11-731 Assignment 2 Report
Low Resource Machine Translation

Chen Xia
chenxia@cs.cmu.edu
Siyuan Wang
siyuanw@cs.cmu.edu

November 10, 2018

Abstract

This paper uses the Ted Talk LRL(low resource languages), Azerbaijani, Belarus, Galician, to translate into English. Starting from the baseline model from assignment 1, and using the following technique to train the model: (1)Bi-source Modeling (2)Pre-trained Multilingual Embeddings (3)Byte-Pair Encoding (4)Back Translation. By using these techniques, we are able to improve our az-en baseline by 4.6 BLEU score, increase our be-en baseline up to 4 BLEU score, and outperform our gl-en baseline by 3.8 BLEU score.

1 Methods

1.1 Baseline Model

Our baseline model is an attentional seq2seq model. The encoder is simply one-layer bidirectional LSTM with 256-dimensional word embeddings and 256-dimensional hidden state. The decoder model is an one-layer LSTM with 256-dimensional word embeddings and 256-dimensional hidden state, as well as input feeding, scheduled sampling and label smoothing.

1.2 Bi-source Modeling

Bi-source modeling method trains an NMT model with two source languages: one primary low-resource language (LRL) of interest, and a highly related high-resource language (HRL) as helper. The intuition of this method is to allow the LRL to learn a better language model from a similar language with a larger corpus. It can also be seen as a form of regularization to avoid overfitting on the limited training corpus of LRL.

This method can be implemented by simply concatenating the training data from the two corpora. However, the LRL corpus can be very small compared to the HRL one, and the model might be overwhelmed by the HRL data. A slight variant is to duplicate the LRL corpus a few times, so that the LRL-HRL ratio can be somewhat balanced.

1.3 Pre-trained Multilingual Embeddings

Pre-trained word embeddings have proven to be invaluable for improving performance in low-resource machine translation task ([3], [5]). MUSE[1] is a library for Multilingual Unsupervised or Supervised word Embeddings. We use MUSE to obtain pre-trained word embeddings trained on Wikipedia corpus. The pre-trained embeddings are used to initialize word embeddings in our NMT

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

model.

We first use monolingual word embedding based on Wiki data.Because we do not have az-en bilingual dictionary.We use unsupervised adversarial training ([2]). From this script, we get the multilingual embedding, and thus, map the source language embedding and target language embedding to the same space. The measurement of how close two embeddings are is using cosine distance.From the closes mapping, we could both get the unsupervised trained mapping dictionary and also multilingual word embeddings. Thus we hope the pretrained multilingual embeddings would give some word alignment information. And we first match the word in embedding matrix and if we did not find them, we then lowercase them. But because of unsupervised training and extremely low resource az language, we still have a lot of missing word. This makes the result even worse.

We then try to co-train az/tr-en, because MUSE already has the trained multilingual embedding between tr and en. So we randomly initialized az embedding, and use tr pretrained multilingual embedding. This just give a little improvement in bleu score.

1.4 Byte-Pair Encoding

NMT typically works with a fixed vocabulary. The translation of out-of-vocabulary words is tricky issue in NMT. [7] introduces a simple yet effective approach, making the NMT model capable of open-vocabulary translation by encoding rare and unknown words as sequences of subword units. This is based on the intuition that various word classes are translatable via smaller units than words. It adapts byte pair encoding (BPE) [1], a compression algorithm, to the task of word segmentation.

Specifically, we used SentencePiece [2] toolkit by Google to pre-process the corpus. SentencePiece is an unsupervised text tokenizer and detokenizer mainly for Neural Network-based text generation systems where the vocabulary size is predetermined prior to the neural model training, which implements subword units. With the BPE'ed subwords as new vocabulary, the vocabulary size could be greatly reduced, and hence result in better training efficiency and hopefully better BLEU score.

1.5 Back Translation

One important thought direction of improving LRL NMT performance is data augmentation. [6] pairs monolingual target language training data with an automatic back-translation, which can be treated as additional parallel training data. They show substantial improvements on en-GE as well as low-resource language pairs. Following this idea, we can reverse the LRL parallel data and train a back-translation model first, and then generate additional low-quality LRL parallel data by translating from the large monolingual target language corpus. The back-translated source language sentences are less reliable than the given ground truth data, but might help improve the accuracy in a low-resource setting.

2 Experiments

2.1 Experimental Setup

Dataset We use the TED talks dataset [3] provided. It consists of 6 languages to English parallel corpora, among which Azerbaijani (az), Belarusian(be), Galician(gl) are the low-resource languages of interest, and Turkish(tr), Russian(ru), Portuguese(pt) are corresponding high-resource helper languages.

Baseline Model and Bi-source We first perform experiments using the model architecture from Assignment 1 as our baseline on Azerbaijani/az, with single source and bi-source(+Turkish/tr) settings.

Pre-trained Multilingual Embeddings Unsupervised training: use az and en wiki pretrained monolingual word embedding, and use adversarial training to map them into same space. This gives us two multilingual embeddings and a best mapping dictionary. Supervised training: use az/tr-en

[2] https://github.com/google/sentencepiece

[3] https://github.com/neulab/word-embeddings-for-nmt

Baseline		Multi Pre-trained Embeddings		BPE az/tr-en		
az-en	az/tr-en	az-en	az/tr-en	src 16k	src 16k / tgt 8k	src&tgt 16k
2.5	5.1	1.5	3.0	**7.1**	5.1	6.3

Table 1: BLEU for baseline model (az-en, az/tr-en); multilingual pre-trained embeddings (az-en, az/tr-en); and BPE on az/tr-en (encode src only to 16k, src to 16k and tgt to 8k separately, src and tgt together to 16k)

Strategy	az/tr-en	be/ru-en	gl/pt-en
BPE-src-16k	**7.1**	7.4	18.5
BPE-src-10k + Back-Translation	6.4	**11.7**	**22.3**
Given Baseline	7.1	11.6	22.0

Table 2: BLEU for BPE on src using 16k size, and BPE on src using 10k size with back translation

embeddings, and az is randomly initialized(also could use monolingual embedding to initialize), use pretrained tr-en multilingual embeddings.

