Hot-start Transfer Learning combined with Approximate Distillation for Mongolian-Chinese Neural Machine Translation

Pengcong Wang[1], Hongxu Hou[2⋆], Shuo Sun, Nier Wu, Weichen Jian, Zongheng Yang, and Yisong Wang

College of Computer Science, Inner Mongolia University
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
[1]32009109@mail.imu.edu.cn, [2]cshhx@imu.edu.cn

Abstract. When parallel training data is scarce, it will affect neural machine translation.For low-resource neural machine translation (NMT), transfer learning is very important, and the use of pre-training model can also alleviate the shortage of data.However, the good performance of common cold-start transfer learning methods is limited to the cognate language realized by sharing its vocabulary. Moreover, when using the pre-training model, the combination of general fine tuning methods and NMT will lead to a serious problem of knowledge forgetting.Both methods have some defects, so this paper optimizes the above two problems, and applies a new training framework suitable for low correlation language to Mongolian-Chinese neural machine translation.Our framework includes two technologies: a) word alignment method under hot-start, which alleviates the problem of word mismatch between the transferred subject and object in transfer learning. b) approximate distillation,not only retains the pre-trained knowledge,but also solves the forgetting problem, so that the encoder of NMT has stronger language representation ability.The results show that BLEU is increased by 3.2, which is better than ordinary transfer learning and multilingual translation system.

Keywords: Transfer learning, Pre-training, Machine translation

1 Introduction

Despite the rapid development of neural machine translation [1] recently, its main improvements and optimizations can not be easily applied to language pairs with insufficient resources.Basic training procedure of NMT does not work well with only a few bilingual data [2], and collecting bilingual resources is difficult for many languages. With fewer parallel corpora and sparse data, it is easy to cause over fitting problems in the training process. The trained model has poor robustness and generalization ability.

In order to solve this problem, unsupervised and transfer learning [3] methods are generally used to improve the quality of the model with the help of external resources.

⋆ Corresponding Author

However, the unsupervised method has no annotation set, so the cross entropy method cannot be used for tuning.Back-translation [4] will produce false corpus, and the increase of false corpus will also produce noise, resulting in inaccurate translation.The concept of transfer learning is: We pre-train an NMT model on a high-resource language pair (parent language pair), and then continue training on the model using the bilingual data of another low-resource language pair (child language pair).This method can alleviate the poor performance of the model caused by less corpus, but it also has some shortcomings.

Some studies [5] show that the most important problem in transfer learning is the vocabulary mismatch between the transfer subject and the transfer object, which seriously limits the translation performance.We regard the source language in the parent language pair as the transfer subject, and the source language in the child language pair as the transfer object. If the words of the subject and object can be correctly corresponding during the transfer, the performance of transfer learning will be greatly improved. In previous studies [6], transfer learning is divided into hot-start method and cold-start method according to whether there is training data of child language pairs when training the parent model.The cold-start method does not use sublanguage for data. In contrast, in the hot-start method, we have available sublanguage pair training data when training the parent model. We can use sublanguage pair knowledge to solve this problem. In this paper, a cross-lingual word embedding method is used to convert words, and a semi-supervised method is used to correctly correspond the two languages without shared sub words. It can alleviate the word mismatch mentioned above and effectively improve the translation quality.

The pre-train models have demonstrated their excellent performance in various natural language processing tasks including translation tasks. Now the common training paradigm is ”pre-train + fine-tune”, which means that specific downstream tasks are tuning on the pre-trained model, so that additional knowledge can be obtained when training downstream tasks.

However, compared with other tasks that directly fine-tune the pre-trained model, NMT has two obvious characteristics, the availability of large training data (10million or more) and the high capability of the baseline NMT model (i.e., Transformer).These two features need a lot of updating steps in the training process in order to adapt to the large capacity model well.However, too many updates will lead to disastrous forgetting [7]. Too many updates in training will forget the general knowledge before training.Since the output of the pre-train model and the encoder output of NMT are essentially language models, this paper does not use the ”pre-train + fine-tune” method, but chooses the approximate distillation method to integrate the pre-trained knowledge into the encoder, so as to enhance the language representation ability of the NMT encoder and avoid the forgetting problem caused by a large number of updates.

