A teacher-student model with the bilingual word embedding for zero-shot machine translation

Yijie Cai 2, a, Min Zhu 1 *

1 School of Data Science & Engineering, East China Normal University, Shanghai, China
2 School of Computer Science and Technology, East China Normal University, Shanghai, China
* Corresponding author’s e-mail: mzhu@cc.ecnu.edu.cn
a 51184506002@stu.ecnu.edu.cn

Abstract. Neural machine translation has achieved state-of-art performance with sufficient data, but it still suffers from the data scarcity problem for low-resource language pairs. Teacher-student model that transfers knowledge from a pivot → target model to a source → target model with the source → pivot parallel data has shown its effectiveness. In this paper, we improve the model with the bilingual word embedding and transformer architecture. Experiments are carried out on three commonly used translation datasets and the result shows the improvement over a baseline teacher-student model.

1. Introduction

Machine translation, which translates natural language into another language automatically, is an important task in natural language processing (NLP). Due to the increasing data, the data-driven method becomes the main method in the machine translation field. Statistical machine translation (SMT) is a traditional data-driven method that studies the translation process from the perspective of probability and statistics. The representative SMT like phrase-based [1] and parse-based [2] model face the problem that they work not well in low count event and can not capture all translation rules by limited manually designed features. With the rise of deep learning, Neural Machine Translation (NMT), which models the translation process in an end-to-end way, has achieved state-of-the-art performance in many translation tasks. [6-8] RNN-based or CNN-based neural network with attention mechanism proof that neural network can automatically capture semantic and syntax features from sentences, that overcomes the disadvantage of SMT. [3-5]

However, NMT has fluent and great results in rich language pairs, but these models underperform phrase-based statistical machine translation (PBSMT) or evenly unsupervised models in low-source conditions. [9] Therefore, exploring new methods to overcome the data inefficiency in NMT has become an important research direction.

A important method to deal with the low-resource problem for NMT is to introduce other rich language corpora. The multilingual translation model tries to fuse several language pairs in one model to relate low-source pairs including network architecture method and universal NMT model training framework. A multilingual model with shared attention is presented to achieve zero-resource translation. [10] Johnson integrates all language pairs into a single dataset with artificial marks which can indicate the language translated. With no change in architecture, the model can translate the pairs...
that is never seen in the dataset. [11] Compared to the multilingual model, pivot-based method tries to build a relationship between two languages without parallel data via a pivot, which is either text or image. [12-13] While the common pivot method divides the translation process into a two-part, Teacher-Student method models source to target translation directly through pseudo data from the pre-trained model, which avoids the error propagation problem. [14]

In this paper, we replace the RNN architecture with the transformer architecture for the teacher-student model. The shared decoder from teacher and student with the bilingual word embedding can further make the model easier to transfer knowledge.

2. Fundamental definition
Neural machine translation captures semantic and syntax features of sentences by neural networks. Compared to the traditional machine translation, it predicts the next target word in an end-to-end way. As shown in figure 1, the basic model consists of an encoder and a decoder. The encoder reads the sentence word by word and encodes the information about the sentence to the decoder. The decoder generates the target word according to the previously generated words and the encoder information. Let \( x \) be a source sentence and \( y \) be a target sentence. The whole model can be denoted as \( P(y|x; \theta) \), where the \( \theta \) is the parameter of the network. Given a source↔target parallel corpus D, the log-likelihood loss function can be written as:

\[
L(\theta) = \sum_{x,y} \log P (y|x; \theta) = \sum_{x,y} \sum_{i} \log P (y_{i}|x, y_{1}, ..., y_{i-1}; \theta)
\]  

Then, the model parameter can be learned by maximizing log-likelihood function:

\[
\theta^* = \arg \max \sum_{x,y} \log P (y|x; \theta)
\]

