Content Word Aware Neural Machine Translation

Kehai Chen, Rui Wang,* Masao Utiyama, and Eiichiro Sumita
National Institute of Information and Communications Technology (NICT), Kyoto, Japan
{khchen, wangrui, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

Neural machine translation (NMT) encodes the source sentence in a universal way to generate the target sentence word-by-word. However, NMT does not consider the importance of word in the sentence meaning, for example, some words (i.e., content words) express more important meaning than others (i.e., function words). To address this limitation, we first utilize word frequency information to distinguish between content and function words in a sentence, and then design a content word-aware NMT to improve translation performance. Empirical results on the WMT14 English-to-German, WMT14 English-to-French, and WMT17 Chinese-to-English translation tasks show that the proposed methods can significantly improve the performance of Transformer-based NMT.

1 Introduction

Neural machine translation (**NMT**) models (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) often utilize the global neural networks to encode all words for learning the sentence representation and the context vector, and computes the accuracy of each generated target word in a universal manner. Meanwhile, each generated target word makes the same contribution to the optimization of the NMT model, regardless of its importance. Actually, there lacks a mechanism to guarantee that NMT captures the information related to word importance when predicting translations.

Intuitively, content words express more important meanings than function words, which indicates their comparative significance. To evaluate this, we randomly masked content or function words with UNK in a source sentence. Figure 1 shows that the BLEU scores of the test set decreased much more substantially when parts of content words were randomly replaced with UNK on the WMT14 English-to-German task, which is in line with the findings in He et al. (2019)'s work.

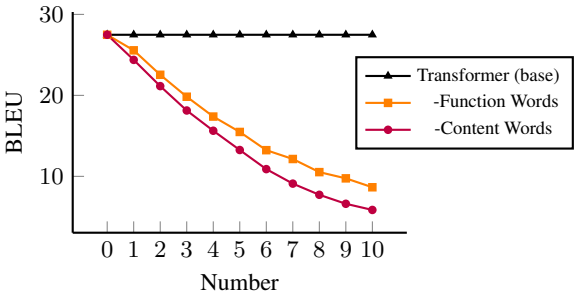


Figure 1: "Number" denotes the number of content or function words that were randomly masked in each sentence of the WMT14 English-to-German translation task.

To address this limitation, we propose a content word-aware NMT model that exploits the results of translation using a sequence of content words learned by a simple content word recognition method. Inspired by the works of (Setiawan et al., 2007, 2009; Zhang and Zhao, 2013), we first divide words in a sentence into content words and other function words depending on term frequency-inverse document frequency (**TF-IDF**) constraints. Two methods are designed to utilize the sequence of content word on the source and target sides: 1) We encode the content words of the source sentence as a new source representation, and learn an additional content word context vector based on it to improve translation performance; 2) A specific loss for content words of the target sentence is introduced to compensate for the original training objection, to obtain a content word-aware NMT model. Empirical results on the WMT14 English-to-German, WMT14 English-to-French, and WMT17 Chinese-to-English tasks show the effectiveness of the proposed method.

* Corresponding author

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 358–364

2 Background: Transformer-based NMT

In Transformer-based NMT (Vaswani et al., 2017), the encoder is composed of a stack of L identical layers, each of which contains two sub-layers. The first sub-layer is a self-attention module (ATT), and the second sub-layer is a position-wise fully connected feed-forward network (FNN). A residual connection (He et al., 2016) is applied between the sub-layers, and layer normalization (LN) (Ba et al., 2016) is performed. Formally, the l-th identical layer of this stack is as follows:

$$\begin{aligned}\overline{\mathbf{H}}^{l} &= \mathrm{LN}(\mathrm{ATT}_e^l(\mathbf{Q}_e^{l-1}, \mathbf{K}_e^{l-1}, \mathbf{V}_e^{l-1}) + \mathbf{H}^{l-1}) \\ \mathbf{H}^{l} &= \mathrm{LN}(\mathrm{FFN}_e^l(\overline{\mathbf{H}}^{l}) + \overline{\mathbf{H}}^{l}).\end{aligned} \tag{1}$$

$\{\mathbf{Q}_e^{l-1}, \mathbf{K}_e^{l-1}, \mathbf{V}_e^{l-1}\}$ are query, key, and value vectors that are transformed from the (l-1)-th layer $\mathbf{H}^{l-1}$. For example, $\{\boldsymbol{Q}^0, \boldsymbol{K}^0, \boldsymbol{V}^0\}$ are packed from the $\mathbf{H}^0$ learned by the positional encoding mechanism (Gehring et al., 2017).

