Chunk-Based Bi-Scale Decoder for Neural Machine Translation

Hao Zhou[*]
Nanjing University
zhouh@nlp.nju.edu.cn

Zhaopeng Tu[*]
Tencent AI Lab
tuzhaopeng@gmail.com

Shujian Huang
Nanjing University
huangsh@nlp.nju.edu.cn

Xiaohua Liu
Huawei Noah's Ark Lab
liuxiaohua3@huawei.com

Hang Li
Huawei Noah's Ark Lab
hangli.hl@huawei.com

Jiajun Chen
Nanjing University
chenjj@nlp.nju.edu.cn

Abstract

In typical neural machine translation (NMT), the decoder generates a sentence word by word, packing all linguistic granularities in the same time-scale of RNN. In this paper, we propose a new type of decoder for NMT, which splits the decode state into two parts and updates them in two different time-scales. Specifically, we first predict a chunk time-scale state for phrasal modeling, on top of which multiple word time-scale states are generated. In this way, the target sentence is translated hierarchically from chunks to words, with information in different granularities being leveraged. Experiments show that our proposed model significantly improves the translation performance over the state-of-the-art NMT model.

1 Introduction

Recent work of neural machine translation (NMT) models propose to adopt the *encoder-decoder* framework for machine translation (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014), which employs a recurrent neural network (RNN) *encoder* to model the source context information and a RNN *decoder* to generate translations, which is significantly different from previous statistical machine translation systems (Koehn et al., 2003; Chiang, 2005). This framework is then extended by an attention mechanism, which acquires source sentence context dynamically at different decoding steps (Bahdanau et al., 2014; Luong et al., 2015).

The decoder state stores translation information at different granularities, determining which segment should be expressed (phrasal), and which word should be generated (lexical), respectively. However, due to the extensive existence of multi-word phrases and expressions, the varying speed of the lexical component is much faster than the phrasal one. As in the generation of "*the French Republic*", the lexical component in the decoder will change thrice, each of which for a separate word. But the phrasal component may only change once. The inconsistent varying speed of the two components may cause translation errors.

Typical NMT model generates target sentences in the word level, packing the phrasal and lexical information in one hidden state, which is not necessarily the best for translation. Much previous work propose to improve the NMT model by adopting fine-grained translation levels such as the character or sub-word levels, which can learn the intermediate information inside words (Ling et al., 2015; Costa-jussà and Fonollosa, 2016; Chung et al., 2016; Luong et al., 2016; Lee et al., 2016; Sennrich and Haddow, 2016; Sennrich et al., 2016; García-Martínez et al., 2016). However, high level structures such as phrases has not been explicitly explored in NMT, which is very useful for machine translation (Koehn et al., 2007).

We propose a chunk-based bi-scale decoder for NMT, which explicitly splits the lexical and phrasal components into different time-scales.[1] The proposed model generates target words in a hierarchical way, which deploys a standard word time-scale RNN (lexical modeling) on top of an additional chunk time-scale RNN (phrasal modeling). At each step of decoding, our model first predict a chunk state with a *chunk attention*, based on which multiple word states are generated with-

[*]Work was done when Hao Zhou was interning and Zhaopeng Tu was working at Huawei Noah's Ark Lab.

[1]In this work, we focus on chunk-based well-formed phrases, which generally contain two to five words.

out attention. The word state is updated at every step, while the chunk state is only updated when the chunk boundary is detected by a *boundary gate* automatically. In this way, we incorporate *soft phrases* into NMT, which makes the model flexible at capturing both *global reordering* of phrases and *local translation* inside phrases. Our model has following benefits:

1. The chunk-based NMT model explicitly splits the lexical and phrasal components of the decode state for different time-scales, which addresses the issue of inconsistent updating speeds of different components, making the model more flexible.
2. Our model recognizes phrase structures explicitly. Phrase information are then used for word predictions, the representations of which are then used to help predict corresponding words.
3. Instead of incorporating source side linguistic information (Eriguchi et al., 2016; Sennrich and Haddow, 2016), our model incorporates linguistic knowledges in the target side (for deciding chunks), which will guide the translation more in line with linguistic grammars.
4. Given the predicted phrase representation, our NMT model could extract attentive source context by *chunk attention*, which is more specific and thus more useful compared to the word-level counterpart.

