MiSS@WMT21: Contrastive Learning-reinforced Domain Adaptation in Neural Machine Translation

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}@nict.go.jp, zhaohai@cs.sjtu.edu.cn

Abstract

In this paper, we describe our MiSS system that participated in the WMT21 news translation task. We mainly participated in the evaluation of the three translation directions of English-Chinese and Japanese-English translation tasks. In the systems submitted, we primarily considered wider networks, deeper networks, relative positional encoding, and dynamic convolutional networks in terms of model structure, while in terms of training, we investigated contrastive learning-reinforced domain adaptation, self-supervised training, and optimization objective switching training methods. According to the final evaluation results, a deeper, wider, and stronger network can improve translation performance in general, yet our data domain adaption method can improve performance even more. In addition, we found that switching to the use of our proposed objective during the finetune phase using relatively small domain-related data can effectively improve the stability of the model's convergence and achieve better optimal performance.

1 Introduction

News translation (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020) is one of the most prominent and appealing tasks in machine translation evaluation (Wu et al., 2020b; Li et al., 2020c). Our MiSS system took part in the WMT21 news translation task, including English $\rightarrow$ Chinese (En $\rightarrow$ Zh), Chinese $\rightarrow$ English (Zh $\rightarrow$ En), and Japanese $\rightarrow$ English (Ja $\rightarrow$ En) translation directions. We developed translation systems for this year's submission to investigate machine translation techniques from two perspectives: model structure and model training. All of the data used by the submitted systems is constrained. Due to a lack of training resources, the English->Japanese translation direction is only investigated from the model structure perspective.

From the perspective of model structure, we choose the Transformer (Vaswani et al., 2017; Li et al., 2021c) model based on self-attention, which is extensively utilized in neural machine translation systems, as our basis (Zhang et al., 2020b; Li et al., 2020d). On this strong foundation, we opt to simply deepen the model by increasing the number of encoder layers or widen the model by increasing the hidden size of the model to obtain a deeper or wider model. When deepening or widening the model, we found that there is no need for additional sophisticated structure design (e.g., layer drop (Fan et al., 2020) / sublayer drop (Li et al., 2021a)) or training strategy when there is adequate training data available. In addition to Transformer architecture, Wu et al. (2019) propose a dynamic convolution structure that can perform competitively or better to the self-attention structure. Follow the practice in WMT20 (Wu et al., 2020a), we also applied the dynamic convolution architecture as another basis.

According to our preliminary results on the development set, domain has a significant impact on performance, despite the fact that we are working with the resource-rich En-Zh and En-Ja language pairs. This year's submissions are mostly concerned with utilizing training approaches to mitigate the impact of domain differences. Specifically, we first use data in all hybrid domains to train the initial NMT model, and then, based on sentence embedding model enhanced by contrastive learning, the parallel/monolingual corpus is filtered monolingually or cross-lingually, and the filtered domain-related parallel corpus is used for further finetuning, and the domain-related monolingual corpus is used for in-domain back-translation enhancement. In addition, we also adopted a self-supervised training method to train the model on the given source text of the test set and its domain-related monolingual text obtained by filtering. In self-supervised

[*]Corresponding author. Zuchao Li was limited technical researcher at NICT when this work was done. This work was partially supported by the Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 154–161
November 10–11, 2021. ©2021 Association for Computational Linguistics

training, we combine our *Data-dependent Gaussian Prior Objective* (D2GPo) objective (Li et al., 2020b) to alleviate the collapse due to non-golden targets. In the finetune stage with the domain-related parallel corpus, we adopted the training strategy of switching the optimization objective from the MLE to our proposed *Dual Skew Divergence* (DSD) (Li et al., 2019). The results demonstrated that switching to the DSD objective resulted in improved convergence.

From the evaluation results, we observe substantial improvements over the strong baseline with 4.3 (En → Zh), 4.8 (Zh → En), 3.2 (Ja → En) BLEU scores on the development sets, respectively. The gains can be attributed to larger model capacity and better training strategies. And the results suggest that the cost of domain adaptation to improve performance is less than the cost of increasing model capacity.