Similarly, the decoder is composed of a stack of L identical layers. Compared with the stacked encoder, it contains an additional attention sub-layer to compute alignment weights for the output of the encoder stack $\mathbf{H}^L$:

$$\begin{aligned}\overline{\mathbf{S}}_i^{l} &= \mathrm{LN}(\mathrm{ATT}_d^l(\mathbf{Q}_i^{l-1}, \mathbf{K}_i^{l-1}, \mathbf{V}_i^{l-1}) + \mathbf{S}_i^{l-1}), \\ \mathbf{C}_i^{l} &= \mathrm{LN}(\mathrm{ATT}_c^l(\overline{\mathbf{S}}_i^{l}, \mathbf{K}_e^{L}, \mathbf{V}_e^{L}) + \overline{\mathbf{S}}_i^{l}), \\ \mathbf{S}_i^{l} &= \mathrm{LN}(\mathrm{FFN}_d^l(\mathbf{C}_i^{l}) + \mathbf{C}_i^{l}),\end{aligned} \tag{2}$$

where $\mathbf{Q}_d^{l-1}, \mathbf{K}_d^{l-1}$, and $\mathbf{V}_d^{l-1}$ are query, key, and value vectors, respectively, that are transformed from the (l-1)-th layer $\mathbf{S}^{l-1}$ in time-step i. $\{\mathbf{K}_e^L, \mathbf{V}_e^L\}$ are transformed from the L-th layer of the encoder. The top layer of the decoder $\mathbf{S}_i^L$ is used to generate the next target word y_i by a linear, potentially multi-layered function:

$$P(y_i|y_{<i}, \mathbf{x}) \propto \exp(\boldsymbol{W}_o \tanh(\boldsymbol{W}_w \mathbf{S}_i^L), \tag{3}$$

where $\boldsymbol{W}_o$ and $\boldsymbol{W}_w$ are projection matrices. To obtain the translation model, the training objection maximizes the conditional translation probabilities over the training data set $\{[\mathbf{X}, \mathbf{Y}]\}$:

$$\mathcal{J}(\theta) = \arg\max_{\theta}\{P(\mathbf{Y}|\mathbf{X}; \theta)\}. \tag{4}$$

3 Content Word Recognition

We explore the effects of content words in a sentence for NMT. Specifically, we propose a content word recognition method based on the TF-IDF (Chen et al., 2019; Zhang et al., 2020). An input sentence of length J_m is treated as a document D_m, and the TF-IDF TI_j for each word d_j in D_m is computed:

$$TI_j = \frac{k_{j,m}}{J_m} \times \log \frac{|M|}{1 + |m : d_j \in D_m|}, \tag{5}$$

where $k_{j,m}$ represents the number of occurrences of the j-th word in the input sentence d_t; $|M|$ is the total number of sentences in the monolingual data; and $|m : d_j \in D_m|$ is the number of sentences including word d_j in the monolingual data. We then select a fixed percentage N (30% in the experiment) of word with high TF-IDF scores in the sentence as content words. Note that we focus on statistics related to word frequency here, instead of the linguistic criteria; this method of approximation eliminates the need for additional language-specific resources.

4 Content Word Aware NMT

In this section, we propose two ways to make use of the information on content words, designing three content word-aware NMT models. The proposed method of content word recognition is first added as an additional module to the encoder to learn the sequence of source content words $\mathcal{X}$ from the input source sentence. $\mathcal{X}$ is mapped and fed into the shared encoder (Li et al., 2020) in Eq.(1) to learn an additional source representation of content words $\boldsymbol{\mathcal{H}}^L$. An multi-head attention module is then introduced to the decoder to learn the context vector $\boldsymbol{\mathcal{C}}_i^l$ based content words at time-step i, and $\boldsymbol{\mathcal{C}}_i^l$ is used to enhance the output $\boldsymbol{\mathcal{S}}_i^l$:

