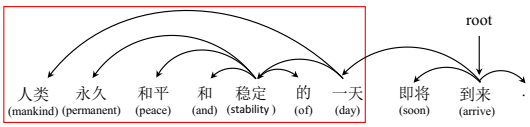


Figure 6: BLEU evaluation on the English-Japanese test set with respect to lengths of source sentence.

4.5 Translation Example

In this section, we give a case study to explain how our method works. Figure 7 gives a translation example from Chinese-English task. HPSMT and RNNsearch refer to the translation results form our SMT and NMT baselines. It is clear that the translations from NMT models are much better than SMT. Compared with RNNsearch, SSNMT generates a better translation. Though RNNsearch keeps most meaning of the source sentence, it fails to identify the subject of the whole sentence. In fact, the subject "一天(day)" has a long attribute which may interfere the translation but can be clearly modeled by the dependency tree (in the rectangle). Taking this kind of information into account, our SSNMT can generate a more correct translation.



[**Source**] 人类 永久 和平 和 稳定 的 一天 即将 到来 .
[**Reference**] the day when mankind can enjoy eternal peace and stability will soon arrive.
[**HPSMT**] human permanent peace and stability of the day dawns .
[**RNNsearch**] human beings permanent peace and stability will soon come here .
[**SSNMT**] the day of mankind 's permanent peace and stability will soon come .

Figure 7: Translation example from Chinese-English task. The top of the figure shows the dependency tree of the source sentence.

5 Related Work

Incorporating linguistic knowledge into machine translation has been extensively studied in Statistic Machine Translation (SMT) [Shen *et al.*, 2008; Liu *et al.*, 2006]. [Liu *et al.*, 2006] proposed a tree-to-string alignment template for SMT to leverage source side syntactic information. [Shen *et al.*, 2008] proposed a target dependency language model for SMT decoder based on the target-side dependency tree.These methods have successfully applied syntactic of either source or target to SMT and show promising improvement.

Recently, the attention-based Neural machine translation (NMT) becomes an emerging translation framework. The attention mechanism in NMT enables the model to translate while aligning each target with the source. However, in most existing NMT models, source sentences are treated as sequences where the syntactic knowledge is neglected. Some effort has been done to incorporate source syntax into NMT to enhance the attention model [Eriguchi *et al.*, 2016b; Hashimoto and Tsuruoka, 2017; Sennrich and Haddow, 2016]. [Eriguchi *et al.*, 2016b] proposed a tree-to-sequence attentional NMT model where source-side parse tree was used and achieved promising improvement. [Sennrich and Haddow, 2016] incorporated linguistic features to improve the NMT performance by appending feature vectors to word embeddings. [Hashimoto and Tsuruoka, 2017] proposed a multi-task framework to learn both source parsing and translation. Difference from previous syntax-based work, in this paper we focus on improve NMT encoder with source-side long-distance word dependencies.

6 Conclusion and Future Work

In this paper, we propose a simple but effective method to incorporate source dependency structure into NMT encoder. Our model can explicitly model word dependencies in the source sentence. Experimental results show that our method can achieve promising improvement over the conventional NMT model and outperform the state-of-the-art tree-to-string NMT model. In the future, along this research direction, we will try to effectively leverage more information from the dependency tree, such as arc-labels, pos-tag.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR 2015*, 2015.

[Chiang, 2005] David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, 2005.

[Cho *et al.*,] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of ENMLP 2014*.

[Eriguchi *et al.*, 2016a] Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. Character-based decoding in tree-to-sequence attention-based neural machine translation. In *Proceedings of WAT2016*, 2016.

[Eriguchi *et al.*, 2016b] Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. Tree-to-sequence attentional neural machine translation. In *Proceedings of ACL 2016*, August 2016.

[Hashimoto and Tsuruoka, 2017] Kazuma Hashimoto and Yoshimasa Tsuruoka. Neural machine translation with source-side latent graph parsing. *arXiv preprint arXiv:1702.02265*, 2017.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8), 1997.

[Isozaki *et al.*, 2010] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of EMNLP*, 2010.

[Jean *et al.*, 2015] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL 2015*, July 2015.

[Kneser and Ney,] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1. IEEE.

[Koehn, 2004] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395. Citeseer, 2004.

[Liu *et al.*, 2006] Yang Liu, Qun Liu, and Shouxun Lin. Tree-to-string alignment template for statistical machine translation. In *Proceedings of ACL 2006*, 2006.

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

[Luong *et al.*, 2015b] Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of ACL 2015*, July 2015.

[Nakazawa *et al.*, 2016] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia, may 2016. European Language Resources Association (ELRA).

[Neubig *et al.*,] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of ACL 2011*.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, 2002.

[Sag *et al.*, 1999] Ivan A Sag, Thomas Wasow, Emily M Bender, and Ivan A Sag. *Syntactic theory: A formal introduction*, volume 92. Center for the Study of Language and Information Stanford, CA, 1999.

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

[Shen *et al.*, 2008] Libin Shen, Jinxi Xu, and Ralph M Weischedel. A new string-to-dependency machine translation algorithm with a target dependency language model. In *ACL*, pages 577–585, 2008.

[Shen *et al.*, 2016] Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. In *Proceedings of ACL 2016*, August 2016.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 2014.

[Tai *et al.*, 2015] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[Wu *et al.*, 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

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

[Zhang and Nivre, 2011] Yue Zhang and Joakim Nivre. Transition-based dependency parsing with rich non-local features. In *Proceedings of ACL 2011*, 2011.