Semantic Perception-Oriented Low-resource Neural Machine Translation

Nier Wu[1], Hongxu Hou[2⋆], Haoran Li[3], Xin Chang[4], and Xiaoning Jia[5]

College of Computer Science-college of Software, Inner Mongolia University, China
[1]wunier04@126.com, [2⋆]cshhx@imu.edu.cn, [3]489633848@qq.com
[4]changxin03@163.com [5]jiaxning@163.com

Abstract. Pre-training method has been proved to significantly improve the performance of low-resource neural machine translation (NMT), while the common pre-training methods (BERT) uses attention mechanism based on Levenshtein distance (LD) to extract language features, which ignored syntax-related information. In this paper, we proposed a machine translation pre-training method with semantic perception which depend on the traditional position-based modeling, we uses semantic role labels (SRL) to annotate sentences with "predicate-argument" structures at the word level, and merge vectorized SRL with word vectors to deepen the model's understanding of deep semantics. In addition, to avoid parameter disaster, we proposed a hierarchical knowledge distillation method to fuse the NMT model and pre-training model to adapt to the output probability distribution of the pre-training model. We validated the method in the LDC En-Zh and CCMT2017 Mongolia-Chinese (Mo-Ch), Uyghur-Chinese (Uy-Ch), Tibetan-Chinese (Ti-Ch) tasks. The results show that compared with baseline, our model achieves significant results, which fully illustrates the generalization of the method.

Keywords: Pre-training · SRL · Machine translation.

1 Introduction

The NMT method based on encoder-decoder framework [1] [2] encodes the source language X into a set of continuous vector representation Z from left to right, and decodes the target language Y from Z in the same way. The model mainly adopts recurrent neural network (RNN) structure to encode the source language in a linear manner, resulting in the feature extraction ability being limited by the linear distance between the word embedding and its context, it means that correlation is inversely proportional to linear distance. In the early stage of training, there is less context information available and semantic relationships cannot be fully learned. Especially in low-resource tasks, the problem is more obvious due to sparse data. In order to alleviate these problems, [3] proposed a parallel encoding method, which is free from the constraints of time series encoding, and the

⋆ Corresponding Author

understanding of language is no longer limited to its linear context information, but to extract features from the global perspective. While the Transformer model [4] based on the self-attention mechanism can learn the weight ratio between words in the language through the self-learning method, so that the model can improve the implicit learning ability of language knowledge. However, whether sequential encoding or parallel encoding is used, the model only learns the explicit structural information of the sentence (LD), and can not mine deeper grammatical information, resulting in the model effect not being significantly improved. Therefore, the NMT method that integrates grammatical information has gradually shows its potential. [5] proposed a syntactic attention-related NMT method to calculate the linear context and syntactic context information corresponding to the current target token by using the dual-attention mechanism, the decoder can accurately predict the target token according to the dual context representation. [6] proposed a novel NMT method for fusing abstract semantic representations (AMR), which used graph recurrent networks (GRN) to represent AMR information. Compared with syntactic tree, AMR greatly retains meaningful words in sentences and ignores non-contributing words, which improved the translation quality. [7] is similar to that of [6] in that it uses graph convolution network (GCN) to model the syntactic dependency tree corresponding to the source language and adds it to the top of the convolutional encoder to integrate word embedding and the syntax tree information.