$$\begin{aligned}\overline{\boldsymbol{\mathcal{S}}}_i^{l} &= \mathrm{LN}(\mathrm{ATT}_d^l(\mathbf{Q}_i^{l-1}, \mathbf{K}_i^{l-1}, \mathbf{V}_i^{l-1}) + \boldsymbol{\mathcal{S}}_i^{l-1}), \\ \mathbf{C}_i^{l} &= \mathrm{LN}(\mathrm{ATT}_c^l(\overline{\boldsymbol{\mathcal{S}}}_i^{l}, \mathbf{K}_e^{L}, \mathbf{V}_e^{L}) + \overline{\boldsymbol{\mathcal{S}}}_i^{l}), \\ \boldsymbol{\mathcal{C}}_i^{l} &= \mathrm{LN}(\mathrm{ATT}_y^l(\overline{\boldsymbol{\mathcal{S}}}_i^{l}, \boldsymbol{\mathcal{K}}_e^{L}, \boldsymbol{\mathcal{V}}_e^{L}) + \overline{\boldsymbol{\mathcal{S}}}_i^{l}), \\ \boldsymbol{\mathcal{S}}_i^{l} &= \mathrm{LN}(\mathrm{FFN}_d^l(\mathbf{C}_i^{l} + \boldsymbol{\mathcal{C}}_i^{l}) + \mathbf{C}_i^{l}),\end{aligned} \tag{6}$$

where $\boldsymbol{\mathcal{K}}_e^L$ and $\boldsymbol{\mathcal{V}}_e^L$ of the content words are transformed from the L-th layer of the encoder. Finally, the top layer of the decoder $\boldsymbol{\mathcal{S}}_i^L$, which is enhanced by the contextual vector of the content words $\boldsymbol{\mathcal{C}}_d^l$, is used as input to the Eq. (3) to compute the probabilities of the next target word y_i at time-step i:

$$P(y_i|y_{<i}, \mathbf{x}) \propto \exp(\boldsymbol{W}_o \tanh(\boldsymbol{W}_w \boldsymbol{\mathcal{S}}_i^L). \tag{7}$$

