Team MICA Project Proposal: Neural Machine Translation Implementation and Comparison

Shasha Lin (sl4964), Akash Kadel (ak6201), Eduardo Fierro (eff254), Millie Dwyer (mmd378)

Project Objective

The encoder-decoder model with attention has been proven effective in neural machine translation. Many modifications have been made to the vanilla word-based encode-decoder model, such as segmenting words into subwords via byte-pair encoding (BPE) [1], and post-processing out-of-vocabulary words through word-alignment algorithm (Luong et al. [2]). Chung et al. [3] also showed that character-level decode without explicit segmentation achieved superior performance compared to subword-level decoder.

For our project, we aim to approach the neural machine translation task by implementing the BPE attention encoder-decoder, as well as Chung et al.'s character-level decoder, and compare their performances on the same dataset (either the WMT [4] dataset used by the papers, or Canadian Parliamentary transcripts [7]).

Models to compare

I: Encoder-Decoder With Attention & BPE

The attention mechanism uses weighted average of context vectors surrounding a word to compute the memory state of decoder, allowing variable-length context information to be learned by the model. BPE allows the model to learn subword units that are useful for learning cognates, loanwords, and morphologically complex, which improves the translation of OOV words or in-vocabulary rare words [5]. Our first model will replicate Sennrich et al.'s model for the WMT 16 translation task by building an attention-based encoder-decoder model with BPE, using subwords as units for language [1].

II: Character-level Decoder

For comparison with BPE, we are going to reproduce Chung et al.'s character-level decoder [3]. Using a character-level decoder addresses problems very similar to those that motivated the proposal of BPE: using character as units does not impose any pre-defined segmentations for a sequence, and allows the model to learn morphemes and lexemes, which generates to rare words. In this paper, Chung et al. also implemented a bi-scale stacked decoder with a faster and a slower layer, respectively, which empirically does not make a big improvement from the regular character-level decoder. If we have time/resources, we will optionally implement the bi-scale decoder.

III: Transformer*

If we have ample time after implementing the first two models, we would like to try implementing the transformer network recently proposed by Vaswani et al. [6] at Google Brain and see how it compares with the RNN based encoder-decoder models described above. The transformer network architecture replaces the recurrent units in the encoder-decoder architecture with self attention at both the input and output side, as well as an attention network connecting the two sides, thereby reducing computation complexity and improving long-range dependency learning.

Dataset

As mentioned above, we are going to use WMT newstest dataset [4] for our initial comparison. Potentially we are also considering aligned Canadian Parliament debate sentences [7], or the European Parliament proceedings parallel corpus [8]. If we use WMT, we would be essentially replicating Chung et al.'s results [3]; whereas Canadian Parliamentary transcripts will allow us to evaluate the two models on a different dataset than the original papers, and see how the performance results published in the papers hold for this dataset. Since most published results all pertain to European language family (En-De, En-Fr, etc), and WMT has recently published an English-Hindi sentence pair dataset, we would potentially also evaluate the models on this new language pair.

References

[1] Sennrich et al., 2016. "Edinburgh Neural Machine Translation Systems for WMT 16." https://arxiv.org/pdf/1606.02891.pdf

[2] Luong et al., 2015. "Addressing the Rare Word Problem in Neural Machine Translation." https://nlp.stanford.edu/pubs/acl15_nmt.pdf

[3] Chung et al., 2016. "A Character-Level Decoder without Explicit Segmentation for Neural Machine Translation." https://arxiv.org/abs/1603.06147

[4] WMT Dataset. http://www.statmt.org/wmt14/translation-task.html

[5] Sennrich et al., 2015. "Neural Machine Translation of Rare Words with Subword Units." https://arxiv.org/pdf/1508.07909.pdf

[6] Vaswani et al., 2017. "Attention Is All You Need." https://arxiv.org/pdf/1706.03762.pdf

[7] Aligned Canadian Parliament debate sentences dataset. https://www.isi.edu/natural-language/download/hansard/

[8] European Parliament proceedings dataset. http://www.statmt.org/europarl/