NICT's Neural Machine Translation Systems for the WAT21 Restricted Translation Task

Zuchao Li[1,2,3], **Masao Utiyama**[4,*], **Eiichiro Sumita**[4], **and Hai Zhao**[1,2,3*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[4]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
charlee@sjtu.edu.cn, {mutiyama, eiichiro.sumita}, zhaohai@cs.sjtu.edu.cn

Abstract

This paper describes our system (Team ID: nictrb) for participating in the WAT'21 restricted machine translation task. In our submitted system, we designed a new training approach for restricted machine translation. By sampling from the translation target, we can solve the problem that ordinary training data does not have a restricted vocabulary. With the further help of constrained decoding in the inference phase, we achieved better results than the baseline, confirming the effectiveness of our solution. In addition, we also tried the vanilla and sparse Transformer as the backbone network of the model, as well as model ensembling, which further improved the final translation performance.

1 Introduction

The performance of machine translation has been greatly improved since it entered the era of Neural Machine Translation (NMT) (Bahdanau et al., 2015; Sutskever et al., 2014; Wu et al., 2016). Different from traditional statistical machine translation (SMT) (Koehn et al., 2003), NMT models are trained end-to-end with contextualized representations to alleviate the locality problem and dense representations to mitigate the sparsity issue. The incorporation of novel structures such as CNN (Gehring et al., 2017) and Transformer (Vaswani et al., 2017) into NMT has brought the performance one step closer to practical translation.

Though NMT can more effectively exploit large parallel corpora, the performance is still insufficient to meet the requirements in some special translation scenarios. The end-to-end NMT models remove many approaches in the SMT paradigm for manually guiding the translation process. One attractiveness of the SMT method is that it provides explicit control over translation output, which is effective in a variety of translation settings, including interactive machine translation (Peris et al., 2017) and domain adaptation (Chu and Wang, 2018), which is also crucial for the practical application of NMT.

Since there is still a need for manual interventions for the new NMT paradigm, much effort is spent in studying how to incorporate this explicit control into the end-to-end neural translation (Arthur et al., 2016). Among these efforts, Constrained Decoding (CD) has gained a lot of attention in this research field, which is a modification to commonly adopted beam search in ordinary NMT models. Hokamp and Liu (2017) proposed *grid beam search*, which expands beam search to include pre-specified lexical constraints. Anderson et al. (2017) used *constrained beam search* to force the inclusion of restricted words in the output, and employed fixed pre-trained word embeddings to facilitate vocabulary expansion to unseen words in training.

While these works accomplish the goal of explicit translation control, the time complexity of their decoding algorithm and resultant decoding speed falls short of the expectations. The complexity of *grid beam search* and *constrained beam search* is linear and exponential to the number of constraints, respectively. These algorithms are thus too inefficient to be practical for large-scale use. To alleviate the shortcomings in constrained decoding, Post and Vilar (2018) proposed a new constrained decoding algorithm with a claimed complexity of $O(1)$ in the number of constraints - *dynamic beam allocation* which allocates the slots in a fixed-size beam. However, their approach still processes sentence constraints sequentially rather than batch processing, limiting the GPU's parallel processing capabilities. Based on Post and Vilar (2018), a *vectorized dynamic beam allocation* approach was proposed in Hu et al. (2019), which which vector-

[*] Corresponding author. This work was completed when Zuchao Li was a fixed-term technical researcher at NICT.

Proceedings of the 8th Workshop on Asian Translation, pages 62–67
Bangkok, Thailand (online), August 5-6, 2021. ©2021 Association for Computational Linguistics

izes the *dynamic beam allocation* for batching and thus leading to improvement in throughput with parallelization. Based on Post and Vilar (2018), Hu et al. (2019) proposed a *vectorized dynamic beam allocation* approach, which vectorizes the *dynamic beam allocation* for batching, resulting in increased throughput with parallelization.

Constrained decoding is a very general method for incorporating additional translation knowledge into the output without modifying the model parameters or training data. However, the model's prediction distribution can be skewed during the decoding process with hard constraints, resulting in poor translation results. When the model is exposed to the restricted translation paradigm during training, the gap between training and inference can be reduced, potentially improving performance. Therefore, in this paper, we propose a training method of *Sampled Constraints as Concentration* (SCC). In this method, training data is the same as the ordinary NMT; only minor modifications on the loss calculation are required to adapt the model to restricted translation.