This work proposes a new framework to adapt the transfer learning of neural machine translation to low-resources languages:

– Cross-lingual word embedding under hot-start is used to alleviate the problem of word mismatch between the transfer subject and the transfer object.

- The approximate distillation method is used to ensure that the NMT model can retain the previously trained knowledge and enhance the generalization ability of the NMT encoder.

2 Background

2.1 NMT

Neural Machine Translation is essentially an encoder decoder system.Typical NMT structures include RNN, LSTM, Transformer, etc. The function of encoder is to encode the source language sequence and extract the information in the source language. Then the information is converted to the target language by the decoder, so as to complete the translation of the language.

The training task is to map the source language sequence $X = \{x_1, x_2...x_n\}$ to the target language sequence $Y = \{y_1, y_2...y_m\}$. The sequence length can be different. In this case here, n and m are the length of source sequence and target sequence respectively.The model is trained on a parallel corpus by optimizing for the cross-entropy loss with the stochastic gradient descent algorithm.

$$L_{nmt} = -\sum_{i=1}^{m} \log p(\theta)(y_i | y_{<i}, X) \tag{1}$$

p(θ) is probability, θ is a set of parameters: source/target word embedding, encoder, decoder, and output layer.The training objective is to minimize the loss in equation (1) to obtain the optimal translation results.In Transformer, the encoder is similar to the decoder in structure. The decoder is essentially a language model of language y. Similarly, the encoder with an additional output layer can also be seen as a language model. Therefore, it is natural to transfer the pre-trained knowledge to the encoder and decoder of NMT.

2.2 Transfer Learning

Generally speaking, transfer learning refers to reusing knowledge from other fields/tasks when facing new problems [8]. Especially when there is not enough training data to solve this problem, transfer learning can play a better role. Because the hidden layer of neural network can implicitly learn the general representation of data, the weight of hidden layer can be copied to another network to transfer knowledge.

In NMT, the earliest transfer learning method was proposed by Zoph et al. [3] . In their work, the parent model was first trained on high resource language pairs. Then, the source word embedding is copied together with the rest of the model, and the ith parent language word embedding is assigned to the ith child language word. Because the parent and child source languages have different vocabularies, this is equivalent to embedding the parent source words and randomly assigning them to the child source words. In other words, even if a word exists in both parent and child vocabularies, it is unlikely that it will be assigned the same embedding in both models.

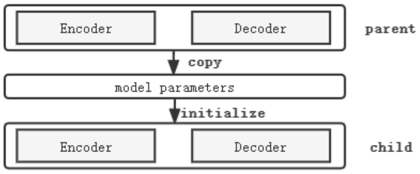


Fig. 1. Schematic diagram of transfer learning

For transfer learning, directly transferring the parameters of the parent model to the child model is not the optimal solution. Because the input language changes from the parent language to the child language, it is equivalent to introducing a completely different data space. The migrated model parameters cannot quickly adapt to the data space of the new language, and the translation effect will become worse. Previous studies have shown that the translation effect of transfer learning is closely related to the correct alignment of word vectors. The higher the alignment, the better the transfer effect.

2.3 Pre-train Techniques

In recent years, unsupervised pre-train of large neural models has recently completely changed natural language processing technology. The most representative model is BERT [9].Generally, there are two methods to use BERT's feature, fine-tune and feature.For the fine-tune method, a simple classification layer is added to the pre-trained model, and all parameters of downstream tasks are jointly fine-tuned, while the feature method keeps the pre-trained parameters unchanged. In most cases, the performance of the fine-tune method is better than feature method.

