Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation

Toan Q. Nguyen & David Chiang

{tnguye28,dchiang}@nd.edu

Computer Science and Engineering, University of Notre Dame

Introduction

The transfer learning approach by Zoph et al. (2016) is a simple yet effective method to improve Neural Machine Translation (NMT) performance on low-resource languages.

1. Train a model on a high-resource language pair
2. Use it to initialize training a child model on a low-resource language pair, possibly unrelated to the parent one
3. Significantly improve performance on Hausa-, Turkish-, and Uzbek-English using French-English as parent

In this work, we explore the opposite scenario:

1. Both parent and child language pairs are low-resource, but related
2. They might share some meaningful lexical similarities

We propose to apply an additional preprocessing step on top of their method:

1. Use Byte-Pair-Encoding (BPE) to learn subword-level representation of data to find common subwords between languages
2. We can now transfer useful word embedding information from parent to child instead of random assignment as is done in Zoph et al.'s

Experiments on three Turkic languages (Uzbek, Turkish, Uyghur) show that while their method helps only slightly, ours significantly improves NMT performance by up to 4.3 BLEU.

Data

Uzbek-English as parent model and Turkish-English/Uyghur-English as child. Data from the LORELEI project.

Our systems

For each parent-child pair:

1. Uyghur is written in Arabic so we transliterate it to Latin before learning the BPE rules (https://cis.temple.edu/~anwar/code/latin2uyghur.html)
2. Learn BPE rules from the union of source and target data of both parent and child to:
   a. Extract the common subwords between source languages
   b. Ensure consistent word splits between source and target languages
3. 2-layer, 512-hidden-unit global attention model with general scoring function and input feeding by Luong et al. (2015). Drop out rate is 0.2 for every system.
4. Number of BPE operations ranges from 5k-60k. Best configuration is chosen according to BLEU score on dev set.
5. Beam search with length normalization (Wu et al. (2016)) on dev and test sets

Baselines

1. Word-based NMT with/without transfer learning
2. BPE-based NMT
3. Same model sizes, training procedures…
4. Vocabulary size:
   a. Word-based:
      - 10k-50k for Turkish-English
      - 5k-20k for Uyghur-English
   b. BPE-based: full vocabulary since it's smaller. To deal with UNK:
      - Train on another copy of data with rare words replaced by UNK
      - Number of epochs is halved accordingly

Results

		baseline		transfer		transfer+freeze	
		BLEU	size	BLEU	size	BLEU	size
Tur-Eng	word-based	8.1	30k	8.5*	30k	8.6*	30k
	BPE	12.4	10k	13.2†	20k	—	—
Uyg-Eng	word-based	8.5	15k	10.6†	15k	8.8*	15k
	BPE	11.1	10k	15.4‡	8k	—	—

Table 1: Whereas transfer learning at word-level does not always help, our method consistently yields a significant improvement over the stronger BPE systems. Scores are case-sensitive **test** BLEU. Key: `size` = vocabulary size (word-based) or number of BPE operations (BPE). The symbols † and ‡ indicate statistically significant improvements with $p < 0.05$ and $p < 0.01$, respectively, while $*$ indicates a statistically insignificant improvement ($p > 0.05$).

