Sentence-Level Agreement for Neural Machine Translation

Mingming Yang[1]*, **Rui Wang**[2], **Kehai Chen**[2], **Masao Utiyama**[2],
Eiichiro Sumita[2], **Min Zhang**[1,3], and **Tiejun Zhao**[1]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[3]School of Computer Science and Technology, Soochow University, Suzhou, China
mmyang@hit-mtlab.net, minzhang@suda.edu.cn, tjzhao@hit.edu.cn
{wangrui, khchen, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

The training objective of neural machine translation (NMT) is to minimize the loss between the words in the translated sentences and those in the references. In NMT, there is a natural correspondence between the source sentence and the target sentence. However, this relationship has only been represented using the entire neural network and the training objective is computed in word-level. In this paper, we propose a sentence-level agreement module to directly minimize the difference between the representation of source and target sentence. The proposed agreement module can be integrated into NMT as an additional training objective function and can also be used to enhance the representation of the source sentences. Empirical results on the NIST Chinese-to-English and WMT English-to-German tasks show the proposed agreement module can significantly improve the NMT performance.

1 Introduction

Neural network based methods have been applied to several natural language processing tasks (Zhang et al., 2016; Li et al., 2018; Chen et al., 2018; Li et al., 2019; He et al., 2018). In neural machine translation (NMT), unlike conventional phrase-based statistical machine translation, an attention mechanism is adopted to help align output with input words (Bahdanau et al., 2015). It is based on the estimation of a probability distribution over all input words for each target word. However, source and target words are in different representation space, and they still have to go through a long information processing procedure that may lead to the source words are incorrectly translated into the target words. Based on this hypothesis, Kuang et al. (2018) proposed a direct bridging model, which directly connects source and target word embeddings seeking to minimize errors in the translation. Tu et al. (2017) incorporated a reconstructor module into NMT, which reconstructs the input source sentence from the hidden layer of the output target sentence to enhance source representation. However, in previous studies, the training objective function was usually based on word-level and lacked explicit sentence-level relationships (Zhang and Zhao, 2019). Although Transformer model (Vaswani et al., 2017) has archived state-of-the-art performance of NMT, more attention is paid to the words-level relationship via self-attention networks.

Sentence-level agreement method has been applied to many natural language processing tasks. Aliguliyev (2009) used sentence similarity measure technique for automatic text summarization. Liang et al. (2010) have shown that the sentence similarity algorithm based on VSM is beneficial to address the FAQ problem. Su et al. (2016) presented a sentence similarity method for spoken dialogue system to improve accuracy. Rei and Cummins (2016) proposed sentence similarity measures to improve the estimation of topical relevance. Wang et al. (2017b; 2018) used sentence similarity to select sentences with the similar domains. The above methods only considered monolingual sentence-level agreement.

In human translation, a translator's primary concern is to translate a sentence through its entire meaning rather than word-by-word meaning. Therefore, in early machine translation studies, such as example-based machine translation (Nagao, 1984; Nio et al., 2013), use the sentence similarity matching between the sentences to be translated and the sentences in the

*Mingming Yang was an internship research fellow at NICT when conducting this work.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3076–3082
Florence, Italy, July 28 - August 2, 2019.

bilingual corpora to extract translation. Inspired by these studies, we establish a sentence-level agreement channel directly in the deep neural network to shorten the distance between the source and target sentence-level embeddings. Specifically, our model can be effectively applied to NMT in two aspects:

- **Sentence-Level Agreement as Training Objective:** we use the sentence-level agreement as a part of the training objective function. In this way, we not only consider the translation of the word level but also consider the sentence level.
- **Enhance Source Representation:** As our model can make the vector distribution of the sentence-level between source-side and target-side closer, we can combine their sentence-level embeddings to enhance the source representation.

Experimental results on Chinese-to-English and English-to-German translation tasks demonstrate that our model is able to effectively improve the performance of NMT.

2 Neural Machine Translation

In this section, we take the Transformer architecture proposed by Vaswani et al. (2017), which is the state-of-the-art translation architecture, as the baseline system.

