**Neural Machine Translation of Rare Words with Subword Units**

**Summary**

The main contribution of this paper is that we show that neural machine translation systems are capable of open-vocabulary translation by representing rare and unseen words as a sequence of subword units. The authors propose a technique novel technique to segment words into "subword units" based on the Byte Pair Encoding (BPE) algorithm in order to handle open vocabularies in Neural Machine Translation (NMT). Using this technique the authors achieve significant improvements over baseline systems without needing to resort to tricks such as UNK replacement and backoff dictionary alignments.

**Strengths**

Translation of out-of-vocabulary words is addressed by BPE by breaking words in subunits based on the intuition that various word classes are translatable via smaller units than words, improving the BLEU scores .

BPE segmentation gives a good balance between the vocabulary size and the decoding efﬁciency, and also sidesteps the need fora special treatment of unknown words

**Observations:**

* Translation of rare words is an open problem. Typically, NMT vocab is limited to fixed 30k-50k words.Intution: Many words are translatable using smaller units than words (e.g. word stems, suffixes, etc).
* Technique(byte pair encoding) is purely a data pre- and post-processing step. Nothing in the model needs to change.
* Basic idea behind algorithm: Start with a vocabulary of all character in the text, then iteratively merge the most frequent pair of characters or character sequences. Thus, the final vocabulary size is num\_chars + num\_merge\_operations where num\_merge\_operations is a hyperparameter of the method.
* Can learn BPE for source and target separately, or joint. Joint BPE has the advantage that words appearing in both source and target are split exactly the same way, making it easier for the model to learn the alignments.

**Weakness:**

* The paper propose to segment words into smaller units and translate just like at the word level, which does not learn to understand relationships among words.
* Vocabulary size of 90k or 60k seems quite large. We wonder if a much smaller BPE vocabulary size (20k, 30k) is enough