2 Model Perspective

With the development of deep learning in NLP (He et al., 2018; Cai et al., 2018; He et al., 2019; Li et al., 2021d), model ensembling can usually produce better results than single models, and the bigger the difference between the models used for ensembling, within a certain limit, the higher the improvement will be. As a result, we chose four distinct typical architectures as the basis for single NMT models and trained them on the same data. The detailed parameters of each model architecture are shown in Table 1.

Deep Transformer Some related works (Zhang et al., 2019; Wang et al., 2019; Li et al., 2020a, 2021a) have revealed that deep networks have great advantages in NMT performance compared to shallow networks recently. Based on the Transformer NMT model architecture, we found that in the presence of sufficient training data, merely increasing the number of stacked layers of the encoder can fulfill the goal of deep Transformer without the use of additional initialization, dropout, or layer skipping techniques.

Wide Transformer Recent researches (Sun et al., 2019; Wu et al., 2020a; Zhang et al., 2020a; Wu et al., 2020b; Meng et al., 2020) have demonstrated that, in addition to deepening the NMT model, widening the model can also effectively improve translation performance, with increasing the feed-forward network (FFN) size in the Transformer model bringing less training and inference cost than increasing the overall hidden size of the model. We took a same practice in our work by increasing the FFN size and established a Wide Transformer baseline.

	Deep Transformer	**Wide Transformer**	**Deep DynamicConv**
Enc. Layers	40	20	20
Dec. Layers	6	6	6
Attn. Heads	16	16	16
Hidden Size	1,024	1,024	1,024
FFN Size	4,096	8,192	4,096

Table 1: Hyper-parameters of different model architectures. Note that Wide Transformer with relative position encoding was also used as baseline models.

Deep DynamicConv Dynamic convolution (DynamicConv) (Wu et al., 2019) was proposed as a replacement for Transformer architecture and has piqued much interest (Wu et al., 2020a) due to its good speed advantage and comparable performance. To enhance the performance of single model, we also deepen the DynamicConv model by increasing the number of encoder layers, denoted as Deep DynamicConv. The original DynamicConv model consists of 7 encoder layers and 6 decoder layers. We deepen the DynamicConv model's encoder layers to Deep DynamicConv. Because the kernel size of each convolution layer in the DynamicConv model differs, we set the kernel sizes of the 16 encoder layers in Deep DynamicConv to `[3, 7, 15, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31]` and leave the other settings unchanged from the original model.

Relative Position Encoding Because self-attention in the convention Transformer model is position-independent, the encoded features must be enhanced with explicit positional information for natural language processing. Absolute position encoding is usually employed in the Transformer NMT model. Shaw et al. (2018) proposed to add relative position encoding (RPE) for improving self-attentional features and shown additional performance gains. We also applied relative position encoding to the Wide Transformer model and created another strong baseline.

We use the identical vocabulary and data to train these four baseline models separately, and then average the best 5 checkpoints in each model's training phase to generate the final model output