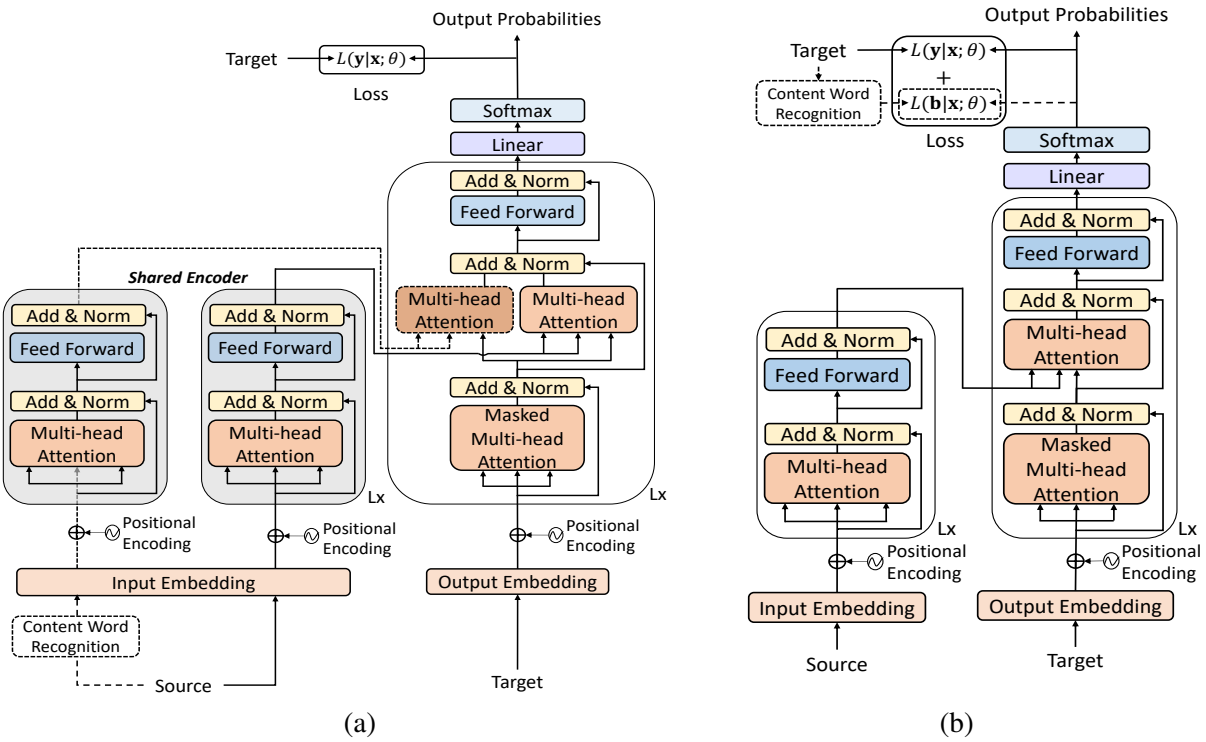


Figure 2: (a) Proposed SCWAContext model; (b) Proposed TCWALoss model.

Note that both the original source representation $\mathbf{H}^L$ and proposed content word based representation $\boldsymbol{\mathcal{H}}^L$ are learned by a shared encoder using our content word recognition module.

4.1 Target Content Word-Aware Loss

Like the source sentence, the target sentence also contains content words. We thus first identify a sequence of content words $\mathbf{b}$ from the target reference translation $\mathbf{y}$ according to the proposed content word recognition method (see Section 3). We then introduce an addition loss term as a measure of the content words, which encourages the translation model to attend to the translation of the content words. Formally, the training objective is revised as:

$$\mathcal{J}(\theta) = \arg\max_{\theta}\{P(\mathbf{y}|\mathbf{x};\theta) + \lambda * P(\mathbf{b}|\mathbf{x};\theta)\}, \quad (8)$$

where λ is a hyper-parameter empirically set to 0.4 in this paper. Note that the introduced content word-aware loss works without any new parameters and influences only the computation of loss during the training of the standard NMT model.

4.2 Proposed Translation Models

Based on the above two strategies, we design three NMT models: 1) **SCWAContext**: The source content words are used to learn an additional context vector to improve the prediction of target word (see Figure 2(a)); 2) **TCWALoss**: The target content words are used to compute an additional loss to guide the training of the translation model (see Figure 2(b)); 3) **BCWAContLoss**: It combines SCWAContext and TCWALoss to capture the content words of both the source and the target sentence to further improve translation performance.

5 Experiments

5.1 Setup

The proposed methods were evaluated on the WMT14 English-to-German (EN-DE), WMT14 English-to-French (EN-FR), and WMT17 Chinese-to-English (ZH-EN) tasks. The EN-DE corpus consists of 4M sentence pairs, the ZH-EN corpus of 22M sentence pairs, and the EN-FR corpus of 36M sentence pairs. We used the case-sensitive 4-gram BLEU score as evaluation metric. The results of the *newstest2014* test sets are reported for the EN-DE and EN-FR tasks, and the *newstest2017* test set is reported for the ZH-EN task. The byte pair encoding algorithm (Sennrich et al., 2016) was applied to encode all sentences to limit the size of the vocabulary to 40K. The other configurations were identical to those in (Vaswani et al., 2017). The poposed models were implemented by using

Systems	EN-DE			ZH-EN		EN-FR	
	BLEU	#Speed	#Param	BLEU	#Param	BLEU	#Param
Existing NMT systems							
Trans.base (Vaswani et al., 2017)	27.3	N/A	65.0M	N/A	N/A	38.1	N/A
+Context-Aware SANs (Yang et al., 2019a)	28.26	N/A	106.9M	24.67	126.8M	N/A	N/A
+Convolutional SANs (Yang et al., 2019b)	28.18	N/A	88.0M	24.80	N/A	N/A	N/A
+BIARN (Hao et al., 2019)	28.21	N/A	97.4M	24.70	107.3M	N/A	N/A
Trans.big (Vaswani et al., 2017)	28.4	N/A	213.0M	N/A	N/A	41.0	N/A
+Context-Aware SANs (Yang et al., 2019a)	28.89	N/A	339.6M	24.56	379.4M	N/A	N/A
+Convolutional SANs (Yang et al., 2019b)	28.74	N/A	339.6M	25.01	N/A	N/A	N/A
+BIARN (Hao et al., 2019)	28.98	N/A	333.5M	25.10	373.3M	N/A	N/A
Our NMT systems							
Trans.base	27.48	13.2K	66.5M	24.28	74.7M	38.32	66.9M
+SCWAContext	28.28+	12.1K	72.8M	24.79+	81.0M	39.41+	73.2M
+TCWALoss	27.94+	14.3K	66.5M	24.65	74.7M	38.89+	66.9M
+BCWAContLoss	28.51+	13.1K	72.8M	24.94+	81.0M	39.56+	73.2M
Trans.big	28.45	11.2K	221.1M	24.55	237.5M	41.21	222.9M
+BCWAContLoss	29.14+	10.1K	246.3M	25.12+	262.7M	42.57+	247.1M