Experiments show that our proposed model obtains considerable BLEU score improvements upon an attention-based NMT baseline on the Chinese to English and the German to English datasets simultaneously.

2 Standard Neural Machine Translation Model

Generally, neural machine translation system directly models the conditional probability of the translation y word by word (Bahdanau et al., 2014). Formally, given an input sequence $\mathbf{x} = [x_1, x_2, \ldots, x_J]$, and the previously generated sequence $\mathbf{y_{<t}} = [y_1, y_2, \ldots, y_{t-1}]$, the probability of next target word y_t is

$$P(y_t|\mathbf{x}) = softmax(f(e_{y_{t-1}}, s_t, c_t)) \quad (1)$$

where $f(\cdot)$ is a non-linear function, $e_{y_{t-1}}$ is the embedding of y_{t-1}; s_t is the *decode state* at the time step t, which is computed by

$$s_t = g(s_{t-1}, e_{y_{t-1}}, c_t) \quad (2)$$

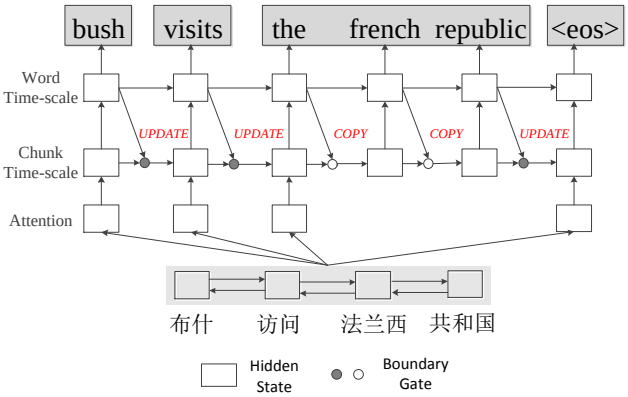

Figure 1: The architecture of the chunk-based bi-scale NMT.

Here $g(\cdot)$ is a transition function of decoder RNN. c_t is the context vector computed by

$$c_t = \sum_{j=1}^{J} \text{ATT}(s_{t-1}, h_j) \cdot h_j = \sum_{j=1}^{J} \alpha_{t,j} \cdot h_j \quad (3)$$

where ATT is an attention operation, which outputs alignment distribution α:

$$\alpha_{t,j} = \frac{exp(e_{t,j})}{\sum_{k=1}^{T_x} exp(e_{t,k})} \quad (4)$$

$$e_{t,j} = v_a^T \tanh(W_a s_{t-1} + U_a h_j) \quad (5)$$

and h is the annotation of $\mathbf{x}$ from a bi-directional RNNs. The training objective is to maximize the likelihood of the training data. Beam search is adopted for decoding.

3 Chunk-Based Bi-Scale Neural Machine Translation Model

Instead of the word-based decoder, we propose to use a chunk-based bi-scale decoder, which generates translation hierarchically with *chunk* and *word* time-scales, as shown in Figure 1. Intuitively, we firstly generate a chunk state with the attention model, which extracts the source context for the current phrasal scope. Then we generate multiple lexical words based on the same chunk state, which does not require attention operations. The boundary of a chunk is determined by a *boundary gate*, which decides whether to update the chunk state or not at each step.

Formally, the probability of next word y_t is

$$P(y_t|\mathbf{x}) = softmax(f(e_{y_{t-1}}, s_t, p_t)) \quad (6)$$

$$s_t = g(s_{t-1}, e_{y_{t-1}}, p_t) \quad (7)$$

here p_t is the chunk state at step t. Compared with Equations 1 and 2, the generation of target word is based on the chunk state instead of the context vector c_t produced by the attention model.