For low-resource tasks, the quality of syntax-based NMT models is limited due to the lack of sufficient parallel corpus and corresponding syntax Treebank. In recent years, profit from the powerful semantic feature extraction ability of BERT, the pre-training method based on BERT has also been gradually applied to many NLP tasks including machine translation. [8] proposed a syntax-infused Transformer and BERT models for machine translation method, which adopted the BERT model to learn the position-aware context representation and regard it as the input of Transformer encoder. [9] proposed a BERT-based machine translation pre-training method, they fed the representation of BERT to all sublayers of NMT model and use the attention mechanism to adaptively control how each layer interacts with the representation of BERT. [10] proposed a knowledge distillation method using dynamic fusion strategy to provide pre-trained knowledge for NMT models. They use an adapter to dynamically transform the general representations in the pre-training model into representations more suitable for NMT model. In addition, the NMT model can fully learn the output distribution of the pre-training model through the knowledge distillation method. However, the pre-training method combined with rich semantic representation has not been widely applied to NMT tasks. With the advantages of BERT methods, this paper proposed a deep semantic perception assisted neural machine translation pre-training method, which includes the following contents.

– We proposed a target-oriented language pre-training method, which uses an improved BERT model to learn the implicit semantic features of the target language (the target language in this paper is Chinese).

– To alleviate the deficiency of syntax treebank, we adopted more easily available SRL labels to represent semantic information. We built a vector lookup table to obtain the vector corresponding to the SRL label, the use BiGRU to encode each vector, and finally splice different forms of SRL labels through the full connection layer.
– We use the hierarchical knowledge distillation method to instruct the output of each layer in the decoder, so that the NMT model can fully learn "prior knowledge" from the pre-training model.

2 Background

Neural Machine Translation The NMT model based on attention mechanism simulates the translation probability $P(y|x)$ of the source language $X = \{x_1, ..., x_n\}$ to the target language $Y = \{y_1, ..., y_m\}$ word by word, as shown in Equation 1.

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

Where $y_{<i}$ indicates the partial translation result before the i-th decoding step, and θ indicates the parameters of the NMT model. The NMT model uses the maximum likelihood estimation method to optimize the parameters θ. For the parallel sentence pair $\{[x^n, y^n]\}_{n=1}^{N}$ in training set, the loss defined as shown in Equation 2.

$$L_{CE} = \underset{\theta}{argmax} \sum_{n=1}^{N} logP(y^n|x^n; \theta) \tag{2}$$

Although the argumentation method is widely used, the problem of exposure deviation still exists, which also directly affects the quality of the NMT model.

NMT assisted by pre-training method The pre-training method transfers knowledge from resource-rich tasks to low-resource tasks. However, the NMT method takes the cross entropy between the two languages as the training goal to optimize the parameters, which is significantly different from the monolingual pre-training model.

Therefore, one approach is to use the resource-rich language pre-training model, and then put source language and the target language into the pre-training model to obtain the corresponding word embedding, and use pre-trained word embedding training NMT model. Another approach is to design a new sequence-to-sequence pre-training task to directly realize bilingual mapping in machine translation. Among them, XLM [11], MASS [12] and BART [13] are both cross-lingual pre-training method based on sequence-to-sequence.

3 Method

This section is mainly divided into the following aspects: semantic perception-assisted pre-training model and hierarchical knowledge distillation training process.

3.1 Semantic perception-assisted pre-training model

Obtain semantic role label SRL mainly takes the sentence as the unit and analyzes the predicate-argument structure of the sentence. Specifically, the task of SRL is to take the predicate as the center, explore the relationship between the various components in the sentence and the predicate, and use semantic roles to describe the relationship (argument). Generally, the process of SRL includes: syntax analysis-candidate argument pruning-argument recognition-argument labeling. According to the results of syntactic analysis, the part that is absolutely not an argument is pruned, and then the binary classification is used to determine whether the remaining part is argument, if so, the semantic category to which it belongs will be marked.

Given a sentence w, various predicate parameter structures are generated. To reveal the multidimensional semantics of sentences, we group different semantic labels corresponding to the same sentence and embed them with text into the next encoding component. The specific method is to input the sentence-predicate pairs (w, v) into the high-speed BiGRU to search for the predicate's argument, and the semantic role corresponding to argument is marked as y. The goal of prediction is to obtain the semantic role label sequence with the highest score among all possibilities Y. As shown in Equation 3.

$$\hat{y} = \underset{y \in Y}{argmax} f(w, y) \tag{3}$$

Where $f(\cdot)$ indicate the nonlinear activation function in the BiGRU. To improve the prediction accuracy, the BIO constraints and semantic role label constraints are added. [14] made a detailed explanation, I won't repeat it.