Table 1: Results of the EN-DE, EN-FR, and ZH-EN tasks. "#Speed" and "#Param" denote the training speed (tokens/second) and the size of model parameters, respectively. "+" after a score indicates that the proposed method was significantly better than the Transformer at significance of $p < 0.01$ (Collins et al., 2005).

the fairseq toolkit (Ott et al., 2019).

5.2 Main Results

Table 1 shows results of the proposed method over our implemented Trans.base/big models which have similar BLEU scores with the original Transformer for the EN-DE and EN-FR tasks. We then make the following observations:

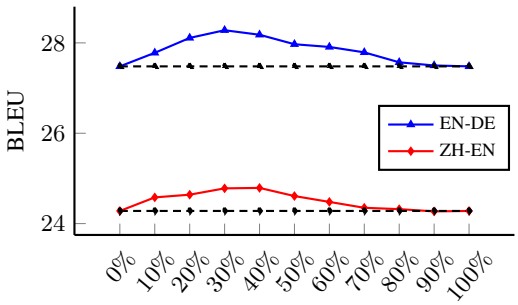
1) All proposed three word-aware NMT models outperformed the baseline Transformer model. This indicates that using information on the importance of words to enhance the translation of content words is helpful for the NMT model.

2) +SCWAContext performed better than +TCWALoss. The NMT model was more sensitive to information on source content words than target content words. +BCWAContLoss outperformed +SCWAContext and +TCWALoss, especially is superior to the existing +Context-Aware, +CSANs, and +BIARN. This suggests that the sequences of content words of both source and the target can be used together to further improve translation performance.

3) The parameters of the proposed models only slightly increased. In addition, Trans.base+BCWAContLoss delivered an comparable performance to Trans.big, which contained many more parameters than Trans.base+BCWAContLoss. This indicates that the improvement in performance did not occur owing to a greater number of parameters. The training speeds of our models were slightly lower than those of Trans.base.

5.3 Evaluating Content Word Recognition

Figure 3 shows the results of the Trans.base+SCWACont based different percentage N of content words in a sentence on the EN-DE and ZH-EN test sets. On both test sets, the highest BLEU scores were obtained with $N = 30\%$. With increasing values of N, the trend of their BLEU scores were similar on both test sets.



The percent of N in the content word recognition method

Figure 3: Results of Trans.base+SCWAContext model on the EN-DE and ZH-EN test set. The dashed line denotes the Trans.base model.

5.4 Evaluating Translation of Content Words

We apply the proposed content word recognition method to the generated translation and the reference translation of test set, and thus extract two short sequences of including 30% of content words.

We compute the accuracy of unigram content word between the extracted two short sequences, as shown in Table 2. The proposed methods outperformed the Trans.base in translating the content words, which is in line with the BLEU. This means that the proposed NMT model improved the generation of target content words.

System	EN-DE	ZH-EN
Trans.base	51.0%	53.8%
+SCWAContext	51.9%	54.6%
+TCWALoss	51.5%	54.2%
+BCWAContLoss	52.1%	54.7%

Table 2: Accuracy of unigram content words on the EN-DE and ZH-EN test sets with 30% of content words.

5.5 Effect of Content Word-Aware Loss

Figure 4 shows the results of +TCWALoss model on the EN-DE and ZH-EN test sets with different hyper-parameter λ. When λ increased from 0 to 0.4, the BLEU scores of +TCWALoss model improved by +0.8 points over Trans.base model. This means that the proposed content word-aware loss is useful for training NMT model. Subsequently, larger values of λ reduced the BLEU scores, suggesting that excessive biased content word translation may be weak at translating function words. Therefore, we set the hyper-parameter λ to 0.4 to control the loss of target content words in our experiments (Table 1).