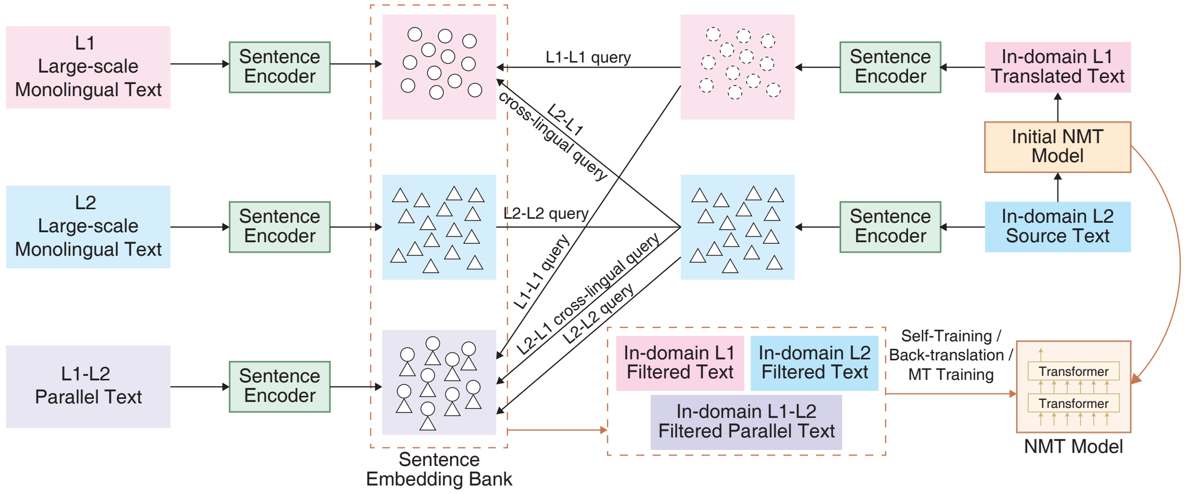

Figure 1: Illustration for contrastive learning-reinforced domain adaptation

in the corresponding stage. According to Wu et al. (2020a)'s experience, the best 5 checkpoints are determined based on the BLEU metric on the development set rather than the perplexity (PPL) metric. Furthermore, we applied the D2GPo objective (Li et al., 2020b) in the training process to obtain more stable convergence and decrease the impacts of overfitting resulting from the training set's noise.

3 Training Perspective

Contrastive Learning-reinforced Domain Adaptation Data domain issues have been found to have a significant impact on machine translation performance (Saunders, 2021). The official training data is of hybrid domain, despite the fact that the evaluation task is news translation. And, while news translation corpora can be deemed to be in the news domain, there are significant variances in news styles within the same domain. As a result, one of the keys to performance enhancement will be how to utilize the data training model that is closer to the evaluation data domain and style.

Using languages L_1 and L_2 as an example, the data that may be used comprises the parallel corpus $D^P_{L_1-L_2}$, as well as their respective large-scale monolingual corpus $D^M_{L_1}$ and $D^M_{L_2}$. Parallel corpora are typically utilized for direct training of NMT models, whereas monolingual corpora are used for back-translation (Edunov et al., 2018) and self-supervised training (Jiao et al., 2021). The domain filtering method can be utilized in these three training procedures to create corpus whose domain is more similar to the development and test sets.

Instead of relying on the co-occurrence probability of the surface tokens in the sentence, we based the domain filtering on the hypothesis that the more similar the sentence representations generated by the Transformer encoder are, the more likely they are to be dispersed in the same domain. Because the current Transformer encoder's representation is based on the bidirectional and full attention of all tokens, the combination and order of tokens have a significant impact on the final representation, the sentence representation is adequate for capturing domain information. As a result, we use the sentence embedding distance to measure the domain similarity.

We leveraged a universal paraphrastic sentence encoder (Wieting et al., 2016; Ethayarajh, 2018; Li and Zhao, 2020) to embed each given sentence to a dense representation. On a large scale monolingual corpus, we train our own monolingual and multilingual sentence encoder, a Transformer that has been pre-trained using masked language modeling (Devlin et al., 2019; Zhang et al., 2020c; Li et al., 2021b), with the XLM toolkit (Conneau et al., 2020) and fine-tuned to maximize cosine similarity between similar sentences. Contrastive learning seeks to acquire effective representation by pulling semantically close neighbors and pushing non-neighbors apart (Hadsell et al., 2006). Since this criterion precisely meets the requirements of sentence representation learning, we use contrastive learning to finetune the pre-trained sentence encoder. Figure 1 illustrates our contrastive learning-reinforced domain adaptation method.

According to the domain adaptation requirements in actual machine translation, the trained sentence encoder needs respond to four scenarios: *Original Input Monolingual Filter*, *Translated Input Monolingual Filter*, *Original Input Cross-lingual Filter*, *Translated Input Cross-lingual Filter*. Because the fourth scenario can be covered by the first, we only employ the first three scenarios in our experiment.

For all scenarios, we first follow Gao et al. (2021)'s approach to perform unsupervised training in which the input sentence itself is used as a positive instance due to there will be some differences between the sentence representations of the two pass input with the presence of the model dropout, and other sentences in the in-batch are used as negative instances.

