Pivot based Transfer Learning for Neural Machine Translation: CFILT IITB @ WMT 2021 Triangular MT

Shivam Mhaskar, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai, India
{shivammhaskar, pb}@cse.iitb.ac.in

Abstract

In this paper, we discuss the various techniques that we used to implement the Russian-Chinese machine translation system for the Triangular MT task at WMT 2021. Neural Machine translation systems based on transformer architecture have an encoder-decoder architecture, which are trained end-to-end and require a large amount of parallel corpus to produce good quality translations. This is the reason why neural machine translation systems are referred to as *data hungry*. Such a large amount of parallel corpus is majorly available for language pairs which include English and not for non-English language pairs. This is a major problem in building neural machine translation systems for non-English language pairs. We try to utilize the resources of the English language to improve the translation of non-English language pairs. We use the pivot language, that is English, to leverage transfer learning to improve the quality of Russian-Chinese translation. Compared to the baseline transformer-based neural machine translation system, we observe that the pivot language-based transfer learning technique gives a higher BLEU score.

1 Introduction

The aim of this work is to improve the quality of Machine Translation (MT) for low-resource, distant and non-English language pairs. One of the major requirements for the good performance of the Neural Machine Translation (NMT) systems is the availability of a large parallel corpus. Such large parallel corpus of good quality is not available for low-resource, distant and non-English language pairs but mostly available for language pairs containing English. This poses a major challenge in developing good quality Machine Translation systems for non-English and distant language pairs. As a result there is a need to come up with additional resources by augmenting parallel corpora or by using knowledge from other tasks using transfer learning for translation of non-English language pairs. In this paper, we focus on leveraging the knowledge from other tasks using transfer learning to improve the performance of NMT systems for low resource language pairs.

In our pivot based transfer learning experiments we try to utilize the resources of English language, that is English-Chinese and English-Russian parallel corpora to improve the quality of Russian-Chinese translation. We implement techniques which efficiently use the resources of the English language for the task of Russian-Chinese translation.

2 Related Work

Recurrent Neural Network (RNN) based encoder decoder architectures (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014) were initially used in NMT systems. Transformer (Vaswani et al., 2017) architecture improved the performance of NMT systems. In order to enable translation between distant and non-English language pairs for which a large amount of parallel corpus is not available, a cascade method can be used. In the cascade method, two models are trained, a source language to English and a English to target language model. Then to translate a source sentence to target sentence, the source sentence is passed through the two models. (Zoph et al., 2016) introduced a transfer learning technique in which a parent model is trained on high resource language pairs, which is then used to initialize the the parameters of a child model which is then trained on low resource language pair data. (Kim et al., 2019) introduced pivot language-based transfer learning techniques in which the encoder and decoder of the model for low resource language pair is initialized using the encoder and decoder of different models trained on high resource language pairs, and this model is then finetuned on low resource language

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 336–340
November 10–11, 2021. ©2021 Association for Computational Linguistics

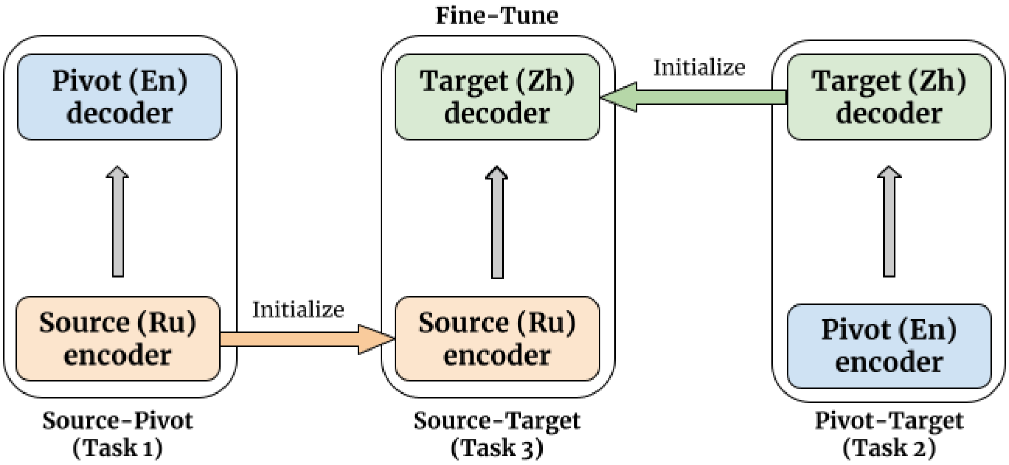

Figure 1: Direct Pivoting (En:English, Ru:Russian, Zh:Chinese)

pair data. Multi-lingual NMT systems (Zoph and Knight, 2016; Firat et al., 2016; Johnson et al., 2017) can also be used to improve the performance of low resource language pair translation as knowledge is transferred from various languages which helps the task of low resource language pair translation.