As an encoder-to-decoder architecture, $\boldsymbol{X} = \{x_1, x_2, ..., x_J\}$ represents a source sentence and $\boldsymbol{Y} = \{y_1, y_2, ..., y_I\}$ represents a target sentence. The encoder-to-decoder model learns to estimate the conditional probability from the source sentence to the target sentence word by word:

$$P(y|x;\theta) = \prod_{i=1}^{I} P(y_i|y_{<i}, x; \theta), \quad (1)$$

where θ is a set of model parameters and $y_{<i}$ denotes a partial translation.

Different from the other NMT, Transformer has the self-attention layers that can operate in parallel. A single self-attention layer has two sub-layers: a multi-head self-attention layer and a feed forward network. The feed forward network consists of two simple fully connected networks with a ReLU activation function in between:

$$\text{FFN}(x) = max(0, xW_1 + b_1)W_2 + b_2, \quad (2)$$

where W_1 and W_2 are both linear transformation networks, b_1 and b_2 are both bias. We define H_{enc} as the sentence representation of $\boldsymbol{X}$ via the self-attention layers in encoder, and H_{dec} as the sentence representation of words $\boldsymbol{Y}$ via embedding layers in decoder.

The parameters of Transformer are trained to minimize the following objective function on a set of training examples $\{(X^n, Y^n)\}_{n=1}^{N}$:

$$L_{mle} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I_y} \log P(y_i^n|y_{<i}^n, H_{enc}, H_{dec}). \quad (3)$$

3 Agreement on Source and Target Sentence

Some studies (Luong et al., 2015; Tu et al., 2016; Chen et al., 2017a,b; Kuang et al., 2018) showed that improving word alignment is beneficial to machine translation. Their idea is based on word-level agreement and make the embeddings of source words and corresponding target words similar. In this paper, we investigate the sentence-level relationship between the source and target sentences. We propose a sentence-level agreement method which can make the sentence-level semantics of the source and target closer. The entire architecture of the proposed method is illustrated in Figure 1.

3.1 Sentence-Level Agreement

First, we need to get the sentence-level representation of the source and target. Some studies showed that the **Mean** operation is an effective method to represent sentence of sequence words (Mitchell and Lapata, 2010; Mikolov et al., 2013; Le and Mikolov, 2014), especially for NMT (Wang et al., 2017a). Motivated by this, we adopt **Mean** to represent the source and target sentences as shown in Figure 1(a).

Denote $\widetilde{H}_{enc}$ is the mean of H_{enc} and $\widetilde{H}_{dec}$ is the mean of H_{dec}. We design a **Sentence Agreement Loss** L_{mse} to measure the distance between the source and target sentence-level vectors:

$$L_{mse} = ||\widetilde{H}_{enc} - \widetilde{H}_{dec}||^2. \quad (4)$$

Finally, our goal is to improve translation with shortening the distance in sentence-level. Thus,