The unsupervised contrastive learning-trained monolingual sentence encoder can be used directly as an evaluator of the similarity of sentences in the same language and to mine similar sentences from the sentence bank. However, for the non-gold translated sentences filtering, we apply the baseline NMT models to translate parallel corpus and to back-translated monolingual corpus to generate pseudo-paraphrase corpus. And then triplet loss is used to fine-tune the unsupervised sentence encoder:

$$\mathcal{L}(x, y) = max(0, \alpha - cos(x, y)) + cos(x, y_n),$$

where positive pairs (x, y) are paraphrases from translation or back-translation, y_n are in-batch negative instances.

Likewise, we still need cross-language filtering, therefore we use parallel corpus instead of synthetic pseudo-restatement corpus and triplet loss for additional finetuning on the multilingual sentence encoder.

As shown in Figure 1, taking the L_2 in-domain source sentences in development set as an example, we first use the initial NMT model to translate these sentences to L_1 translated text. The different trained sentence encoder is then used to encode these sentences and the large-scale monolingual or parallel corpus based on different scenarios respectively. Then, using the faiss toolkit[1], a query procedure is performed to locate related in-domain monolingual or parallel corpora with similarity calculation and ranking.

Back-translation and Self-supervised Training Using the in-domain monolingual and parallel corpus, we may train the initial model using back-translation and self-supervised training approaches. For back-translation, we leverage the original multiple NMT models to translate these monolinguals into various pseudo-parallel corpora, and then combine them with the in-domain parallel corpus to finetune the NMT model. For self-supervised training, we use a variety of models to perform ensemble translation on the in-domain monolingual text as the translation target and combine the in-domain translation corpus to fine-tune the model. In the specific implementation, we perform back-translation and self-supervised training consecutively such that the self-supervised training stage can exploit the stronger NMT model trained during the back-translation stage.

Optimization Objective Switching Training It is easier to fall into a local optimum in the process of back-translation and self-supervised training because there are relatively fewer in-domain data and input or output in part of the data utilized is not gold. According to our experience in (Li et al., 2019), switching the training objective to the adversarial learning objective after MLE training converges might help jump out of the local optimal state and get better performance. Follow this practice, in the back-translation and self-supervised training stages, we first employ MLE target training to converge on a development set and then switch to Li et al. (2019)'s DSD loss for further training:

$$\begin{aligned}\mathcal{L}_{DSD} = -\frac{1}{n}\sum_{i=1}^{n}[\beta(t)\mathbf{y_i}\log((1-\alpha)\hat{\mathbf{y}}_\mathbf{i} + \alpha\mathbf{y_i}) \\ -(1-\beta(t))\hat{\mathbf{y}}_\mathbf{i}\log(\hat{\mathbf{y}}_\mathbf{i}) \\ +(1-\beta(t))\hat{\mathbf{y}}_\mathbf{i}\log((1-\alpha)\mathbf{y_i} + \alpha\hat{\mathbf{y}}_\mathbf{i})],\end{aligned}$$

where $\mathbf{y_i}$ is the i-th token in the target sequence $\mathbf{y}$, $\hat{\mathbf{y}}_\mathbf{i}$ is the i-th predicted token, α is a hyper-parameter in α-skew divergence (Lee, 1999), and $\beta(t)$ is the controllable weight from the PID controller.

4 Data Setup

English↔Chinese In the English↔Chinese translation, we used all official parallel corpus, including ParaCrawl v7.1, News Commentary v16, Wiki Titles v3, UN Parallel Corpus V1.0, CCMT Corpus and WikiMatrix. For English, we use the tokenization tool provided by *Moses*[2], and