The basic steps of the tuning method in NMT scenarios: train the language model on a large number of unlabeled text data, then initialize the NMT encoder with the pre-trained language model, and use a marked data set for tuning. However, this process may lead to catastrophic forgetting. After fine-tuning, the model performance on the language modeling task will be significantly reduced. This may hinder the ability of the model to use pre-trained knowledge. To solve this problem, we introduce a distillation method to improve the model.

3 Methods

3.1 Word Alignment Under Hot-start

The biggest challenge of cross language transfer is vocabulary mismatch. When we replace only one source language, the NMT encoder will see a completely different input sequence. The pre-trained encoder weight does not match with the source embedding. Therefore, when we want to reuse the parent model parameters to train child language pairs, we need to solve the vocabulary mismatch between the transfer subject and the transfer object. However, the cold-start method is not applicable to the two languages that have nothing to do with subwords. Therefore, this paper uses the hot-start method to solve this problem. Before training, a Cross-lingual word embedding alignment method is used to match the words of the subject and object and align them correctly.

In this work, we use the method proposed by Patra et al. [10] and integrate it with transfer learning. Before model training, by embedding and aligning the words of the two languages, the parent model can recognize the transfer of child language pairs during training, so that the parameter migration can quickly adapt to the data distribution of the new language, which is impossible for the cold-start method.

Set $X = \{x_1, x_2 ... x_n\}$ and $Y = \{y_1, y_2 ... y_m\}$, They are two groups of word embedding from the source language and the target language respectively. Then set $S = \left\{(x_s^1, y_s^1) ... (x_s^k, y_s^k)\right\}$,$S$ represent the word embedding that has been bilingual aligned. We combine unsupervised distribution matching, alignment of known word pairs and weak orthogonal constraints to learn the linear mapping matrix W that maps X to Y.

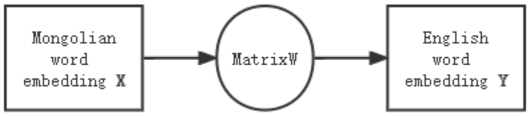


Fig. 2. Cross-lingual word embedding mapping from child language to parent language.

Unsupervised method: Given X and Y, the objective of unsupervised loss is to match the distributions of these two embedding spaces. We used an adversarial distribution matching target, similar to the work of Conneau et al. [11]. Specifically, a source to target mapping matrix W is learned to trick a discriminator D, which is trained to distinguish between WX and Y. We parameterize our discriminator with MLP, or optimize the mapping matrix and discriminator with corresponding objectives:

$$L_{D|W} = -\frac{1}{n} \sum_{x_i \in X} \log\left(1 - D(WX_i)\right) - \frac{1}{m} \sum_{x_i \in Y} \log D(X_i) \tag{2}$$

$$L_{W|D} = -\frac{1}{n} \sum_{x_i \in X} \log D(Wx_i) \tag{3}$$

Aligning Known Word Pairs: Given aligned bilingual word embeddings S.Our task is to minimize a similarity function(f_s) and maximize the similarity between the corresponding matched word pairs. Specifically, loss is defined as:

$$L_{W|S} = -\frac{1}{|S|} \sum_{(x_i^s, y_i^s) \in S} f_s(Wx_i^s, y_i^s) \tag{4}$$

Weak orthogonal constraint: Given an embedding space X, Patra et al. define a consistency loss that maximizes a similarity function f_a between x and W^TWx , $x \in X$. [10] This cyclic consistency loss $L_{W|O}$ encourages orthogonality of the W matrix based on the joint optimization:

$$L_{W|O} = -\frac{1}{|X|} \sum_{x_i \in X} f_a(x_i, W^T W x_i) \tag{5}$$

The above loss terms are used in conjunction with supervised and unsupervised losses, allowing the model to adjust the trade-off between orthogonality and accuracy based on joint optimization. This is particularly useful in embedded spaces that do not conform to orthogonality constraints.The final loss function of the mapping matrix is:

$$L = L_{W|D} + L_{W|S} + L_{W|O} \tag{6}$$

It enables the model to utilize the available distribution information from the two embedded spaces, so as to use all available monolingual data. On the other hand, it allows the correct alignment of tag pairs in the form of a small seed dictionary. Finally, orthogonality is encouraged. We can think of and as opposed to each other. Co optimization of the two helps the model strike a balance between them.

