# **Baidu Neural Machine Translation Systems for WMT19**

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, Haifeng Wang

Baidu Inc. No. 10, Shangdi 10th Street, Beijing, 100085, China {sunmeng09, jiangbojian, xionghao05, hezhongjun, wu\_hua, wanghaifeng}@baidu.com

#### Abstract

In this paper we introduce the systems Baidu submitted for the WMT19 shared task on Chinese⇔English news translation. Our systems are based on the Transformer architecture with some effective improvements. Data selection, back translation, data augmentation, knowledge distillation, domain adaptation, model ensemble and re-ranking are employed and proven effective in our experiments. Our Chinese→English system achieved the highest case-sensitive BLEU score among all constrained submissions, and our English→Chinese system ranked the second in all submissions.

### 1 Introduction

The Transformer model (Vaswani et al., 2017), which exploits self-attention mechanism both in the encoder and decoder, has significantly improved the translation quality in recent years. It is also adopted by most participants as the basic Neural Machine Translation (NMT) system in the previous translation campaigns (Bojar et al., 2018; Niehues et al., 2018). In this year's translation task, we focus on the improvement of single system, and propose three novel Transformer variants:

- Pre-trained Transformer: We train a big Transformer language model (Radford et al., 2018; Devlin et al., 2018; Dai et al., 2019; Sun et al., 2019) on monolingual corpora, and use the language model as the encoder of the Transformer model.
- Deeper Transformer: We increase the encoder layers to better learn the representation of the source sentences. Specifically, we increase the number of encoder layers from 6 to 30 for the base version, and from 6 to 15 layers for the big version.

• Bigger Transformer: According to the previous experiments, the performance of the Transformer model is largely dependent on the dimensions of feed forward network. To further improve the performance, we increase the inner dimension of feed-forward network from 4,096 to 15,000 for big version.

In addition, we develop effective approaches to exploit additional monolingual data and generate augmented bilingual data. To use the monolingual data, back translation (Sennrich et al., 2015a) is employed on large corpora including News Corpus and Gigaword. We also use an iterative approach (Zhang et al., 2018) to extend the back translation method by jointly training source-totarget and target-to-source NMT models. For bilingual data augmentation, a target-to-source baseline system is used to translate the target of the bilingual corpus as the synthetic data. Moreover, the sequence-level knowledge distillation (Hassan et al., 2018) mechanism is employed to boost the performance by means of using the model decoding from right to left (Right-to-Left) and the aforementioned Transformer variants to generate synthetic data for training the NMT model (Wang et al., 2018).

The remainder of paper is structured as follows: Section 2 describes the detailed overview of our training strategy. Section 3 shows the experimental settings and results. Finally, we conclude our work in Section 4.

### 2 System Overview

Figure 1 depicts the overall process of our submissions in this year's evaluation task, in which we train our advanced Transformer models on the bilingual corpus together with synthetic corpora, fine-tune them on the well-selected in-domain data, and generate the ensemble model for the final

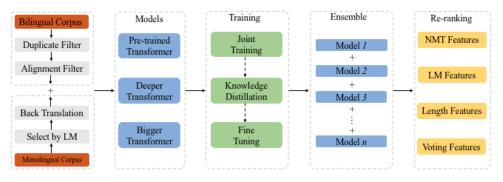


Figure 1: Architecture of Baidu NMT system

re-ranking strategy. In this section, we will introduce each step in details.

It is worth noting that our advanced Transformer model requires larger GPU memory to train due to the large number of training parameters. Hence we train our models on machines with 8 NVIDIA V100 GPUs each of which has 32 GB memory, to avoid out-of-memory issues. In training phase, we limit the number of source and target tokens per batch to 4,096 per GPU for deeper and bigger Transformer models (at most 526,052,128 parameters), while the token batch size is 3,072 for pre-trained Transformer model due to GPU memory limitation.

# 2.1 Pre-trained Transformer

Recent empirical improvements with language models have showed that unsupervised pretraining (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018; Dai et al., 2019; Sun et al., 2019) on very large corpora is an integral part of many NLP tasks. We implement a big Transformer language model using PaddlePaddle<sup>1</sup>, an end-to-end open source deep learning platform developed by Baidu. It provides a complete suite of deep learning libraries, tools and service platforms to make the research and development of deep learning simple and reliable. The language model is pre-trained only with masked language model task (Taylor, 1953; Devlin et al., 2018; Sun et al., 2019) on a monolingual corpus of the source language.