Pre-training model combined with SRL As shown in Figure 1, pre-training model includes text encoding module and SRL encoding module. The text encoding module is similar to the common BERT. For sentence $X = \{x_1, ..., x_n\}$, we employ BERT model to capture the context information of each word segment and generate the corresponding context word embedding sequence.

For a sentence that contains the m semantic role label sequences associated with the predicate, $T = \{t_1, ..., t_m\}$, and the i-th label sequence t_i contains n labels, which can be expressed as $t_i = \left\{lbl_1^i, ..., lbl_n^i\right\}$, because semantic labels belong to the word level, the number of labels is equal to the sentence length. We construct a vector table that maps the semantic role labels to the corresponding vectors $\left\{v_1^i, ..., v_n^i\right\}$, and feed the vector to BiGRU to capture the hidden state

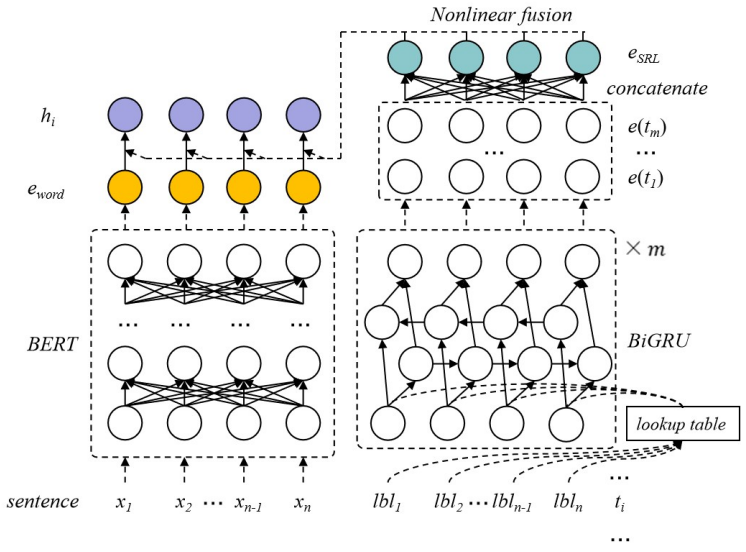

Fig. 1. Pre-training model combined with SRL. Word embedding (yellow circle) and semantic role representation (cyan circle) are obtained by BERT and BiGRU respectively. Then the new word embedding representation (purple circle) is obtained by nonlinear fusion method and fed to the NMT model.

representation of the semantic labels, so as to extract the feature of the label sequences. As shown in Equation 4.

$$e(t_i) = BiGRU(v_1^i, ..., v_n^i) \tag{4}$$

Where $0 < i \leq m$, we assume that L_i represents the set of label sequences corresponding to the i-th predicate token x_i, and the vector representation is defined as $e(L_i) = \{e(t_1), ..., e(t_m)\}$. Finally, we concatenate the m sequences of label representation and feed them to a fully connected layer to obtain an accurate label representation, as shown in Equation 5.

$$\begin{aligned} e_{concat}(L_i) &= W[e(t_1), ..., e(t_m)] + b, \\ e_{SRL} &= \{e_{concat}(L_1), ..., e_{concat}(L_n)\} \end{aligned} \tag{5}$$

Where W indicates the weight, and b represents bias. e_{SRL} represents the embedding of the semantic label sequence corresponding to each predicate in a sentence.

The original sequence can be expressed as $e_{word} = \{e(x_1), ..., e(x_n)\}$. Then, word embedding and SRL embedding are concatenated by function $h = e_{SRL} \diamond e_{word}$.