Since a chunk may correspond to multiple words, we employ a *boundary gate* b_t to decide the boundary of each chunk:

$$p(b_t) = softmax(s_{t-1}, e_{y_{t-1}}) \quad (8)$$

b_t will be 0 or 1, where 1 denotes this is the boundary of a new chunk while 0 denotes not. Two different operations would be executed:

$$p_t = \begin{cases} p_{t-1}, & b_t = 0 \text{ (COPY)} \\ g(p_{t-1}, e_{p_{t-1}}, pc_t), & b_t = 1 \text{ (UPDATE)} \end{cases}$$

In the COPY operation, the chunk state is kept the same as the previous step. In the UPDATE operation, $e_{p_{t-1}}$ is the representation of last chunk, which is computed by the LSTM-minus approach (Wang and Chang, 2016):

$$e_{p_{t-1}} = m(s_{t-1}, e_{y_{t-1}}) - m(s_{t'}, e_{y_{t'}}) \quad (9)$$

here t' is the boundary of last chunk and $m(\cdot)$ is a linear function. pc_t is the context vector for chunk p_t, which is calculated by a *chunk attention* model:

$$pc_t = \sum_{j=1}^{T_s} \text{ATT}(p_{t-1}, h_j) \cdot h_j \quad (10)$$

The chunk attention model differs from the standard word attention model (i.e., Equation 3) at: 1) it reads chunk state p_{t-1} rather than word state s_{t-1}, and 2) it is only executed at boundary of each chunk rather than at each decoding step.

In this way, our model only extracts source context once for a chunk, and the words in one chunk will share the same context for word generation. The chunk attention mechanism adds a constrain that target words in the same chunk shares the same source context.

Training To encourage the proposed model to learn reasonable chunk state, we add two additional objectives in training:

Chunk Tag Prediction: For each chunk, we predict the probability of its tag $P(l_k|\mathbf{x}) = softmax\big(f(p_t, e_{p_t}, c_t)\big)$, where l_k is the syntactic tag of the k-th chunk such as *NP* (noun phrase) and *VP* (verb phrase), and t is time step of its boundary.

Chunk Boundary Prediction: At each decoding step, we predict the probability of chunk boundary $P(b_t|\mathbf{x}) = softmax(s_{t-1}, e_{y_{t-1}})$.

Accordingly, given a set of training examples $\{[\mathbf{x}_n, \mathbf{y}_n]\}_{n=1}^N$, the new training objective is

$$J(\theta, \gamma) = \arg\max \sum_{n=1}^{N} \Big\{ \log P(\mathbf{y_n}|\mathbf{x_n}) + \log P(\mathbf{l_n}|\mathbf{x_n}) + \log P(\mathbf{b_n}|\mathbf{x_n}) \Big\} \quad (11)$$

where $\mathbf{l_n}$ and $\mathbf{b_n}$ are chunk tag sequence and boundary sequence on $\mathbf{y_n}$, respectively.

4 Experiments

We carry out experiments on a Chinese-English translation task. Our training data consists of 1.16M[2] sentence pairs extracted from LDC corpora, with 25.1M Chinese words and 27.1M English words, respectively. We choose the NIST 2002 (MT02) dataset as our development set, and the NIST 2003 (MT03), 2004 (MT04) 2005 (MT05) datasets as our test sets. We also evaluate our model on the WMT translation task of German-English, newstest2014 (DE14) is adopted as development set and newstest2012, newstest2013 (DE1213) are adopted as testing set. The English sentences are labeled by a neural chunker, which is implemented according to Zhou et al. (2015). We use the case-insensitive 4-gram NIST BLEU score as our evaluation metric (Papineni et al., 2002).

In training, we limit the source and target vocabularies to the most frequent 30K words. We train each model with the sentences of length up to 50 words. Sizes of the chunk representation and chunk hidden state are set to 1000. All the other settings are the same as in Bahdanau et al. (2014).

4.1 Results on Chinese-English

We list the BLEU score of our proposed model in Table 1, comparing with Moses (Koehn et al., 2007) and dl4mt[3] (Bahdanau et al., 2014), which are state-of-the-art models of SMT and NMT, respective. For Moses, we use the default configuration with a 4-gram language model trained on the target portion of the training data. For dl4mt, we also report the results (dl4mt-2) by using two

[2] 3LDC2002E18, LDC2003E14, the Hansards portion of LDC2004T08, and LDC2005T06.

[3] https://github.com/nyu-dl/dl4mt-tutorial

System	MT02	MT03	MT04	MT05	Ave.
Moses	30.10	28.82	31.22	27.78	29.48
dl4mt	31.66	29.92	32.76	28.88	30.81
dl4mt-2	31.01	28.74	31.71	27.95	29.85
This Work	**33.43**	**32.06**	**34.21**	**30.01**	**32.42**

Table 1: BLEU scores for different systems.