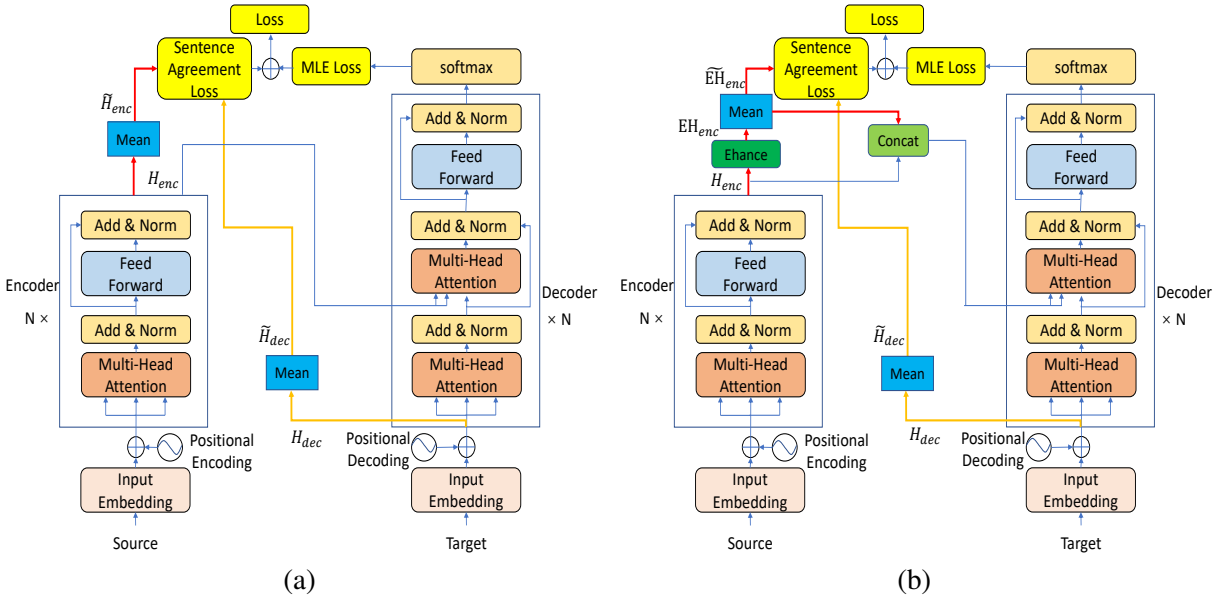

Figure 1: (a) Architecture of Sentence-Level Agreement Loss; (b) Architecture of Enhance Source Representation.

the final objective of our model is composed of two parts, the formula is as follows:

$$L = L_{mle} + L_{mse}. \tag{5}$$

3.2 Enhance Source Representation

Sentence-level agreement helps make the target-side sentence representation closer to the source. Intuitively, we can also use this mechanism to strengthen the source representation to improve the translation. Further, we propose a simple and efficient architecture in Figure 1(b).

First, we map H_{enc} to the target-side vector EH_{enc} through a simple feed forward network TFFN by eq.(2):

$$EH_{enc} = \text{TFFN}(H_{enc}). \tag{6}$$

In particular, we use a Tanh activation function instead of ReLU in the feed forward network. The value range of Tanh is -1 to 1, which indicates some information should be counterproductive. Our **Enhanced Sentence Agreement Loss** LE_{mse} is to measure the distance between the source and target sentence-level vectors:

$$LE_{mse} = ||\widetilde{EH}_{enc} - \widetilde{H}_{dec}||^2, \tag{7}$$

where $\widetilde{EH}_{enc}$ is the mean of EH_{enc}. Le and Mikolov (2014) use concatenation as the method to combine the sentence vectors to strengthen the capacity of representation. We also use the same method to combine H_{enc} and $\widetilde{EH}_{enc}$:

$$CH_{enc} = \text{Concat}(H_{enc}, \widetilde{EH}_{enc}). \tag{8}$$

In this way, we can enhance the source representation with a sentence-level representation closer to the target-side. The updated translation training objective is:

$$LE_{mle} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I_y}\log P(y_i^n | y_{<i}^n, CH_{enc}, H_{dec}). \tag{9}$$

Thus, the final objective is as follows:

$$LE = LE_{mle} + LE_{mse}. \tag{10}$$

4 Experiments

4.1 Dataset

For Chinese-English (ZH-EN) translation, our training data for the translation task consists of 1.25M Chinese-English sentence pairs extracted from LDC corpora[1]. The NIST02 testset is chosen as the development set, and the NIST03,

[1]The corpora include LDC2002E18, LDC2003E07,LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

#	Model	NIST					WMT
		03	04	05	06	Avg	14
	Existing NMT Systems						
1	EDR (Tu et al., 2017)	N/A	N/A	33.73	34.15	N/A	N/A
2	DB (Kuang et al., 2018)	38.02	40.83	N/A	N/A	N/A	N/A
	Our NMT Systems						
3	Transformer(Base)	45.57	46.40	46.11	44.92	45.75	27.28
4	+loss$_{mse}$	46.71†	47.23†	47.12†	45.78†	46.71	28.11†
5	+loss$_{mse}$ + $enhanced$	46.94†	47.52†	47.43†	46.04†	46.98	28.38†
6	Transformer(Big)	46.73	47.36	47.15	46.82	47.01	28.36
7	+loss$_{mse}$	47.43†	47.96	47.78	47.39	47.74	28.71
8	+loss$_{mse}$ + $enhanced$	47.68†	48.13†	47.96†	47.56†	47.83	28.92†