We use all the available resources of WMT19 as the pre-training corpus. For the Chinese language model, we use the concatenation of Chinese Gigaword, Chinese News Crawl, XMU and the Chinese part of CWMT and UN corpus. For the English language model, we use the concatenation of English Gigaword, English News Crawl and the English part of CWMT and UN corpus. There are 45 million Chinese sentences and 170 million English sentences in our pre-training corpora.

To use the pre-trained language model as encoder of NMT and enable the open-vocabulary translation, we learn a BPE (Sennrich et al., 2015b) model with 30K merge operations. We use Adam with learning rate of 1e-4,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , L2 weight decay of 0.001, and learning rate warmup over the 10,000 steps. We train the big Transformer language model with 24 layers, setting the hidden size to 1,024 and the number of self-attention heads to 16. Both Chinese and English pre-training took 7 days to complete.

In the fine-tuning procedure of the translation task, we employ a pre-trained language model as encoder of NMT, and the parameters of decoders are learned during fine-tuning. The decoder has 6 self-attention layers, and the hidden size is 1024, which is same with the decoder of standard big Transformer. During fine-tuning, we only fix the parameters of the language model for the first 10,000 steps.

### 2.2 Deeper Transformer

According to the previous literatures, the model tends to specialize in word sense disambiguation and tends to focus on local dependencies in lower layers but finds long dependencies on higher ones while increasing the size of layers in the encoder (Tang et al., 2018; Domhan, 2018; Raganato and Tiedemann, 2018). Meanwhile, inspired by the success of pre-trained Transformer, that translation results can benefit from very deep architectures of encoder, we introduce the deeper Transformer. But vanishing-gradient problem is encoder depth,

https://github.com/paddlepaddle/
paddle

the standard Transformer failed to train. To alleviate the vanishing-gradient problem, we design a particular residual connections. Specificially, the outputs of all preceding layers are used as inputs for each layer, as opposed to the standard Transformer model in which the residual connection is employed between two adjacent layers.

In our experiments, both the big Transformer with 15 encoder layers and the base transformer with 30 encoder layers obtain significant improvements compared with the standard big Transformer on Chinese $\rightarrow$ English translation task, whereas the improvement is not remarkable on English $\rightarrow$ Chinese translation task.

### 2.3 Bigger Transformer

Motivated by the success of increasing the model size on the language modeling (Devlin et al., 2018) and NMT (Vaswani et al., 2017) tasks, we propose bigger Transformer which has larger inner dimension of feed-forward network than the standard big Transformer. Specifically, we increase the inner dimension of feed-forward network from 4,096 to 15,000 constrained by the GPU memory capacity. To overcome the overfitting problem, we set attention dropout and relu dropout from 0.1 to 0.3, increasing the value of label smoothing from 0.1 to 0.2. Note that the specific settings are only employed for the bigger Transformer.

In addition, we explore the effectiveness of increasing hidden size with respect to the Transformer model. However, the results indicate that the model with increased hidden size performs worse than the model with big feed-forward network. Nevertheless, we retain the model with different hidden size as one diverse system for the generation of the final ensemble model, which has shown effective performance in our further experiments.

### 2.4 Large-scale Back-Translation

In recent work, Edunov et al. (2018) proposed an effective approach to improve the translation quality by exploiting back-translation mechanism on the large-scale monolingual corpus. Following their work, we also train our model on the synthetic bilingual corpus to further improve the performance. However, the provided monolingual data contains a certain amount of noise and out-ofdomain data which may affect the translation quality implicitly. Therefore, we use a language model to select high-quality and in-domain data from the large amount of monolingual data according to the perplexity score.

After training language models on different types of monolingual data (i.e., News crawl, Gigaword), we select 96M English sentences and 23M Chinese sentences according to LM scores, since Chinese monolingual corpus provided by WMT 19 is much less than that of English. The selected English sentences are translated and divided into 12 portions. For the 23M Chinese sentences, we translate and divide the sentences into 3 portions, resulting in 8M synthetic parallel sentence pairs in each portion. We further evaluate the performance of the similar model training on a different bilingual corpus which consists of the original bilingual corpus and the generated synthetic bitext. According to the BLEU score of translation results on the WMT 18 news translation dev set, we select the top 4 most effective portions for training Chinese $\rightarrow$ English system and the top 2 portions for training English→Chinese system. In the final submission, the selected synthetic portions are used to enhance individual baseline models by the following joint training technique, respectively.