3.2 Hierarchical knowledge distillation training process

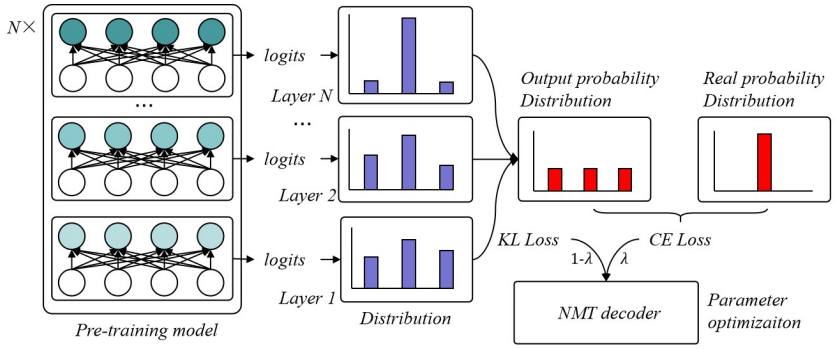

Fig. 2. Hierarchical distillation.

Model	**ALBERT**	**BERT**	**GPT-2**	**XLNet**	**T5**
Parameters	12mil.	0.1bn	1.5bn	2bn	11bn

Table 1. Parameter scale of various pre-training models.

Table 1 shows several mainstream pre-training models at present. It can be seen that the parameters scale of the model is very large, resulting in the method of parameter transfer or word embedding transfer method are difficult to achieve in neural machine translation model. Therefore, we proposed a hierarchical knowledge distillation method, which selectively extracts the output of a specific layer from the pre-training model and guides NMT model training according to its probability distribution. As shown in Figure 2.

Generally, there is a temperature hyper-parameter τ in knowledge distillation method, and a smoother output distribution probability can be learned by increasing τ. The output probability is shown in Equation 6.

$$p_{prt} = \frac{exp(z_{prt_i}/\tau)}{\sum_j exp(z_{prt_j}/\tau)} \tag{6}$$

Where z_{prt_i} represents hidden layer state, the equation also applies to the NMT model. Generally, a well pre-trained model will generate distributions with high probability for a few words, leaving others with probabilities close to zero. By increasing τ we expose extra information to the NMT model. In addition, for each layer of the pre-training model, it has certain representation ability. When some intermediate layers already have high distribution probability or confidence, the subsequent layers do not need to calculate KL divergence, which called "adaptive inference". While reducing resource consumption, it can still maintain a high prediction probability, so that the distillation data of the pre-training model has better indicative ability. The calculation of KL divergence is shown in Equation 7.

$$D_{KL}(p_{prt}||p_{nmt}) = \sum_{i=1}^{N} p_{prt}(i) \cdot log\frac{p_{prt}(i)}{p_{nmt}(j)} \tag{7}$$

Where p_{prt} and p_{nmt} represent the probabilities of the pre-training model and NMT model respectively, the pre-training model is regarded as a teacher, NMT model is regarded as student. i represents the sentence number. Therefore, we define the KL divergence (relative entropy) loss between the output probability distribution of the pre-trained model and the NMT model, as shown in Equation 8.

$$L_{KL}(p_{prt_0}, ..., p_{prt_{N-2}}, p_{nmt}) = \sum_{i=0}^{N-2} \tau^2 D_{KL}(p_{prt_i} || p_{nmt}) \tag{8}$$

Where $p_{prt_0}, ..., p_{prt_{N-2}}$ represents the output probability of various intermediate layers in the pre-training model. Our goal is to minimize the KL divergence loss of the selected intermediate layers in the pre-training model and the NMT model, so that the two distributions are gradually similar.