In our submission to WAT'21 (Nakazawa et al., 2021) restricted translation task, we chose Transformer (Vaswani et al., 2017) as our baseline because of its high performance and scalability. Although there are some variants, our previous experiments have shown there are not too many approaches that can be both concise and effective. At the same time, though multi-head self-attention in Transformer can model extremely long dependencies, deep layer attention tends to overconcentrate on a single token, resulting in inadequate use of local information and difficulty representing long sequences. To address this disadvantage, we employ the PRIME Transformer (Zhao et al., 2019) with a multi-scale sparse attention mechanism as a second baseline. The models in the two architectures are ensembled to improve the overall results. Our final system uses a combination of the SCC training method and the constrained decoding of Hu et al. (2019), which makes our system leverages soft constrained (inside the model) and benefit from hard restrictions (external decoding).

2 Our System

In this section, we describe the methods used in our system in detail. Our system is made up of four components: the Transformer model, the Sparse Transformer model, the SCC training approach, and the constrained decoding algorithm. In translation, given the source input sequence $X = \{w_1, w_2, ..., w_m\}$, its target translation is $Y = \{y_1, y_2, ..., y_n\}$, the parameter of the NMT model is θ, then the probability form of the translation process can be written as:

$$P(Y|X, \theta) = \prod_{i=1}^{n} P(y_i|y_{<i}, X, \theta),$$

where $y_{<i}$ denotes the tokens generated before time step i.

2.1 Transformer Model

Transformer model (Li et al., 2021) is a encoder-decoder architecture entirely built on multi-head self-attention which is responsible for learning representations of global context. With an input representation H, a multi-head self-attention (MHA) layer first projects H into three representations, key K, query Q, and value V. Then, it uses a self-attention mechanism to get the output representation:

$$head_k = \text{Attn}(H) = \sigma(QW^Q, KW^K, VW^V)W^O$$
$$\text{MHA}(H) = \text{Concat}(head_1, \cdots, head_{\mathcal{K}})W^O,$$

where $Q = \text{Linear}_Q(H)$, $K = \text{Linear}_K(H)$, $V = \text{Linear}_V(H)$, W^O, W^Q, W^K, and W^V are projection parameters. The self-attention operation σ is the dot-production between key, query, and value pairs:

$$\sigma(Q_1, K_1, V_1) = \text{Softmax}(\frac{Q_1 K_1^T}{\sqrt{d_k}})V_1,$$

where $d_k = d_{model}/\mathcal{K}$ is the dimension of each head. The encoder of the Transformer model consists of a stack of multiple layers with MHA structure (Self-MHA$_{enc}$) where the residual mechanism and layer normalization are used to connect two adjacent layers. Similar to the encoder, each decoder layer decoder is composed of two MHA structures: Self-MHA$_{dec}$ and Cross-MHA, since it not only encodes the input sequence but also incorporates the source representation. Then the processing flow of the model can be written as:

$$H_{enc} = \text{Self-MHA}_{enc}(X),$$

$$H_{dec} = \text{Self-MHA}_{dec}(\text{IncMask}([\text{BOS}, y_1, \cdots, y_{n-1}])),$$
$$P(Y|X) = \text{Softmax}(\text{Linear}(\text{Cross-MHA}(H_{dec}, H_{enc}))),$$

where IncMask$(\cdot)$ represents the incremental masking strategy.

2.2 Sparse Transformer Model

According to the evaluation in recent research (Tang et al., 2018), it has shown that the vanilla Transformer has surprising shortcomings in long sequence encoding even the Transformer is designed to modeling long dependencies. Vanilla Transformer works well for short sequence translation, but performance drops as the source sentence length increases because only a small number of tokens are represented by self-attention, resulting in difficulty for translation. Replacing the dense self-attention mechanism with a sparse attention mechanism will alleviate the difficulties in long sentence translation; we chose the PRIME Transformer (Zhao et al., 2019) as our another base model. Compared to vanilla Transformer, PRIME Transformer generates the output representation of layer i in a fusion way:

$$\begin{aligned} H^i = {} & H^{i-1} + \text{MHA}(H^{i-1}) \\ & + \text{Conv}(H^{i-1}) + \text{Pointwise}(H^{i-1}), \end{aligned}$$

where H^{i-1} is the output of layer $i-1$. Conv$(\cdot)$ refers to dynamic convolution with multiple kernel sizes, which is employed to capture local context:

$$\begin{aligned} \text{Conv}_k(H) &= \text{DepthConv}_k(HW^V)W^{out} \\ \text{DepthConv}_k(H) &= \sum_{j=1}^{k} \big(\text{Softmax}(\sum_{c=1}^{d} W^Q_{j,c} H_{i,c}) \\ &\quad \cdot H_{i+j-\lceil \frac{k+1}{2} \rceil, c}\big), \\ \text{Conv}(H) &= \sum_{i=1}^{\mathcal{K}} \frac{\exp{(\alpha_i)}}{\sum\limits_{j=1}^{n} \exp{(\alpha_j)}} \text{Conv}_{k_i}(X) \end{aligned}$$