### 2.5 Joint Training and Data Augmentation

In the work of Zhang et al. (2018), they proposed a novel method for better usage of monolingual data from both source side and target side by jointly optimizing a source-to-target (S2T) model and a target-to-source (T2S) model, training with several iterations. In each iteration, the T2S model is responsible for generating synthetic parallel training data for S2T model using target-side monolingual data, while S2T model is employed to generate synthetic parallel training data for T2S model using source-side monolingual data. After training on the additional synthetic data, the performance of both T2S model and S2T model can be further improved. In the next iteration, the two improved models can potentially generate better synthetic parallel data. This procedure can be applied in several iterations until no further improvement can be obtained.

In addition, we also augment the training data by exploring the bilingual corpus rather than the monolingual corpus. Specifically, we translate the sentences in the target language back into the source language by diverse training models, such as Left-to-right model and Right-to-left model. This procedure can be viewed as one alternative solution for alleviating the exposure bias problem (Ranzato et al., 2016).

# 2.6 Knowledge Distillation

The early adoption of knowledge distillation (Kim and Rush, 2016) is for model compression, where the goal is to deliver a compact student model that matches the accuracy of a large teacher model or the ensemble of models. In our knowledge distillation approach, we translate the source side of the bilingual data with a Right-to-Left (R2L) (Liu et al., 2016) model teacher and different architecture NMT teachers to use the translations as additional training data for the student network. Considering that distillation from a bad teacher model is likely to hurt the student model and thus result in inferior accuracy, we selectively use distillation in the training process. In particular, the sentences generated by a teacher model are filtered if BLEU scores are below a threshold  $\tau$ . According to our previous empirical results, we select English translations with BLEU score higher than 30 and Chinese translations with BLEU score higher than 42.

There are two kinds of teacher models to help a student model improve translation performance:

- R2L Teacher: The idea is to reverse the target sentences of bilingual corpus and train a R2L model. Then we employ R2L model to translate the source sentences of the bilingual corpus and reverse the translated sentences. The pseudo corpus is added to the real bilingual corpus in order to enhance the L2R model. The paradigm can be regarded as a kind of knowledge transfer method which provides complementary information for student model to learn.
- Hybrid Heterogeneous Teacher: Pre-trained Transformer, deeper Transformer and bigger Transformer represent a source sentence at different granularities, therefore it is intuitive that each model can learn effective knowledge from other models. For each individual model, we use the other two models as the teacher model to further improve the performance.

# 2.7 Fine-tuning with In-domain Data

Domain adaptation plays an important role in improving the performance towards given testing

Source	Chn→En	En→Chn
CWMT	6.7M	6.7M
UN	9M	3.5M
Wiki Titles	-	0.6M
Total	15.7M	10.8M

Table 1: Statistics of the bilingual training data (Chn indicates Chinese while En indicates English).

data. The dominant approach for domain adaptation is training on large-scale out-of-domain data and then fine-tuning on the in-domain data (Luong and Manning, 2015). Thus the effectiveness of the domain adaptation depends on the selected in-domain data.

According to our previous empirical results, using the WMT 18 dev set to fine-tune the models straightforwardly achieves the best results. In our final submission, we set the batch size to 1,024 and fine-tune the model for a few iterations on the WMT 18 dev set. It is surprising to find a gain of almost +2 BLEU improvement on WMT 18 Chinese $\rightarrow$ English test set. However, on WMT 18 English $\rightarrow$ Chinese test set, the improvement is not significant.

In WMT 17 and 18, the source side of both dev set and test set are composed of two parts: documents created originally in Chinese and documents created originally in English. We split both the dev set and test set into original Chinese part and original English part according to tag attributes of SGM files. Finally, we translate each specific test part with the model finetuned on the corresponding dev set. Experiments show significant improvement with this method, that is, 2.23 BLEU improvements on Chinese→English test set and 0.5 BLEU improvements on English->Chinese test set. This indicates that the translation quality is affected by the original sources of the language. Consider the English -> Chinese task, if the English sentences are created from native English corpus, then the corresponding Chinese sentences are translation style, so the model fine-tuned on these parallel sentences is more inclined to decode with translation style. Similarly, if the Chinese sentences are created from native Chinese corpus, the fine-tuned English-Chinese model decodes with more native style.