To improve the utilization of the intermediate layers and ensure that the output probability distribution can optimize the NMT model, we set a threshold U, when the result is less than the threshold, we can distill the output data in advance. Otherwise, we need to calculate the output distribution of subsequent layers and repeat the process until the last layer of the pre-training model. The calculation of threshold U is shown in Equation 9.

$$U = \frac{\sum_{i=1}^{N} p_{prt}(i) log p_{prt}(i)}{log \frac{1}{N}} \tag{9}$$

The variables in the Equation 9 have been explained above and will not be repeated. For convenience, we use entropy as the threshold value. If the entropy value is large, the confidence is low, and if the entropy value is small, the result can be output, which not only saves subsequent calculation resources but also improves the inference speed. For this reason, the objective function of the model can be regarded as the weighted sum of cross-entropy loss (See equation 2) and relative entropy loss (See equation 8), as shown in Equation 10.

$$L = \lambda L_{CE} + (1 - \lambda) L_{KL} \tag{10}$$

We set λ to 0.5.

4 Experiments

4.1 Datasets and configuration

We conducted experiments on English-Chinese (En-Zh) and three low-resource translation tasks (Mo-Zh, Uy-Zh, Ti-Zh). For En-Zh task, the training set consist of 1.2 million bilingual sentences from LDC corpus[1], we use NIST02 as validation set, and NIST03-06 as the test set. For low-resource translation task, the data sets are provided by CCMT2017, as shown in Table 2.

[1] http://www.ldc.upenn.edu/

	Training	Valid	Test
Mo-Zh	64752	500	500
Uy-Zh	542796	1000	1000
Ti-Zh	30004	500	500

Table 2. Data sets for three low-resource machine translation tasks.

In addition, we limit the bilingual vocabulary to 35K words and limit the length of sentences to 80. We utilized BLEU scores[2] to evaluate the quality. The parameters are updated by stochastic gradient descent (SGD), and the learning rate is dynamically adjusted by adam and the initial value is set to 0.0001. Word embedding dimension and hidden layer set 512, the beam size is 8, we apply dropout to avoid over-fitting, with dropout rate being 0.2, set U to 0.2. Since the pre-training model combined with SRL is used to guide the distribution probability of the target language prediction, we perform semantic role labeling on the target language (Chinese), and the label sets comes from Chinese Binzhou Proposition Bank (CPB)[3].

Our pre-training model is improved on SemBERT[4]. For the NMT model, we improved the Transformer[5] to implement our approach. To verify the effectiveness of the model, we also adopted two NMT methods combined with the pre-training model as the baseline,**BERT4NMT [9]**: A NMT method that integrated BERT pre-training model, and extracted the knowledge of pre-training model by introducing BERT-based attention mechanism.**AK4NMT [10]**: They used fusion strategy to transformed pre-trained word embedding into a more suitable representation for NMT task, and employed the distillation method to learn the output probability distribution of the pre trained model. We employed two TITAN X to train the model and obtained by averaging the last 5 checkpoints for the translation tasks.

4.2 Results and analysis

Results According to Table 3, our model compared to the traditional Transformer, BLEU scores improved by 1.75, 0.49, 1.81, 1.28 and 2.58, respectively, and the average BLEU scores increased by 1.18. Table 4 shows the experimental

Model	NIST02	NIST03	NIST04	NIST05	NIST06	AVG
Transformer	37.19	36.66	37.06	34.89	35.23	35.96
BERT4NMT [9]	38.01	36.67	38.45	34.92	34.02	36.01
AK4NMT [10]	37.92	36.25	38.08	35.53	36.29	36.54
Our model	**38.94**	**37.15**	**38.87**	**36.17**	**37.81**	**37.5**

Table 3. En-Zh translation results in LDC corpus.

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl
[3] http://verbs.colorado.edu/chinese/cpb/
[4] https://github.com/cooelf/SemBERT
[5] https://github.com/tensorflow/tensor2tensor

results of three low-resource translation tasks, it can be seen that our method has also improved 4.18, 3.61 and 3.59 BLEU scores in three low-resource machine translation tasks.