Attention	MT02	MT03	MT04	MT05	Ave.
Word	32.69	31.36	33.55	29.77	31.56
Chunk	33.43	32.06	34.21	30.01	32.42

Table 2: Results with different attention models.

decoder layers (Wu et al., 2016) for better comparison.

As shown in Table 1, our proposed model outperforms different baselines on all sets, which verifies that the chunk-based bi-scale decoder is effective for NMT. Our model gives a 1.6 BLEU score improvement upon the standard NMT baseline (dl4mt). We conduct experiment with dl4mt-2 to see whether the neural NMT system can model the bi-scale components with different varying speeds automatically. Surprisingly, we find that dl4mt-2 obtains lower BLEU scores than dl4mt. We speculate that the more complex model dl4mt-2 may need more training data for obtaining reasonable results.

Effectiveness of Chunk Attention As described in Section 3, we propose to use the *chunk attention* to replace the word level attention in our model, in which the source context extracted by the chunk attention will be used for the corresponding word generations in the chunk. We also report the result of our model using conventional word attention for comparison. As shown in Table 2, our model with the chunk attention gives higher BLEU score than the word attention.

Intuitively, we think chunks are more specific in semantics, thus could extract more specific source context for translation. The chunk attention could be considered as a compromise approach between encoding the whole source sentence into decoder without attention (Sutskever et al., 2014) and utilizing word level attention at each step (Bahdanau et al., 2014). We also draw the figure of alignments by chunk attention (Figure 2), from which we can see that our chunk attention model can well explore the alignments from phrases to words.

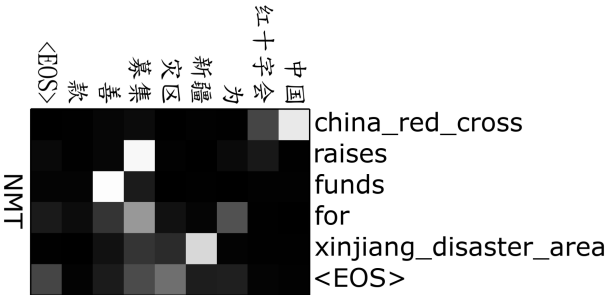

Figure 2: Alignments with chunk attention.

	MT02	MT03	MT04	MT05
Boundary	89.97	88.81	89.64	89.25
Label	47.00	44.75	45.54	44.41

Table 3: Accuracies of predicted chunk boundary and chunk label.

Predictions of the Chunk Boundary and Chunk Label We also compute predicted accuracies of chunk boundaries and chunk labels on the auto-chunked development and testing data (Table 3). We find that the chunk boundary could be predicted well, with an average accuracy of 89%, which shows that our model could capture the phrasal boundary information in the translation process. However, our model could not predict chunk labels as well as chunk boundaries. We speculate that more syntactic context features should be added to improve the performance of predicting chunk labels.

Subjective Evaluation Following Tu et al. (2016, 2017a,b), we also compare our model with the dl4mt baseline by subjective evaluation. Two human evaluators are asked to evaluate the translations of 100 source sentences randomly sampled from the test sets without knowing which system

Model	dl4mt	Our Work
Adequacy	3.26	3.35
Fluency	3.69	3.71
Under-Translation	50%	47%
Over-Translation	32%	26%

Table 4: Subjective evaluation results.

System	DE-14	DE-1213
dl4mt	16.53	16.78
This Work	17.40	17.45

Table 5: Results on German-English

the translation is translated by. The human evaluator is asked to give 4 scores: adequacy score and fluency score, which are between 0 and 5, the larger, the better; under-translation score and over-translation score, which are set to 1 when under or over translation errors occurs, otherwise set to 0.

We list the averaged scores in Table 5. We find that our proposed model improves the dl4mt baseline on both the translation adequacy and fluency aspects. Specifically, the over translation error rate drops by 6%, which confirms the assumption in the introduction that splitting the fast and slow varying components in different time-scales could help alleviate the over translation errors.

4.2 Results on German-English

We evaluate our model on the WMT15 translation task from German to English. We find that our proposed chunk-based NMT model also obtains considerable accuracy improvements on German-English. However, the BLEU score gains are not as significant as on Chinese-English. We speculate that the difference between Chinese and English is larger than German and English. The chunk-based NMT model may be more useful for bilingual data with bigger difference.