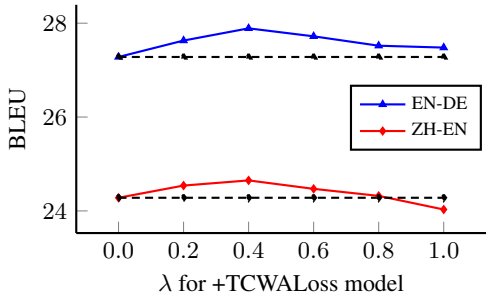


Figure 4: BLEU scores of the +TCWALoss model on the EN-DE and ZH-EN test sets with different values of λ. The dashed line denotes the result of the Trans.base model.

5.6 Content Word Recognition based on Function Word Frequency

Instead of directly identify content words, we identify the function words as the T most frequent words in the corpus. Furthermore, after we remove the function words in a sentence $\mathbf{x}=\{x_1, \cdots, x_J\}$, all the remaining words will be treated as a sequence (maintain the original order) of content words $\mathcal{X}$ according to the (Setiawan et al., 2007, 2009; Zhang and Zhao, 2013)'s work. Figure 5 shows the results of Trans.base+SCWAContLoss on the EN-DE and ZH-EN test sets with different number of the top T function words. Trans.base+SCWAContLoss obtained the highest BLEU scores on the both test sets over the Trans.base on modeling T = 256.

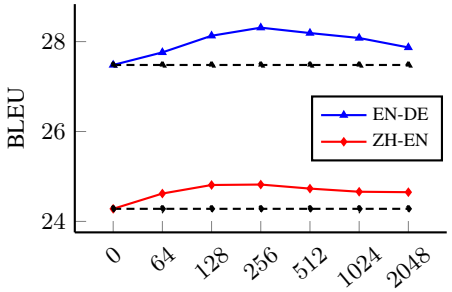


Figure 5: BLEU scores of Trans.base+SCWAContLoss on the EN-DE and ZH-EN test sets with different number of function words T.

6 Conclusion and Future Works

This paper explored the importance of word for NMT. We divided words of one sentence into content and function words through word frequency-related information. Our proposed NMT models, that are easy to implement and not much time and space cost, are introduced to the training and inference, and can improve the representation and translation of content words. In future work, we will investigate the impact of fine-grained word categories (such as nouns, verbs, and adjectives) on the translation performance and design specific methods according to these categories.

Acknowledgments

We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions. Masao Utiyama is partly supported by JSPS KAKENHI Grant Number 19H05660. Rui Wang was partially supported by JSPS grant-in-aid for early-career scientists (19K20354): "Unsupervised Neural Machine Translation in Universal Scenarios" and NICT tenure-track researcher startup fund "Toward Intelligent Machine Translation".

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations 2015*, San Diego, CA.

Kehai Chen, Rui Wang, Maosao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Neural machine translation with sentence-level topic context. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1970–1984.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Jie Hao, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. 2019. Modeling recurrence for transformer. 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 1198–1207, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. 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 953–962, Hong Kong, China. Association for Computational Linguistics.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020. Explicit sentence compression for neural machine translation. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

Hendra Setiawan, Min-Yen Kan, and Haizhou Li. 2007. Ordering phrases with function words. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 712–719, Prague, Czech Republic. Association for Computational Linguistics.

Hendra Setiawan, Min-Yen Kan, Haizhou Li, and Philip Resnik. 2009. Topological ordering of function words in hierarchical phrase-based translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 324–332, Suntec, Singapore. 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*, pages 3104–3112. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Baosong Yang, Jian Li, Derek F. Wong, Lidia S. Chao, Xing Wang, and Zhaopeng Tu. 2019a. Context-aware self-attention networks. In *AAAI Conference on Artificial Intelligence*, pages 387–394, Hawaii, USA.

Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019b. Convolutional self-attention networks. 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 4040–4045, Minneapolis, Minnesota. Association for Computational Linguistics.

Jingyi Zhang and Hai Zhao. 2013. Improving function word alignment with frequency and syntactic

information. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2211–2217. AAAI Press.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.