3. Proposed approach
3.1. Teacher-student framework
Although NMT achieves success in rich-resource language pairs, it still performs very poorly in low-resource or zero-resource translation scenarios. One existing pivot-based method, teacher-student modelling (TSM), has shown its effectiveness in the zero-shot scenario [14] where there is a lot of source↔pivot and pivot↔target bilingual data but no source↔target parallel data. First, TSM trains a common NMT (teacher model) with pivot-to-target data, which translates pivot language into target language. Through the teacher model, TSM generates a lot of pseudo data(source-to-target) with
source-to-pivot data. The generating process method is inspired by the assumption that if two sentences in different languages have the same meaning, their translations in other languages are the same. Then, TSM directly trains the result NMT (student model) with the pseudo data. In order to further improve the performance of the student model, TSM uses the soft labels of the teacher model to guide the learning of the student model, which is based on a new assumption:

If a source sentence $x$ is a translation of a pivot sentence $z$, then the probability of generating a target word $y$ from $x$ should be close to that from its counterpart $y$, given the already obtained partial translation $y_{j-1}$.

Let $P(y|x, y_{j-1}; \theta_{xy})$ be the result probability of student model and let $P(y|z, y_{j-1}; \theta^{*}_{zy})$ be the result probability of teacher model. $\theta_{xy}$ is the parameter of the student model, and $\theta^{*}_{zy}$ is the parameter of the teacher model which is fixed. According to the assumption, these two probabilities are close so the KL divergence of these probabilities is close to zero:

$$KL(P(y|x, y_{j-1}; \theta_{xy}) || P(y|z, y_{j-1}; \theta^{*}_{zy})) \approx 0 \quad (3)$$

Then, the training objective can be defined by the total average value of KL divergence on the whole dataset:

$$L(\theta_{xy}) = \sum_{x,z,y} E_p(y|z) J(x, y, z; \theta_{xy}, \theta^{*}_{zy}) \quad (4)$$

3.2. Shared transformer architecture

The architecture of the original model is RNN-based which is time-consuming in the training process. The transformer architecture has fast training speed and can capture the long dependence between two words. Due to the two advantages, we replace the RNN architecture with the transformer architecture. Figure 2 shows our model architecture. We share the decoder between the teacher and student model to transfer knowledge. The parameters from the decoder of the teacher model are the knowledge to translate from the encoding information. Despite the differences in the source language, the shared parameters may be beneficial to the student model.

(a)The RNN-based architecture               (b)The transformer architecture with shared decoder

Fig.2 A comparison between two model architecture. In the shared architecture, the purple block is the shared decoder between the two models. The blue block is the encoder of the teacher model and the red one is the encoder of the student model. In the original architecture, the blue block is the teacher model and the red one is the student model. Both of them are whole translation models.
3.3. Bilingual word embedding

The bilingual word embedding technique [15] is applied to the process of initializing word vectors. In the original method, the teacher and student model have independent word space. The different word representation makes encoding information hard to keep consistency. The bilingual word embedding technique reconstructs the initial word vectors in the student model. They are in the same coordinate system with vectors from the teacher model.

4. Experiment and result

In this section, the proposed approach is evaluated on three language pairs. Each language pair has the same pivot language which is English. The training and test data are chosen from Europarl v7 and IWSLT. Table 1 shows the size of the dataset. The property represents the use of the dataset. In pivot property, the data is used to train the teacher model for the student models in the same corpus. The student models are trained with data in source property.

Several baselines are compared to our approach:

1. Direct bridge way: this method uses pivot language to train source→pivot and pivot→target translation models. The source sentences are translated into pivot sentences with source→pivot model. The pivot→target model translates the pivot sentences into target sentences.