[1] https://github.com/facebookresearch/faiss

[2] https://github.com/moses-smt/mosesdecoder

	Systems	En→Zh		Zh→En		En→Ja		Ja→En	
		Dev	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*
	Transformer-big	31.67	–	33.26	–	23.31	–	21.61	–
	Deep Transformer	32.48	–	34.18	–	24.68	–	22.78	–
①	++ID-BT	35.30	–	38.94	–	–	–	24.46	–
②	++ID-ST	35.95	–	39.18	–	–	–	**25.82**	–
	Wide Transformer	32.67	–	34.01	–	24.27	–	23.20	–
③	++ID-BT	35.37	–	38.82	–	–	–	24.55	–
④	++ID-ST	**36.15**	–	39.13	–	–	–	25.71	–
	Deep DynamicConv.	32.39	–	33.68	–	24.08	–	21.91	–
⑤	++ID-BT	35.01	–	38.66	–	–	–	24.37	–
⑥	++ID-ST	36.03	–	39.05	–	–	–	25.66	–
	Wide Transformer w/ RPE	32.52	–	34.35	–	**24.76**	–	22.78	–
⑦	++ID-BT	35.55	–	38.91	–	–	–	24.48	–
⑧	++ID-ST	36.08	–	**39.20**	–	–	–	25.71	–
	Baseline Ensemble	32.79	31.9	34.47	27.8	24.79	42.6	23.15	23.8
	Ensemble: ① + ③ + ⑤ + ⑦	35.62	35.7	38.98	32.4	–	–	24.63	26.4
	Ensemble: ② + ④ + ⑥ + ⑧	**36.41**	**36.2**	**39.25**	**32.6**	–	–	**25.99**	**27.0**

Table 2: BLEU evaluation results on the WMT 2021 development and test sets. The BLEU in the development set is a word-level MultiBLEU score, but the BLEU in the test set is from the official evaluation. Due to a lack of resources, En→Ja only completed the baseline training and ensemble submission.

for Chinese, we use *pkuseg* (Luo et al., 2019) as the word segmentor. We adopt a joint byte pair encoding (BPE) (Sennrich et al., 2016) with 44K operations for subword vocabulary in English and Chinese. Punctuation normalization is not employed to preprocess the training data in order to prevent complex post-processing of punctuation restoration. For English post-processing, we use the script in *Moses* to de-tokenize the translation, whereas for Chinese, we employ sacremoses[3] for de-segmentation.

English↔Japanese In the English↔Japanese translation, data for training were combined from ParaCrawl v7.1, News Commentary v16, Wiki Titles v3, WikiMatrix, The Kyoto Free Translation Task Corpus, and TED Talks. Similarly, the Japanese sentences are segmented using the Mecab[4] segmentor, while the English sentences are processed using the Moses tokenizer. The size of the English and Japanese joint BPE is also set to 44K. In post-processing, Moses script and sacremoses are also employed for detokenization.

We merged the whole news-crawl corpus for monolingual data. However, in Chinese and Japanese, news-crawl corpus alone is insufficient to train the sentence encoder, so we sampled some data from the common-crawl corpus and eventually produced the data in English, Chinese, and Japanese 100M sentences each. For pre-processing, we exclude sentences that are more than 175 words long, and the word ratio between the source and the target greater than 1:2 or 2:1.

5 Model Training

All of our NMT models are built using the Fairseq toolkit. Except for the switching training phase, all models are optimized with Adam optimizer, and SGD optimizer is utilized for optimization training when switching to DSD loss. During the baseline model training process, the learning rate is scheduled using the inverse sqrt scheduler with 4000 warm-up steps, maximum learning rate 5e-4, and betas (0.9, 0.98). Each model is trained on 8 NVIDIA V100 GPUs, with batch size limited to 8192 tokens per GPU. FP16 is emploted to save GPU memory and speed up calculations. To increase the virtual batch size, we set the gradient update steps to 8 during the training phase. The label smoothing and dropout values are both set to 0.1. In the finetuning stage, we utilize a smaller batch size, 4,096 tokens per GPU, and train the model at a fixed learning rate of 1e-4. Sentence encoder models are developed with the XLM toolkit, and the architecture is based on the BERT-base. The hidden size, heads, hidden layers, and FFN size are 768/12/12/3072 respectively. During training, a early stop mechanism is applied in which the training will stop when the PPL on the development set does not decrease after 25 epochs.