3 Approaches

In this section, we discuss the various approaches we used to build a Russian-Chinese MT system. We mainly focus on pivot-based transfer learning techniques, in which we use the resources of English to improve the quality of Russian-Chinese translation.

3.1 Baseline

The baseline Russian-Chinese model is a NMT model based on Transformer architecture. The model is trained on Russian-Chinese parallel data.

3.2 Cascade Model

The cascade model makes use of the resources of English language to train a Russian-Chinese MT system. In this approach, we train two NMT models, a source to pivot (Russian-English) model and a pivot to target (English-Chinese) model. The source Russian sentence is first translated into English using the Russian-English model. Then this English sentence is translated into Chinese using the English-Chinese model. In this way, the cascade model translates the Russian sentence to Chinese by passing it through the two NMT models. There are a few disadvantages in this cascade model based approach,

1. The source sentence is passed through two different NMT models to produce the target sentence. This doubles the decoding time for the generation of the output sentence which is very inefficient.

2. The errors in translation are propagated from first (source-pivot) model to the second (pivot-target) model.

These disadvantages of the cascade model approach make it an undesirable approach to utilize the resources of the pivot language. In order to overcome these disadvantages, we need to train a single source-target model which utilizes the resources of the pivot language. In the following pivot language-based transfer learning technique, direct pivoting, we overcome these disadvantages. In this technique, we train a single source-target model while utilizing the resources of the pivot language.

3.3 Direct Pivoting

In this technique, we first train two separate NMT models, a source-pivot model and a pivot-target model. As demonstrated in Figure 1, we first separately train a Russian-English (source-pivot) model (task 1) and a English-Chinese (pivot-target) model (task 2) on their respective parallel corpus. Then we use the encoder of the Russian-English (source-pivot) model and the decoder of the English-Chinese (pivot-target) model to initialize the encoder and decoder of the Russian-Chinese (source-target) model respectively. Finally, we fine-tune the Russian-Chinese (source-target) model on the Russian-Chinese parallel corpus.

As in this technique we are training a single source-target (Russian-Chinese) model, there is no problem of double decoding time. The parameters of the encoder and decoder of the source-target (Russian-Chinese) model are not randomly initialized, they are trained on the source-pivot and pivot-target translation task respectively. The initialized encoder and decoder of the source-target (Russian-Chinese) model have already learned some representation or knowledge from the previous tasks. This knowledge helps in the source-target (Russian-Chinese) translation task. In this way this approach utilizes the resources of the pivot (English) language which assists in the translation task from source to target (Russian-to-Chinese).

4 Experiments

In this section, we discuss the details of all the experiments that we carried out to implement the Russian-Chinese MT system.

4.1 Dataset

The NMT systems were trained on the parallel corpora provided by the WMT 2021 organizers. We used the Russian-Chinese, Russian-English, and the Chinese-English parallel corpus. We used a subset of the provided parallel corpora for training the models. Byte Pair Encoding (BPE) (Sennrich et al., 2015) is used as a segmentation technique. The words in the data are broken down into sub-words using the BPE technique. For the baseline model the number of BPE merge operations used were 16000 for the source and target data. For the direct pivoting model, the source and target vocabulary are combined English-Russian and English-Chinese vocabulary, respectively. So, the BPE codes are computed by combining the source side Russian and English data for source and the target side English and Chinese data for target. The number of BPE merge operations used were 32000 for the source and target data. The detailed corpora statistics are mentioned in Table 1.