in which DepthConv$(\cdot)$ is the depth convolution structure proposed in Wu et al. (2019). And Pointwise$(\cdot)$ refers to a position-wise feed-forward network:

$$\text{Pointwise}(H) = max(0, HW_1 + b_1)W_2 + b_2.$$

where W_1, b_1, W_2, and b_2 are learnable parameters.

2.3 Sampled Constraints as Concentration Training

The predicted probability in ordinary NMT is $y_i \sim P(y_i|X, \theta)$. Because of the inclusion of the constrained word sequence C in restricted translation, the probability distribution becomes $y_i \sim P(y_i|X, C, \theta)$. To adapt the restricted translation for the NMT model rather than just influencing the search process, we expose the constrained word sequence C as additional context like source input.

Since the parallel training data only contains the source and target language sequences, we obtain the constrained word sequence for training via random dynamic sampling from the reference target translation. This not only alleviates the burden of constrained word annotation but also has the potential to minimize overfitting.

Specifically, in the model, we use the Self-MHA$_{dec}$ to encode the input constrained sequence to obtain its representation:

$$H_{cst} = \text{Self-MHA}_{dec}(C).$$

It is worth noting that we remove the positional encoding of constrained sequence since the order of restricted word sequence is usually inconsistent with the target translation; additionally, we also remove the incremental mask because the whole sequence is exposed to the decoder as an additional context at the same time. The probabilistic form of restricted translation accordingly changes to:

$$\begin{aligned} P(Y|X) = \text{Softmax}(\text{Linear}(&\text{Cross-MHA}(H_{dec}, H_{enc}) + \\ &\text{Cross-MHA}(H_{dec}, H_{cst})))). \end{aligned}$$

Because sampled constrained words are exposed to the decoder, to enforce the inclusion of these words in the translation, we place additional penalties on the loss of these sampled positions to achieve the goal of restrict translation with soft constraints on the model:

$$\begin{aligned} \mathcal{L}_{\text{SCC}} = -\sum_{i=1}^{m} \big(& (1 + \gamma \mathbb{1}(y_i \in C)) \\ & logP(y_i|X; C; y_{<i}; \theta) \big), \end{aligned}$$

where $\mathbb{1}(\cdot)$ is the indicator function and γ is the penalty factor.

2.4 Lexically Constrained Decoding

Beam search (Koehn, 2010) is a common approximate search algorithm for sequence generation task. Lexically constrained decoding is a modification to the beam search algorithm, which is proposed to enforce hard constraints that force a given constrained sequence to appear in the generated sequence. Specifically, beam search maintains a beam B_t on time step t, which contains only the b most likely partial sequences, where b is known as the beam size. The beam B_t is updated by retaining the b most likely sequences in the candidate set E_t generated by considering all possible next word predictions:

$$E_t = \big\{(\hat{Y}_{t-1}, w) \mid \hat{Y}_{t-1} \in B_{t-1}, w \in \mathcal{V}\big\},$$

Model	BLEU			RIBES			AMFM
	jum	*kyt*	*mec*	*jum*	*kyt*	*mec*	–
Transformer-big	41.67	41.82	41.84	81.05	81.32	81.50	74.95
Transformer-big + SCC + CD*	48.92	49.24	49.25	82.79	83.15	83.57	79.15
Sparse Transformer-big + SCC + CD*	50.93	51.18	51.21	83.27	83.52	84.00	79.91
Ensemble*	51.07	51.32	51.36	83.68	83.99	84.41	79.99

Table 1: Results on ASPECT En→Ja test sets. $*$ indicates that the official evaluation results are reported.

Dataset	Sentences
ParaCrawl-v5.1	10.12M
Wiki Titles v2	3.64M
ASPEC	3.01M

Table 2: Training data statistics.

Model	BLEU	RIBES	AMFM
Transformer-big	28.18	67.79	58.69
Transformer-big + SCC + CD*	35.26	74.44	64.16
Sparse Transformer-big + SCC + CD*	36.83	75.84	65.29
Ensemble*	37.01	75.38	65.15

Table 3: Results on ASPECT Ja→En test sets. $*$ indicates that the official evaluation results are reported.

where $\hat{Y}_{t-1}$ is the generated sequence in time step $t-1$ and $\mathcal{V}$ is the target vocabulary.