In the final submission, we take the following steps to avoid overfitting: 1) We employ the en-

Settings	Big	Pre-trained	Deeper	Bigger	
	Transformer	Transformer	Transformer	Transformer	
Baseline	25.86	-	-	-	
+ Back Translation	26.72	27.68	26.83	27.54	
+ Joint Training	26.95	27.79	27.01	27.61	
+ Knowledge Distillation	27.4	28.11	27.43	27.88	
+ Fine-tuning	29.39	29.87	29.82	30.11	
+ Ensemble	31.59				
+ Re-ranking*	31.83				

Table 2: BLEU evaluation results on the WMT 2018 Chinese→English test set (\* denotes the submitted system).

semble models to translate the WMT 19 test set, and use the translations as additional synthetic fine-tuning corpus. 2) We fine-tune the final system on the mixture of the additional synthetic corpus and the selected in-domain corpus.

# 2.8 Model Ensemble

Model ensemble is a widely used technique to boost the performance by combining the predictions of several models at each decoding step. In our previous experiments, we find that the improvement is slight while integrating the predictions of multiple models with similar model architecture. Instead, we train our models with different model architectures training on different versions of training data, increasing the model diversity for the model ensemble. The experimental results indicate that this method achieves absolute improvements over the single system (at most a 1.7 BLEU point improvements).

### 2.9 Re-ranking

In order to get better translation results, we generate n-best hypotheses with an ensemble model and then train a re-ranker using k-best MIRA (Cherry and Foster, 2012) on the validation set. K-best MIRA is a version of MIRA (Chiang et al., 2008) that works with a batch tuning to learn a re-ranker for the n-best hypotheses. The features we use for re-ranking are:

- NMT Features: Ensemble model score and Right-to-Left model score.
- Language Model Features: Multiple n-gram language models and backward n-gram language models.
- Length Features: Length ratio and length difference between source sentences and hypotheses.

• Weighted Voting Features: Average of BLEU scores calculated between each hypothesis and the other hypotheses.

# **3** Experiments and Results

All of our experiments are carried out on 32 machines with 8 NVIDIA V100 GPUs each of which have 32 GB memory. For all models, we average the last 20 checkpoints to avoid overfitting. We use cased BLEU scores calculated with Moses<sup>2</sup> mteval-v12a.pl script as evaluation metric. Following the organizers' suggestion, News dev 2018 is used as the development set and News test 2018 as the test set.

### 3.1 Pre-processing and Post-processing

The Chinese data has been tokenized using the Jieba tokenizer<sup>3</sup>. For English data, punctuation normalization, aggressive tokenization and truecasing are applied orderly to all sentences with the scripts provided in Moses. We also filter the parallel sentences which are duplicated or bad alignment scores obtained by fast-align (Dyer et al., 2013), and then we have a preprocessed bilingual training data consisting of 18M parallel sentences.

In post-processing phase, the English translations are true-cased and de-tokenized with the scripts provided in Moses. We use simple rules to normalize the punctuations and Arabic numerals in the Chinese translations.

#### **3.2** Chinese→English

For Chinese $\rightarrow$ English task, we do not use all of the 18M preprocessed parallel sentences, in that there is much out-of-domain data in UN corpus. Table 1 shows that the 6.7M CWMT corpus and 9M UN corpus which are selected ran-

<sup>&</sup>lt;sup>2</sup>http://www.statmt.org/moses/ <sup>3</sup>https://github.com/fxsjy/jieba

Settings	Big	Pre-trained	Deeper	Bigger	
	Transformer	Transformer	Transformer	Transformer	
Baseline	39.2	-	-	-	
+ Back Translation	43.33	43.7	42.19	44	
+ Joint Training	43.86	44.12	42.5	44.78	
+ Knowledge Distillation	44.25	44.6	42.86	45.27	
+ Fine-tuning	44.72	44.75	42.94	45.79	
+ Ensemble	46.42				
+ Re-ranking*	46.51				

Table 3: BLEU evaluation results on the WMT 2018 English→Chinese test set (\* denotes the submitted system).

domly are used as our bilingual training set for Chinese $\rightarrow$ English task. We learn a BPE (Sennrich et al., 2015b) model with 30K merge operations, in which 46.4K and 31K sub-word tokens are adopted as Chinese and English vocabularies separately. We set beam size to 12 and alpha to 1.1 during decoding.