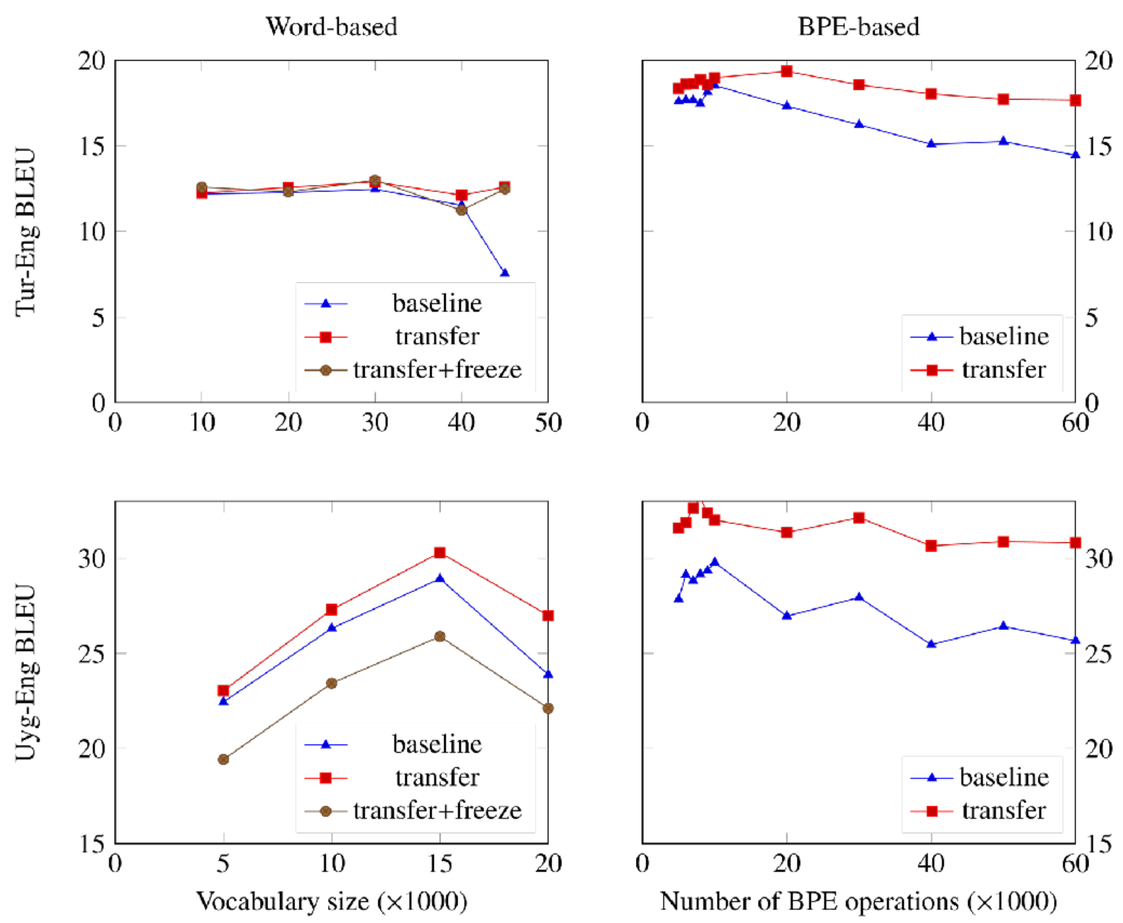

Figure 1: Tokenized **dev** BLEU scores for various settings as a function of the number of word/subword types. Key: baseline = train child model only; transfer = train parent, then child model; +freeze = freeze target word embeddings in child model.

Conclusions

In this paper, we have shown that the transfer learning method of Zoph et al. (2016), while appealing, might not always work in a low-resource context. However, by combining it with BPE, we can improve NMT performance on a low-resource language pair by exploiting its lexical similarity with another related, low-resource language. Our results show consistent improvement in two Turkic languages. Our approach, which relies on segmenting words into subwords, seems well suited to agglutinative languages; further investigation would be needed to confirm whether our method works on other types of languages.

References

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Proc. EMNLP.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Proc. EMNLP.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.

Acknowledgments

This research was supported in part by University of Southern California subcontract 67108176 under DARPA contract HR0011-15-C-0115. Nguyen was supported by a fellowship from the Vietnam Education Foundation. We would like to express our great appreciation to Dr. Sharon Hu for letting us use her group's GPU cluster (supported by NSF award 1629914), and to NVIDIA corporation for the donation of a Titan X GPU.

UNIVERSITY OF NOTRE DAME
College of Engineering