Analysis Compared with two NMT models based on pre-training methods, our pre-training method combines SRL to obtain additional information. Meanwhile, due to the limitation of the MLE algorithm, the NMT model only relies on real translations to predict the target word, making the output distribution of the model more concentrated and cannot be effectively generalized to other words. Therefore, the distillation data output by the pre-trained model is better used to optimize the NMT model by adjusting the temperature hyper-parameter τ, and then extract more semantic representations.

Model	**Mo-Zh**	**Uy-Zh**	**Ti-Zh**
Transformer	27.18	32.41	24.29
BERT4NMT [9]	29.03	34.96	25.85
AK4NMT [10]	29.51	35.77	26.09
Our Model	**31.36**	**36.02**	**27.88**

Table 4. The BLEU scores for three low-resource machine translation tasks.

4.3 Ablation experiment

The ablation experiment in this paper is mainly used to observe the effect of the proposed method on the quality of the translation, including whether the pre-training model is combined with SRL, whether the encoder and decoder of the NMT model use pre-trained word embedding, and whether knowledge distillation or hierarchical knowledge distillation is used. See table 5 for details.

Pre-training	**Module**	**En-Zh**	**Mo-Zh**	**Uy-Zh**	**Ti-Zh**
BERT	Emb-Enc	36.1	28.15	32.98	25.26
	Emb-All	36.02	28.28	33.05	25.47
	+KD	36.39	29.78	33.96	26.17
	+H-KD	36.92	30.11	34.79	26.85
SRLBERT	Emb-Enc	36.59	29.72	33.18	26.12
	Emb-All	36.72	29.26	33.25	25.87
	+KD	37.03	30.75	34.98	27.17
	+H-KD	**37.5**	**31.36**	**36.02**	**27.88**

Table 5. The Ablation Experiment.

According to Table 5, when using general pre-training word embedding, whether the word embedding applied to the encoder (Emb-Enc) or encoder-decoder (Emb-All), it does not significantly improve the quality of translation. However, when the word embedding generated by the BERT model integrated with semantic role labels (SRLBERT) is used, the quality of the model has been improved to a certain extent. Meanwhile, compared with the NMT model using

general knowledge distillation (KD), the hierarchical knowledge distillation (H-KD) method has better representation ability and translation effect. This shows that the intermediate representation of the pre-trained model also has accurate semantic feature extraction ability.

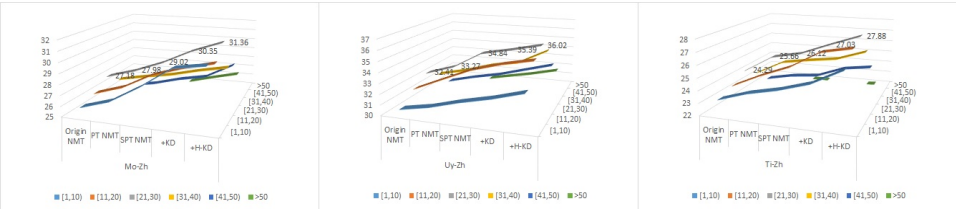

Fig. 3. BLEU scores for sentences of different length.

4.4 Case study

We observed the BLEU scores with different length and the quality of translation obtained by various methods. According to Figure 3, when the sentence length is about 20 to 30 tokens, the model has better performance. It can be seen that the NMT model using pre-training word embedding combined with semantic role labeling (SPTNMT) and hierarchical knowledge distillation (H-KD) methods has better generalization performance and translation fidelity.