5 Related Work

NMT with Various Granularities. A line of previous work propose to utilize other granularities besides words for NMT. By further exploiting the character level (Ling et al., 2015; Costa-jussà and Fonollosa, 2016; Chung et al., 2016; Luong et al., 2016; Lee et al., 2016), or the sub-word level (Sennrich and Haddow, 2016; Sennrich et al., 2016; García-Martínez et al., 2016) information, the corresponding NMT models capture the information inside the word and alleviate the problem of unknown words. While most of them focus on decomposing words into characters or sub-words, our work aims at composing words into phrases.

Incorporating Syntactic Information in NMT Syntactic information has been widely used in SMT (Liu et al., 2006; Marton and Resnik, 2008; Shen et al., 2008), and a lot of previous work explore to incorporate the syntactic information in NMT, which shows the effectiveness of the syntactic information (Stahlberg et al., 2016). Shi et al. (2016) give some empirical results that the deep networks of NMT are able to capture some useful syntactic information implicitly. Luong et al. (2016) propose to use a multi-task framework for NMT and neural parsing, achieving promising results. Eriguchi et al. (2016) propose a string-to-tree NMT system by end-to-end training. Different to previous work, we try to incorporate the syntactic information in the target side of NMT. Shonosuke et al. (2017) concurrently propose to use chunk-based encoder for NMT, which utilizes the chunk structure to efficiently capture long-distance dependencies and cope with the problem of free word-order languages like Japanese.

6 Conclusion

We propose a chunk-based bi-scale decoder for neural machine translation, in which way, the target sentence is translated hierarchically from chunks to words, with information in different granularities being leveraged. Experiments show that our proposed model outperforms the standard attention-based neural machine translation baseline. Future work includes abandoning labeled chunk data, adopting reinforcement learning to explore the boundaries of phrase automatically (Mou et al., 2016). Our code is released on `https://github.com/zhouh/chunk-nmt`.

Acknowledge

We would like to thank the anonymous reviewers for their insightful comments. We also thank Lili Mou for helpful discussion and Hongjie Ji, Zhenting Yu, Xiaoxue Hou and Wei Zou for their help in data preparation and subjective evaluation. This work was partially founded by the Natural Science Foundation of China (61672277, 71503124) and the China National 973 project 2014CB340301.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

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.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 1724–1734. https://doi.org/10.3115/v1/D14-1179.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1693–1703. http://www.aclweb.org/anthology/P16-1160.

R. Marta Costa-jussà and R. José A. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 357–361. https://doi.org/10.18653/v1/P16-2058.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 823–833. https://doi.org/10.18653/v1/P16-1078.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation. *arXiv preprint arXiv:1609.04621* .

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1700–1709. http://aclweb.org/anthology/D13-1176.

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.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017* .

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586* .

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 609–616.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico.

Thang Luong, Hieu Pham, and D. Christopher Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1412–1421. https://doi.org/10.18653/v1/D15-1166.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *ACL*. pages 1003–1011.

Lili Mou, Zhengdong Lu, Hang Li, and Zhi Jin. 2016. Coupling distributed and symbolic execution for natural language queries. *arXiv preprint arXiv:1612.02741* .

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 83–91. http://www.aclweb.org/anthology/W16-2209.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162.

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

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1526–1534. http://aclweb.org/anthology/D16-1159.

Ishiwatari Shonosuke, Yao Jingtao, Liu Shujie, Li Mu, Zhou Ming, Yoshinaga Naoki, Kitsuregawa Masaru, and Jia Weijia. 2017. Chunk-based decoder for neural machine translation. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne. 2016. Syntactically guided neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 299–305. https://doi.org/10.18653/v1/P16-2049.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017a. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics* 5:87–99.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017b. Neural machine translation with reconstruction. In *Proceedings of AAAI 2017*. pages 3097–3103.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 76–85. https://doi.org/10.18653/v1/P16-1008.

Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional lstm. In *Proceedings of ACL*. volume 1, pages 2306–2315.

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

Hao Zhou, Yue Zhang, Shujian Huang, and Jiajun Chen. 2015. A neural probabilistic structured-prediction model for transition-based dependency parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1213–1222. http://www.aclweb.org/anthology/P15-1117.