In lexically constrained decoding, a finite-state machine (FSM) is used to impose the constraints. For each state $s \in S$ in the FSM, a corresponding search beam B^s is maintained similar to the beam search:

$$E_t^s = \bigcup_{s' \in S} \{(\hat{Y}_{t-1}, w) \mid \hat{Y}_{t-1} \in B_{t-1}^{s'}, w \in V, \\ \delta(s', w) = s\},$$

where $\delta : S \times V \mapsto S$ is the FSM state-transition function that maps states and predicted words to states.

2.5 System Details

Our implementation of the Transformer models and lexically constrained decoding algorithm are based on the Fairseq toolkit[1]. We follow the settings and pre-processing methods in our previous models and systems (He et al., 2018; Li et al., 2018; He et al., 2019; Li et al., 2019; Zhou et al., 2020; Li et al., 2020b,d,c; Zhang et al., 2020). We use Transformer-big as our basic model, which has 6 layers in both the encoder and decoder, respectively. For each layer, it consists of a multi-head attention sublayer with 16 heads and a feed-forward sublayer with an inner dimension 4096. The word embedding dimensions and the hidden state dimensions are set to 1024 for both the encoder and decoder. In the training phase, the dropout rate is set to 0.1.

Our model training consists of two phases. In the first NMT pre-training phase, the ParaCrawl-v5.1 (Esplà et al., 2019) and Wiki Titles v2 datasets are used. Then we finetune the model using the ASPEC training data in the second domain finetune phase. Table 2 shows the data statistics for each dataset. In both phases, cross-entropy with label smoothing of 0.1 and D2GPo (Li et al., 2020a) are employed as the training loss criterions. We use Adam (Kingma and Ba, 2015) as our optimizer, with parameters settings $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-8}$. The initial learning rate is set to 10^{-4} for NMT pre-training and 10^{-5} for domain finetuning. The models are trained on 8 GPUs for about 500,000 steps. In our systems, we follow standard practice and learn a subword (Sennrich et al., 2016) encoding with 40K joint merge operations.

3 Results

Table 1 shows the official results evaluated on ASPEC En→Ja test set. Comparing the results of the vanilla Transformer-big model and Transformer-big+SCC+CD, restricted translation under +SCC+CD has brought a very large performance improvement, which illustrates the performance advantage of restricted translation. Similar to ordinary NMT, sparse Transformer achieves better results than Transformer-big in restricted translation, which demonstrates that Sparse Transformer is a general model structure. A further increase in performance is achieved after ensembling on these two models. This benefits from the models of the distinct architectures of the two models. In general, the improvement brought about by the same architecture is less. We show the results of ASPEC En→Ja test set in Table 3. By comparison, the conclusion is essentially consistent with Table 2.

[1] https://github.com/pytorch/fairseq

4 Conclusion

In this paper, we present our NMT systems for WAT21 restricted translation shared tasks in English ↔ English. By integrating the following techniques: Sparse Transformer, Sampled Constraints as Concentration, and Lexically Constrained Decoding, our final system achieves substantial improvement over baseline systems which show the effectiveness of our approaches.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5350–5359, Hong Kong, China. Association for Computational Linguistics.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2061–2071, Melbourne, Australia. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2411, Brussels, Belgium. Association for Computational Linguistics.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020a. Data-dependent gaussian prior objective for language generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao.

2020b. Explicit sentence compression for neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8311–8318. AAAI Press.

Zuchao Li, Zhuosheng Zhang, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2021. Text compression-aided transformer encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020c. SJTU-NICT's supervised and unsupervised neural machine translation systems for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 218–229, Online. Association for Computational Linguistics.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020d. Reference language based unsupervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, Online. Association for Computational Linguistics.

Zuchao Li, Hai Zhao, Zhuosheng Zhang, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2019. SJTU-NICT at MRP 2019: Multi-task learning for end-to-end uniform semantic graph parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 45–54, Hong Kong. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. Association for Computational Linguistics.

Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2017. Interactive neural machine translation. *Comput. Speech Lang.*, 45:201–220.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4263–4272, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Guangxiang Zhao, Xu Sun, Jingjing Xu, Zhiyuan Zhang, and Liangchen Luo. 2019. MUSE: parallel multi-scale attention for sequence to sequence learning. *CoRR*, abs/1911.09483.

Junru Zhou, Zuchao Li, and Hai Zhao. 2020. Parsing all: Syntax and semantics, dependencies and spans. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4438–4449, Online. Association for Computational Linguistics.