Table 1: Translation results for Chinese-English and English-German translation task. "†": indicates statistically better than Transformer(Base/Big) ($\rho < 0.01$).

#	Model	BLEU WMT14	Param	Speed (tokens/s)	
				Train	Decode
1	Transformer(Base)	27.28	93.3M	9,950	150
2	+loss$_{mse}$	28.11	93.3M	9,850	150
3	+loss$_{mse}$ + $enhanced$	28.38	93.9M	9,780	146
4	Transformer(Big)	28.36	274.7M	4,340	95
5	+loss$_{mse}$	28.71	274.7M	4,300	95
6	+loss$_{mse}$ + $enhanced$	28.92	276.8M	4,150	88

Table 2: The efficiency analysis on English-German task. "Param" denotes the trainable parameter size of each model (M=million) and Beam=10.

NIST04, NIST05, NIST06 datasets are testsets. We use the case-insensitive *4*-gram NIST BLEU score as our evaluation metric (Papineni et al., 2002). The training data of English-German (EN-DE) translation is from WMT14, which consists of 4.5M sentence pairs. We use byte-pair encoding (Sennrich et al., 2016b) to segment words. The newstest2013 was used as a development set and the newstest2014 as test sets that are evaluated by SacreBLEU (Post, 2018).

To efficiently train NMT models, we train each model with sentences of length up to 50 words. In this way, about 90% and 89% of ZH-EN and EN-DE parallel sentences are covered in the experiments. In addition, we use byte pair encoding (Sennrich et al., 2016a) with 32K merges to segment words into sub-word units for all languages to alleviate the out-of-vocabulary problem. We evaluate the proposed approaches on our re-implemented Transformer model (Vaswani et al., 2017). We test both the Base and Big models, which differ at the dimensionality of input and output (512 vs 1024), the number of attention head (8 vs 16) and the inner-layer size (2048 vs 4096). We set 6 layers for encoder and decoder. All the models were trained on a single NVIDIA P100 GPU, which is allocated a minibatch of 4096 tokens. About 200K minibathes are trained.

4.2 Performance

Table 1 shows the performances measured in terms of BLEU score. On ZH-EN task, Transformer(Base) outperforms the existing systems EDR (Tu et al., 2017) and DB (Kuang et al., 2018) by 11.5 and 6.5 BLEU points. With respect to BLEU scores, all the proposed models (Row 4-5) consistently outperform Transformer(base) by 0.96 and 1.23 BLEU points. The big models (Row 7-8) also achieve similar improvement by 0.73 and 0.82 BLEU points on a larger parameters model. These findings suggest a sentence-level agreement between source-side and target-side is helpful for NMT. Further, we use it to enhance the source representation is an effective way to improve the translation. In

Model	NIST		WMT	
	BLEU	sim(‰)	BLEU	sim(‰)
Transformer(Base)	45.75	13.7	27.28	36.2
+loss$_{mse}$	46.71	19.8	28.11	48.3
+loss$_{mse}$ + $enhanced$	46.98	26.9	28.38	57.6
Transformer(Big)	47.01	13.5	28.36	41.5
+loss$_{mse}$	47.74	18.3	28.71	56.4
+loss$_{mse}$ + $enhanced$	47.83	23.2	28.92	68.2

Table 3: Source-to-target sentence-level similarity analysis on Chinese-English and English-German translation task.

addition, the proposed methods gain similar improvements on EN-DE task.