[3] https://github.com/alvations/sacremoses

[4] https://github.com/taku910/mecab

6 Results and Analysis

Table 2 shows the results on the development sets as well as the official evaluation results on the WMT21 test sets. First, when comparing Deep Transformer, Wide Transformer, and Transformer-big, we observed that increasing the number of model layers or widening the model to increase the number of model parameters can result in large performance benefits. Second, Deep DynamicConv has shown comparable results to Deep Transformer in multiple data sets, demonstrating that DynamicConv is a viable replacement option for Transformer. Third, the Deep Transformer w/ RPE model outperforms Deep Transformer model in most circumstances, demonstrating that machine translation benefits from additional relative position encoding information. Fourth, in-domain back-translation (ID-BT) and in-domain self-supervised training (ID-ST) improve the model's performance substantially more than the increased model parameters, indicating that the data domain is a primary factor limiting translation performance. Furthermore, these enhancements demonstrate that our domain adaption approach of contrast learning-reinforced is a effective approach. Finally, we performed ensemble on the four finetuned baselines and received even higher results, demonstrating that the models of the four architectures differ from each other.

7 Conclusion

In this paper, we introduce our MiSS translation system, which participated in the WMT21 news translation task. We developed a new contrast learning-reinforced domain adaptation strategy in this work, and the experimental findings suggest that this method may significantly increase translation performance. Furthermore, we conducted experiments on a range of model architectures. Our domain adaption strategy improved these strong baseline models significantly, illustrating the method's generality and indicating that the performance deficiency is not due to a specific model structure.

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2753–2765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Kawin Ethayarajh. 2018. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Edouard Grave, and Armand Joulin. 2020. Reducing transformer depth on demand with structured dropout. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society.

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.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2840–2850, Online. Association for Computational Linguistics.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, College Park, Maryland, USA.

Bei Li, Ziyang Wang, Hui Liu, Quan Du, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2021a. Learning light-weight translation models from deep transformer. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13217–13225. AAAI Press.

Bei Li, Ziyang Wang, Hui Liu, Yufan Jiang, Quan Du, Tong Xiao, Huizhen Wang, and Jingbo Zhu. 2020a. Shallow-to-deep training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 995–1005, Online. Association for Computational Linguistics.

Yian Li and Hai Zhao. 2020. Learning universal representations from word to sentence. *arXiv preprint arXiv:2009.04656*.

Zuchao Li, Kevin Parnow, Hai Zhao, Zhuosheng Zhang, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2021b. Cross-lingual transferring of pre-trained contextualized language models. *CoRR*, abs/2107.12627.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020b. 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, Zhuosheng Zhang, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2021c. Text compression-aided transformer encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zuchao Li, Hai Zhao, Shexia He, and Jiaxun Cai. 2021d. Syntax Role for Neural Semantic Role Labeling. *Computational Linguistics*, pages 1–46.

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, Yingting Wu, Fengshun Xiao, and Shu Jiang. 2019. Controllable dual skew divergence loss for neural machine translation. *arXiv preprint arXiv:1908.08399*.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain Chinese word segmentation. *CoRR*, abs/1906.11455.

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, and Hao Zhou. 2020. WeChat neural machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 239–247, Online. Association for Computational Linguistics.

Danielle Saunders. 2021. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *arXiv preprint arXiv:2104.06951*.

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.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 374–381, Florence, Italy. 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.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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.

Liwei Wu, Xiao Pan, Zehui Lin, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2020a. The volctrans machine translation system for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 305–312, Online. Association for Computational Linguistics.

Shuangzhi Wu, Xing Wang, Longyue Wang, Fangxu Liu, Jun Xie, Zhaopeng Tu, Shuming Shi, and Mu Li. 2020b. Tencent neural machine translation systems for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 313–319, Online. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention. 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 898–909, Hong Kong, China. Association for Computational Linguistics.

Yuhao Zhang, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, Yongyu Mu, Jingnan Zhang, Xiaoqian Liu, Xuanjun Zhou, Yinqiao Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2020a. The NiuTrans machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 338–345, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020b. Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020c. Semantics-aware BERT for language understanding. 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 9628–9635. AAAI Press.