3.2 Approximate Distillation

The transfer learning initializes the child model with the trained parent model parameters, and then fine-tune the new training set.Since the new training set is generally a low-resource language and the corpus is relatively small, it may not be able to fully learn the source language knowledge of the sub language pairs.The commonly used auxiliary method adopts the pre-train model to learn the source language knowledge, and then initializes the knowledge to NMT for fine tune. Moreover, the use of fine-tune of large pre-train model will reduce the speed of NMT [12]. In this regard, we use distillation method to integrate the knowledge obtained from the pre-train model into the NMT encoder, retain the previous knowledge, and improve the language representation ability of the encoder.

As shown in the figure, first use the pre-train language model(PLM) to train the source language monolingual, and the trained knowledge is stored in the hidden layer in the form of matrix.Then, the hidden layer of the PLM is taken as the teacher [13], and the hidden layer state of the translation model encoder is taken as the student for knowledge fusion. (The PLM and NMT encoder are essentially language models, so it is reasonable to integrate the knowledge of the two language models.)

$$L_{ad} = -\left\|\hat{h}^{lm} - h_l\right\|_2^2 \tag{7}$$

L_{ad} is the mean square error loss of the two hidden layer states, $\hat{h}^{lm}$ is the state of the last hidden layer of the PLM, and h_l is the state of the lth hidden layer of the encoder.By punishing the loss of mean square error between the PLM and the state of the hidden layer of NMT encoder, the states of the two hidden layers are gradually

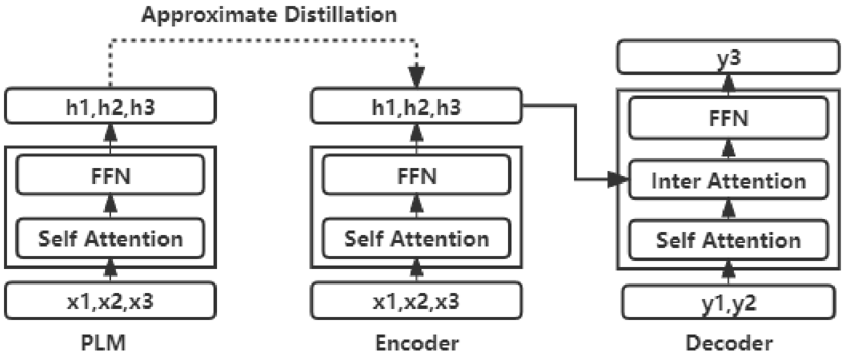


Fig. 3. The frame of approximate distillation.

close.In the experiment, the hidden state of the PLM is frozen, and the last layer (top layer) of the encoder is set to h_l. We find that adding the supervision signal to the top encoder layer is the best.During training, distillation loss can be used together with traditional cross entropy loss:

$$L_{all} = \alpha \cdot L_{nmt} + (1 - \alpha) \cdot L_{ad} \tag{8}$$

L_{nmt} is the cross entropy loss of the translation model. In the above formula, the total loss L_{all} combines L_{nmt} and L_{ad}, and α is a hyperparameter to balance the translation preference of the translation system [14]. In this way, the knowledge of PLM and the NMT can be combined to make better use of the pre-trained knowledge, so that the NMT encoder can obtain stronger representational capacity and generalization ability.

4 Experiment

4.1 Settings

We conducted experiments on English-Chinese (en-zh) and a low-resource translation task (mo-zh). For the en-zh task, the train set consists of 2 million bilingual sentences from the casic2015 corpus. We use NIST02 as the validation set and nist03-06 as the test set. For low-resource tasks, the dataset is provided by ccmt2019, as shown in Table 2. Mongolian monolingual comes from Wikipedia and news, with a total of 700m words.