Language pair	Number of sentences
Russian-Chinese	10M
Russian-English	10M
English-Chinese	10M

Table 1: Corpora statistics of all the language pairs

4.2 Models

For all the experiments, Transformer architecture was used. The encoder of the Transformer consisted of 6 encoder layers and 8 encoder attention heads. The encoder used embeddings of dimension 512. The decoder of the Transformer consisted of 6 decoder layers and 8 decoder attention heads. For the implementation of all models, fairseq (Ott et al., 2019) library was used.

4.3 Training Setup

For all experiments, the transformer model from fairseq library was used. The optimizer used was adam with betas (0.9, 0.98). The inverse square root learning rate scheduler was used with an initial learning rate of 5e-4 and 4000 warm-up updates. The criterion used was label smoothed cross entropy with label smoothing of 0.1. The dropout probability value used was 0.3 for all layers. For the baseline model, the size of source (Russian) and target (Chinese) vocabulary is 16876 and 29500, respectively. For the direct pivoting model, the size of source (combined Russian-English) and target (combined English-Chinese) vocabulary is 34020 and 47052, respectively. The best model for all the techniques was chosen by calculating the BLEU (Papineni et al., 2002) scores on the development set provided by the WMT 2021 organizers and the choosing the model with best BLEU score.

4.4 Baseline

The baseline model is a transformer model trained on Russian-Chinese (source-target) parallel corpus.

4.5 Cascade Model

The cascade model consists of two NMT models trained separately. The first model is a Russian-English model trained on Russian-English parallel corpus. The second model is a English-Chinese model trained on English-Chinese parallel corpus. For translating a Russian sentence to Chinese, the sentence is passed through two models.

4.6 Direct Pivoting

The direct pivoting model uses a shared vocabulary of Russian-English (source-pivot) on the encoder side and English-Chinese (pivot-target) on the decoder side. This is done to ensure that the

Model	BLEU score
Baseline	18.2
Cascade	17.2
Direct Pivoting	18.8

Table 2: BLEU scores of Russian-Chinese NMT system using different techniques

encoder and decoder parameters are transferable as transformers are fixed vocabulary models. The Russian-English (source-pivot) model is trained on Russian-English parallel data and the English-Chinese (pivot-target) model is trained on English-Chinese parallel data. Then the encoder of Russian-English model and decoder of English-Chinese model is used to initialize the encoder and decoder of Russian-Chinese (source-target) model. Finally, we fine-tune the Russian-Chinese model on the Russian-Chinese parallel data.

5 Results and Analysis

The evaluation of the models were performed on the basis of the BLEU scores. These BLEU scores were calculated and provided by the WMT 2021 organizers. The BLEU scores were calculated on a test set provided by WMT 2021 organizers, which consisted of 1751 sentences. Table 2 shows the BLEU scores of all the models. The baseline Russian-Chinese model produced a BLEU score of 18.2. The cascade model in which the Russian sentence is first translated to English using Russian-English model and then the English sentence is translated to Chinese using the English-Chinese model, produced a BLEU score of 17.2. The possible reason for this decrease in BLEU score is that the errors made by the Russian-English model are propagated to the English-Chinese model, which further introduced its own errors. As the source sentence is passed through the two model each model introduces its own errors, which decreases the BLEU score.

The direct pivoting model produced a BLEU score of 18.8 which improved the BLEU score by 0.6 points over the baseline model. This increase in BLEU score is because the encoder and decoder of the Russian-Chinese model are not randomly initialized; but they are initialized from the encoder and decoder of Russian-English and English-Chinese model respectively. Then the model is fine-tuned on Russian-Chinese parallel corpus. The encoder and decoder have already learnt some representations which helps in the task of Russian-Chinese translation. Also as this is a single NMT model, there is no problem of propagation of errors or double decoding time.

6 Conclusion and Future Work

In this work, we implement and compared pivot language-based transfer learning technique to improve the task of translation between non-English language pair, that is Russian-Chinese. We observe that pivot language-based transfer learning technique improves the BLEU score over the baseline model and is an efficient way to use the resources of the pivot language. We also observe that the pivot language-based transfer learning technique mitigates the problems of double decoding time and error propagation present in simple cascade-based models.

In future, we plan to explore various data augmentation techniques that can make use of the resources of the English language to augment data for the task of translation of non-English language pair translation. We also plan to use various language model pretraining techniques like Masked Sequence to Sequence Pre-training (MASS) to pretrain the encoder and decoder before using them for the downstream task of translation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

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

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.