Byte-Pair Encoding To understand the best way to apply BPE for our NMT model, we experiment with the following three setups on az/tr-en corpus: 1. BPE src only (size 16000); 2. BPE src/tgt separately (src size 16000, tgt size 8000); 3. BPE src/tgt together (size 16000). Also, we run setup 1 (BPE src only) on all three language pairs (az/tr-en, be/ru-en, gl/pt-en).

Back Translation To do back translation for each language pair, we first train a statistical machine translation system using Moses [4] on en-az, en-be, en-gl parallel corpora separately. Then we use the trained SMT model to translate from English sentences in the whole corpora except the ones with the target language and its helper language to avoid contaminating high-quality ground truth data. Finally, the amount of LRL in the train corpus is about the same as or slightly more than the amount of helper HRL. With the augmented train corpus, we train with almost the same setup as the previous experiment, except BPE size adjusted to 10000 due to time constraint.

2.2 Experimental Results

Baseline Model and Bi-source As shown in Table 1, the baseline model is the model from last assignment. The setting is encoder-decoder model with dot-product attention. Also, we use label smoothing, schedule sampling and input feeding. It achieves 2.5 BLEU score in az-en translation, which is lower than given baseline. And we decide to focus on more techniques on low resource translation to benefit more from this assignment, so we do not pay much attention to tune our vanilla model. Then we concatenate az and tr, and co-train them together. Finally we get 5.1 BLEU score on az-en translation.

Pre-trained Multilingual Embeddings As shown in Table 1, the multilingual embeddings experiment starts with unsupervise-trained az-en monolingual embeddings. It achieves only 1.50 bleu score due to large missing vocabulary. Then we use supervised az/tr-en and it increases 0.50 BLEU score. And we also deal with missing vocabulary by lowercase them. Then it gives us 1 more point BLEU score. We finally get 3.0 BLEU score on this experiment.

Byte-Pair Encoding and Back Translation Three BPE setups are performed on az/tr-en to figure out the best configuration. As shown in Table 1, doing BPE only on source language reports the best BLEU score. This is probably because BPE is more suitable for sparse languages with more low-freq words. Specifically, in this train parallel corpus, az+tr has 102982 words with frequency >= 2, whereas en has 39617. The English vocabulary is small enough to keep the original words as decoder output with a decent accuracy. However, we cannot completely rule out the potential influence factor of BPE size. Unfortunately, due to time and resource constraint, we were not able to perform extensive experiments on BPE size.

[4] http://www.statmt.org/moses/

As for the back-translation experiment, we realized a smaller BPE size might perform better according to [4], so we changed BPE size to 10k at the same time (unfortunately due to time constraint). As shown in Table 2, we achieve significant improvement on BLEU for be and gl. We hypothesize there are three reasons for this improvement. First, smaller BPE size on source languages helps. Second, with more LRL produced by back translation, the model can learn a better encoder more suitable for the LRL language itself. Last but not least, as we train a single BPE model on the concatenated Bi-source corpus, previously the LRL only takes a very small percentage as compared to the helper HRL. Now with more LRL enabled by back translation, the BPE vocabulary could be more balanced. As for the slight decrease on az, this might be due to the low quality of the SMT system for en-az, which only achivies 3.1 BLEU. Hence, it also shows that for back translation to help, the ground truth parallel corpus should have a reasonable size and achieve a reasonable back translation system in the first place.

3 Conclusion

Despite a lower baseline due to legacy reasons from last assignment, we are able to achieve up to 4.6 BLEU improvement with byte-pair encoding (BPE) and back translation techniques.

From our point of view, the main reason of not that desired performance of BPE is due to the BPE vocabulary size. If we truncate our vocabulary size from 16000 to 10000, the performance would be better.

In conclusion, BPE helps the NMT system by reducing the vocabulary size and segmenting rare words to subword units. Also, back translation gives us much better performance because it's kind of data augmentation technique and it enables our models to generalize.

References

[1] P. GAGE, *A new algorithm for data compression*, The C Users Journal, 12 (1994), pp. 23–38.

[2] G. LAMPLE, A. CONNEAU, L. DENOYER, AND M. RANZATO, *Unsupervised machine translation using monolingual corpora only*, arXiv preprint arXiv:1711.00043, (2017).

[3] M. NEISHI, J. SAKUMA, S. TOHDA, S. ISHIWATARI, N. YOSHINAGA, AND M. TOYODA, *A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size*, in Proceedings of the 4th Workshop on Asian Translation (WAT2017), 2017, pp. 99–109.

[4] G. NEUBIG AND J. HU, *Rapid adaptation of neural machine translation to new languages*, arXiv preprint arXiv:1808.04189, (2018).

[5] Y. QI, D. S. SACHAN, M. FELIX, S. J. PADMANABHAN, AND G. NEUBIG, *When and why are pre-trained word embeddings useful for neural machine translation?*, arXiv preprint arXiv:1804.06323, (2018).

[6] R. SENNRICH, B. HADDOW, AND A. BIRCH, *Improving neural machine translation models with monolingual data*, arXiv preprint arXiv:1511.06709, (2015).

[7] ——, *Neural machine translation of rare words with subword units*, arXiv preprint arXiv:1508.07909, (2015).