4.3 Efficiency Analysis

In Table 2, we analyze the efficiency of the proposed methods. $loss_{mse}$ (Row 2 and 5) gets better translation without any added parameters, only decrease approximately 1% train speed. It shows that sentence-level agreement is helpful for translation. Compared with Row 1 and 4, $loss_{mse} + enhanced$ (Row 3 and 6) increases little parameters about 0.6M and 2.1M, train and decode speed drop very little. However, it has greatly improved the translation performance. In particular, by comparing Row 3 and 4, we find that our proposed methods achieve a similar performance with the Transformer(Big) and gain a faster speed with fewer parameters. It indicates that enhancing source representation with a sentence-level representation is an effective method for improving translation performance.

4.4 Sentence-Level Similarity Analysis

We further study how the proposed models influenced sentence-level similarity in translation. For this, we follow the method of Lapata and Barzilay (2005) to measure sentence similarity. First, each sentence is represented by the mean of the distributed vectors of its words. Second, the similarity between source and target sentences is determined by the cosine of their means:

$$sim = cos(\widetilde{H}_{enc}, \widetilde{H}_{dec}). \quad (11)$$

As Table 3 shows, the sentence-level similarity of the proposed method is higher than the corresponding baselines. In addition, there is a correlation between NMT performance (BLEU) and the sentence-level similarity. This indicates that the proposed method can improve the sentence-level similarity between source and target sentences and the performance of NMT.

5 Conclusion

In this work, we have presented a sentence-level agreement method for NMT. Our goal is to bring the sentence representation of the source-side and the target-side closer together. At the same time, we can utilize this information to enhance source representation. Our study suggests the source-to-target sentence-level relationship is very useful for translation. In future work, we intend to apply these methods to other natural language tasks.

Acknowledgments

The corresponding authors are Rui Wang and Min Zhang. This work was partially conducted under the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of the Ministry of Internal Affairs and Communications (MIC), Japan. 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". Min Zhang is partially supported by the National Natural Science Foundation of China via No. 61525205.

References

Ramiz M. Aliguliyev. 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Syst. Appl.*, 36(4):7764–7772.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly

learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA.

Kehai Chen, Rui Wang, Masao Utiyama, Lemao Liu, Akihiro Tamura, Eiichiro Sumita, and Tiejun Zhao. 2017a. Neural machine translation with source dependency representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2852, Copenhagen, Denmark. Association for Computational Linguistics.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2017b. Context-aware smoothing for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–20, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2018. Syntax-directed attention for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4792–4799, New Olreans, LA.

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.

Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong. 2018. Attention focusing for neural machine translation by bridging source and target embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1776. Association for Computational Linguistics.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: models and representations. In *International Joint Conference on Artificial Intelligence*, pages 1085–1090.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214, Santa Fe, New Mexico, USA.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform semantic role labeling. *CoRR*, abs/1901.05280.

Xu Liang, Dongjiao Wang, and Ming Huang. 2010. Improved sentence similarity algorithm based on vsm and its application in question answering system. In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, volume 1, pages 368–371.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34:1388–1429.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. Of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, New York, NY, USA. Elsevier North-Holland, Inc.

Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2013. Combination of example-based and smt-based approaches in a chat-oriented dialog system. *Proc. of ICE-ID*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Marek Rei and Ronan Cummins. 2016. Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 283–288, San Diego, CA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Bo-Hao Su, Ta-Wen Kuan, Shih-Pang Tseng, Jhing-Fa Wang, and Po-Huai Su. 2016. Improved tf-idf weight method based on sentence similarity for spoken dialogue system. *2016 International Conference on Orange Technologies (ICOT)*, pages 36–39.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 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)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

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, *Proceedings of Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada.

Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2018. Sentence selection and weighting for neural machine translation domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1727–1741.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Huan Zhang and Hai Zhao. 2019. Minimum divergence vs. maximum margin: An empirical comparison on seq2seq models. In *Proceedings of the Seventh International Conference on Learning Representations*, New Orleans, USA.

Zhisong Zhang, Hai Zhao, and Lianhui Qin. 2016. Probabilistic graph-based dependency parsing with convolutional neural network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1382–1392, Berlin, Germany.