Table 1. Dataset distribution

language	Train set	Validation set	Test set
mn-zh	256,754	2,000	2,000

All NMT models in our experiments follow the basic 6-layer transformer architecture of Vaswani et al. [1].Each source language adopts byte pair encoding [15], 30K merge operation, while the target language adopts 50k bpe merge encoding. The training was conducted using sockeye [16] and Adam optimizer with default parameters. The maximum sentence length is set to 100 and the batch size is set to 4,096 words. When the confusion on a verification set does not improve on the 12 checkpoints, the training is stopped.We set the checkpoint frequency of the parent model to 10,000 updates and the child model to 1,000 updates.The teacher model of knowledge distillation is trained by Bert, and the model in the experiment is $BERT_{base}$, which follows the structure proposed by Devlin et al. [9], l=12, h=768, a=12, total parameters=110m. Set the hyperparameter α to 0.5 during knowledge fusion.

We first train the collected Mongolian and English monolingual corpus into word embedding. In order to learn cross language mapping, we use a semi-supervised framework, and the parameters basically follow the settings of Patra et al. [10]. The unsupervised method uses muse, the data set is composed of Mongolian and English dictionaries in the corpus, and the weak supervised method uses a set of aligned word embedding. After learning the final mapping matrix, the words in the source language are mapped to the target space, and their nearest neighbors are selected as the final result according to the CSLS [11] distance. We compared it with the multilingual translation model. In multilingual training, we trained a single and shared NMT model [17]. For each subtask, we learned the joint BPE vocabulary of all source and target languages in the parent / subtask through 32K merge operations. The training data of subtasks are oversampled, so the proportion of parent / child training samples of each small batch is about 1:1.

4.2 Results And Analysis

Results: From table 2, in low-resource tasks, our method improved the scores of 3.2 and 1.7 BLEU respectively compared with traditional transformer and multilingual translation system.

Table 2. Comparison of experimental results.

System	Method	BLEU
Vaswani et al. [1]	Transformer base	27.4
Johnson et al. [17]	Multilingual	28.9
	+Transfer Learning(cold-start)	28.7
Ours	+Cross-lingual Word Embedding	29.5
	+Asymptotic Distillation	30.6

Analysis: In the first part of our experiment, we adopted the cold-start method of transfer learning, and directly transferred parameters without using sublanguage pairs.It is observed from the experiment that the cold-start method is also effective for low-resource languages, but it is less effective than the hot-start method using cross-lingual word embedding.It also further shows that the higher the degree of lexical matching between the subject and object of transfer, the better the effect of transfer learning.Finally, the approximate distillation method is added. Compared with the Transformer, it has a 3.2 BLEU improvement. We believe that the distillation method can enable the encoder of NMT to fuse additional information.

4.3 Ablation Test

In this section, we will further study our method in detail, compare it with their similar variants, and conduct general ablation studies.

Pre-trained word embedding type In Table 3, we analyze the cross-lingual impact of pre-trained embedding. We try not to transfer word embedding in transfer learning, but use pre-trained monolingual word embedding to replace the original word embedding.We observe that monolingual embedding without cross language mapping also improves transfer learning, but it is significantly worse than our proposed mapping to parent (en) embedding.You can also use learning mapping on the target (zh) side.Target mapping embedding is not compatible with the pre trained encoder, but directly guides the sub model to establish the connection between the new source and target.It also improves the system, but our method is still the best of the three embedding types.

Table 3. The experimental performance of different types of cross-lingual word embedding.

Pre-trained embedding	BLEU%
None	4.8
Monolingual	6.3
Cross-lingual (en-mo)	7.7
Cross-lingual (zh-mo)	7.2

Encoder vs Decoder As shown in Table 4, the effect of integrating the pre-trained knowledge into the encoder is good, but the effect is low in the decoder.Since BERT contains bidirectional information, the fusion of pre-trained knowledge into decoder may lead to inconsistency between training and reasoning.Gpt-2 uses limited self attention, where each token can only focus on its left context. Therefore, it is natural to introduce gpt-2 into the NMT decoder.It may be that the decoder is not a typical language model, it only contains information from the source language.