	[illegible]
Reference	被告 的 辩护 律师 想 把 案件 发生 时 被告 精神 失常 ，为 被告 不 承担 法律 责任 进行 辩护 。
Transformer (NMT)	被告 辩护 律师 把 案件 被告 神经 失准 ， 为 不 承担 法律 责任 辩护 。
Pre-trainingNMT(PT-NMT)	被告 辩论 士 想 案件 发生 完毕 被告 失落 ,使 被告 不 承担 责任 辩护 。
SRL_PT-NMT	被告 辩论 律师 把 案件 发生 后 被告 精神 失落 ,使 被告 无法 承担 法律 辩护 。
+KD	被告 的 辩论 律师 把 案件 发生 被告 精神 错乱 ,使 被告 无法 承担 责任 进行 辩论 。
+H-KD	被告 的 辩护 律师 把 案件 发生 时 被告 精神失常 ，为了 被告 不用 承担 法律责任 辩解 。

Fig. 4. Case analysis.

As shown in Figure 4, taking Mo-Zh translation tasks as an example, our method significantly improves the translation fluency and faithfulness compared with the translation generated by the pre-training model based on BERT. It can be seen that our method pays more attention to semantic coherence in the process of context generation. Meanwhile, due to the use of pre-trained word embedding combined with semantic roles labels, our proposed NMT model can effectively representation the context when predicting verbs or nouns and the words with reference relations. In addition, the hierarchical knowledge distillation method can also be used to provide more choices for translation, so as to significantly improve the generalization ability of the model.

5 Conclusion

We proposed a pre-training method that integrates semantic role labeling, and embed the words generated by the pre-training model into the NMT model for training. Meanwhile, to improve the prediction accuracy of the decoder and improve generalization ability, we proposed a hierarchical knowledge distillation method to guide the NMT model to learn the output probability distribution of the pre-trained model, so that the NMT model can comprehensively learn the probability distribution of the translation. Experiments show that our method has shown significant effects on large-scale corpus translation tasks and low-resource translation tasks. In the future, we will continue to study syntax-based pre-training methods and merge with NMT model to improve translation quality.

References

1. Ilya Sutskever, Oriol Vinyals, Quoc V. Le: Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, pp. 3104-3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks
2. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio: Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations, ICLR 2015, San Diego. http://arxiv.org/abs/1409.0473
3. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin: Convolutional Sequence to Sequence Learning. Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, pp. 1243-1252. http://proceedings.mlr.press/v70/gehring17a.html
4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention is All you Need. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5998-6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need
5. Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Tiejun Zhao: Syntax-Directed Attention for Neural Machine Translation. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Sympo-

sium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 4792-4799. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16060
6. Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, Jinsong Su: Semantic Neural Machine Translation using AMR. Trans. Assoc. Comput. Linguistics, 2019. vol:7, pp. 19-31. https://transacl.org/ojs/index.php/tacl/article/view/1474
7. Diego Marcheggiani, Joost Bastings, Ivan Titov: Exploiting Semantics in Neural Machine Translation with Graph Convolutional Networks. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pp. 486-492. doi:10.18653/v1/n18-2078
8. Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, Lawrence Carin: Syntax-Infused Transformer and BERT models for Machine Translation and Natural Language Understanding. CoRR, vol. abs/1911.06156 (2019). http://arxiv.org/abs/1911.06156
9. Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, Tie-Yan Liu: Incorporating BERT into Neural Machine Translation. 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. https://openreview.net/forum?id=Hyl7ygStwB
10. Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, Weihua Luo: Acquiring Knowledge from Pre-Trained Model to Neural Machine Translation. 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. pp. 9266-9273. https://aaai.org/ojs/index.php/AAAI/article/view/6465
11. Alexis Conneau, Guillaume Lample: Cross-lingual Language Model Pretraining. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 7057-7067. https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html
12. Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu: MASS: Masked Sequence to Sequence Pre-training for Language Generation. Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. vol. 97. pp. 5926-5936. http://proceedings.mlr.press/v97/song19d.html
13. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 7871-7880. doi:10.18653/v1/2020.acl-main.703
14. Shexia He, Zuchao Li, Hai Zhao: Syntax-aware Multilingual Semantic Role Labeling. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 5349-5358. doi:10.18653/v1/D19-1538