12 portions of sentences are selected from huge volumes of English monolingual data, and we carry out a large number of experiments in which the Transformer models are trained with each portion. And then 4 most effective portions are selected. Due to the extensive training time and the approaching deadline for submissions, pre-trained transformer, deep Transformer(base Transformer with 30 encoder layers) and bigger Transformer are trained on the combination of real bilingual data and the synthetic data directly. For each different architecture model, we train 4 more systems with different portions of monolingual data and different parameters in order to obtain more diverse models. For comparison, we only report results on the WMT 2018 test set with the same portion of monolingual data.

Table 2 shows that the translation quality is largely improved using proposed techniques. We observe solid improvement of 0.86 BLEU for the baseline system after back translation. Joint training and knowledge distillation yield improvements over all the different architecture models, approximating 0.34-0.68 BLEU improvements toward single models. It is also clear that the fine-tuning technique brings substantial improvements compared with the baseline systems.

In our experiments, the ensemble models consists of 8 single models: 1 Transformer, 2 pretrained Transformers, 2 deeper Transformers and 3 bigger Transformers. As shown in the Table 2, the ensemble models also outperform the best single model by 1.49 BLEU score. However, the improvement of re-ranking is relatively slight, and we attribute this to the strong performance of ensemble models. Our WMT 2019 Chinese $\rightarrow$ English submission achieves a cased BLEU score of 38.0, winning the first place among all submissions.

### **3.3** English → Chinese

As listed in the Table 1, the parallel training data for English $\rightarrow$ Chinese translation task consists of about 6.7M sentence pairs from the filtered CWMT Corpus, 3.5M sentence pairs from the UN Parallel Corpus, 0.6M sentence pairs from the Wiki Titles Corpus. For the UN data, we train a 5-gram KN language model on the Chinese sides of the CWMT data and select 3.5M sentence pairs according to their perplexities. The size of the English vocabulary and the Chinese vocabulary are 31K and 48.6K respectively after BPE operation. We use beam search with a beam size of 12, and set alpha 0.8.

From the Table 3, we can observe: 1) We obtain +4.13 BLEU score when adding the synthetic parallel data to the training set of the Transformer. 2) We further gain +0.92 BLEU score after applying joint training and knowledge distillation for the Transformer system. 3) The improvement from the fine-tuning technique is relative slight for the pre-trained Transformer and deeper Transformer, whereas it is effective for the Transformer and bigger Transformer, with about 0.5 BLEU score improvements.

Notably, the ensemble models consist of pretrained Transformers and bigger Transformers. We omit the deeper Transformer model due to its worse performance on this translation task. On the WMT 2019 English $\rightarrow$ Chinese task, our submission achieves 42.4 cased BLEU score, winning the second place in the translation task.

# 4 Conclusion

This paper presents the Baidu NMT systems for WMT 2019 Chinese↔English news translation tasks. We investigate various different architectures of Transformer to build numerous strong single systems. We exploit effective strategies to better utilize parallel data as well as monolingual data. We find significant gains from combining multiple heterogeneous systems due to the diversity. Finally, our submission of Chinese→English news task achieves the highest cased BLEU score and our submission of English→Chinese achieves the second best cased BLEU score among all the constrained submissions.

# **5** Acknowledgements

We thank Shikun Feng at Baidu for providing the pre-trained language model. We thank the anonymous reviews for their careful reading and their thoughtful comments.

#### References

- Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.
- Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics.
- David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the conference on empirical methods in natural language processing*, pages 224–233. Association for Computational Linguistics.
- Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

- Tobias Domhan. 2018. How much attention do you need? a granular analysis of neural machine translation architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1799–1808.
- Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.
- Yoon Kim and Alexander M Rush. 2016. Sequencelevel knowledge distillation. *arXiv preprint arXiv:1606.07947*.
- Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on targetbidirectional neural machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 411–416.
- Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.
- Jan Niehues, Ronaldo Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT) 2018.*
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. URL https://s3us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.

- Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformerbased machine translation. In *Proceedings of the* 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 287–297.
- Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.
- Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*.
- Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Mingxuan Wang, Li Gong, Wenhuan Zhu, Jun Xie, and Chao Bian. 2018. Tencent neural machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 522–527, Belgium, Brussels. Association for Computational Linguistics.
- Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.