Vocabulary Size Table 5 shows the effect of different vocabulary sizes on translation. We changed the number of source side BPE merges and fixed the target vocabulary. The better result is to use 20K or 30K merges, which indicates that the vocabulary should be small in order to maximize the quality of translation. Fewer BPE merges result in more language independent tags. Cross-lingual embedding makes it easier to find overlaps in the shared semantic space. However, if the vocabulary is too small, we may lose too many language specific details necessary in the translation process.

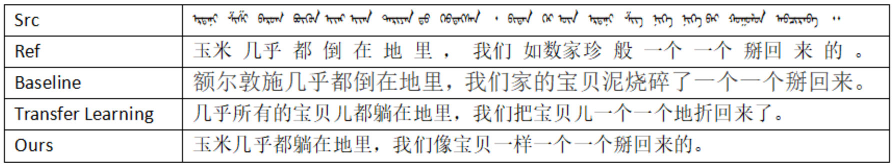
Table 4. Different Transformer modules and different PLM were used for approximate distillation ablation test.

PLM to module	BLEU
BERT to Transformer Encoder	29.5
BERT to Transformer Decoder	26.8
GPT-2 to Transformer Encoder	28.3
GPT-2 to Transformer Decoder	27.7

Table 5. Baseline translation results for different vocabulary sizes.

BPE merges	BLEU
20k	27.1
30k	27.4
40k	26.6
50k	26.3

4.4 Case Analysis



Src	[illegible]
Ref	玉米 几乎 都 倒 在 地 里 ， 我们 如数家珍 般 一个 一个 掰回 来 的 。
Baseline	额尔敦施几乎都倒在地里，我们家的宝贝泥烧碎了一个一个掰回来。
Transfer Learning	几乎所有的宝贝儿都躺在地里，我们把宝贝儿一个一个地折回来了。
Ours	玉米几乎都躺在地里，我们像宝贝一样一个一个掰回来的。

Fig. 4. Translation effects of different tasks.

It can be seen from the figure that the translation of this method basically conforms to the standard translation in terms of accuracy and fluency, so as to control the details of translation.In the case analysis, the words “玉米” and “宝贝” in Mongolian are very similar and easy to be confused.Translating these two words correctly makes the translation more accurate. And the words “几乎” and “掰” more reflect the fluency of language and express more accurately. It is proved that this method can improve the accuracy and fluency of translation.

5 Conclusion

The main contributions of this paper include: we propose a transfer learning framework based on hot-start. On the basis of transfer learning, we alleviates the problem of vocabulary mismatch between two languages without shared subwords.Meanwhile, in order to give full play to the role of the PLM and improve the generalization ability of the NMT encoder, we use the approximate distillation method to guide the NMT model to learn the output probability distribution of the PLM.In this way, the NMT model can master the knowledge probability distribution of the PLM and the NMT encoder at the same time. Experiments show that this method has a significant impact on low-resource translation tasks.

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
2. Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.
3. Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.
4. Benjamin Marie, Raphael Rubino, and Atsushi Fujita. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, 2020.
5. Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. In neural machine translation, what does transfer learning transfer? Association for Computational Linguistics, 2020.
6. Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*, 2018.
7. Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.
8. Markus Freitag and Yaser Al-Onaizan. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*, 2016.
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
10. Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. *arXiv preprint arXiv:1908.06625*, 2019.
11. Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.
12. Sergey Edunov, Alexei Baevski, and Michael Auli. Pre-trained language model representations for language generation. *arXiv preprint arXiv:1903.09722*, 2019.
13. Yichen Zhu and Yi Wang. Student customized knowledge distillation: Bridging the gap between student and teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5057–5066, 2021.
14. Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385, 2020.
15. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
16. Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*, 2017.

17. Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google‘‘s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.