2. Original teacher-student model: In order to verify our approach, we choose the original model as our baseline.

In the training process, tensor2tensor which is a transformer-based toolkit is used to build our experiment. The number of the layers is set to six both in encoder and decoder. The dimension of all hidden layers is 512 and the number of the multi-head is 8. For the regularization, the dropout with different probability is employed for word embedding, attention mechanism and ReLU activation process. The layer post-process and preprocess are set between two layers. The post-process contains dropout and residual structure and the preprocess contains layer normalization. In the pre-trained model, the input embedding and output embedding are shared. The training batch token was set to 1024 and the training optimizer was Adam with parameters $\beta_1 = 0.9$, $\beta_2 = 0.9$ and $\epsilon = 10^{-9}$. The initial value of the learning rate is set to 0.1. The schedule of the learning rate consists of warm-up and decay. The evaluation metric is case-insensitive.

Table 2 shows the experiment results. For simplicity, we denote the direct bridge way as Bridge-RNN and Bridge-TR. The two different models represent RNN and transformer architecture for the same method. Similarly, the teacher-student method is denoted to TS-RNN and TS-TR. Bilingual word embedding is also applied to the TS-TR model to verify the performance of our shared model. The TS-TR performs better than TS-RNN by an average of 2.485 BLEU. Bridge-TR also has better performance than Bridge-RNN. The two results indicate that the transformer may be appropriate for the pivot method. From table 2, it is found that the teacher-student method has better performance with the same architecture. That means the one-step translating process may avoid the error propagation in the two-step translating process like direct bridge way. In the last row of table 2, it is found that the TS-TR model with the bilingual word embedding has the best result than all baselines. Word embedding in the same space lets the encoders from two models output similar information to the shared decoder. It shows that the shared structure in decoder with word embedding initializing can improve the transferring performance in the teacher-student framework.

Table 1. Parallel training data statistics

| corpus     | language | train | dev  | test | property |
|------------|----------|-------|------|------|----------|
| Europarl v7 | ES-EN    | 840K  | 2000 | 2000 | source   |
| Europarl v7 | DE-EN    | 960K  | 2000 | 2000 | source   |
Table 2. BLEU score on several language pairs

| method   | ES-FR | DE-FR | RO-FR |
|----------|-------|-------|-------|
| Bridge-RNN | 29.79 | 23.7  | 27.64 |
| TS-RNN    | 33.86 | 27.03 | 28.34 |
| Bridge-TR | 34.28 | 28.22 | 29.32 |
| TS-TR     | 36.13 | 29.73 | 29.67 |
| TS-TR-BWE | 37.08 | 30.22 | 29.98 |

5. conclusion
In this paper, we study the teacher-student framework in zero-shot machine translation. In order to improve the original model, we adopt a new architecture and introduce an effective initializing technique. The experiment on three language pairs shows that the shared transformer architecture with the bilingual word embedding achieves high performance in the zero-shot translation scenario. However, the proposed approach is not suitable in very remote language pairs like Chinese to French. In future work, we will explore new techniques to expand the application scope of our algorithm.

References
[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 48–54. Association for Computational Linguistics, 2003.
[2] David Chiang, Kevin Knight, and Wei Wang. 11,001 new features for statistical machine translation. In Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics, pages 218–226. Association for Computational Linguistics, 2009.
[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
[4] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1243–1252. JMLR. org, 2017.
[5] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. Advances in NIPS, 2014.
[6] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025, 2015.
[7] Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. arXiv preprint arXiv:1512.02433, 2015.
[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.
[9] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. arXiv
preprint arXiv:1706.03872, 2017.

[10] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. arXiv preprint arXiv:1601.01073, 2016.

[11] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viegas, 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.

[12] Yun Chen, Yang Liu, and Victor OK Li. Zero-resource neural machine translation with multi-agent communication game. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[13] Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. Neural machine translation with pivot languages. arXiv preprint arXiv:1611.04928, 2016.

[14] Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. A teacher-student framework for zero-resource neural machine translation. arXiv preprint arXiv:1705.00753, 2017.

[15] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Herve Jégou. Word translation without parallel data. arXiv preprint arXiv:1710.04087, 2017.