Summary of "Addressing the Rare Word Problem in Neural Machine Translation" Paper

`Addressing the Rare Word Problem in Neural Machine Translation.md`

Addressing the Rare Word Problem in Neural Machine Translation

Introduction

- NMT(Neural Machine Translation) systems perform poorly with respect to OOV(out-of-vocabulary) words or rare words.
- The paper presents a word-alignment based technique for translating such rare words.
- Link to the paper

Technique

- Annotate the training corpus with information about what do different OOV words (in the target sentence) correspond to in the source sentence.
- NMT learns to track the alignment of rare words across source and target sentences and emits such alignments for the test sentences.
- As a post-processing step, use a dictionary to map rare words from the source language to target language.

Annotating the Corpus

Copy Model

- Annotate the OOV words in the source sentence with tokens *unk1*, *unk2*,..., etc such that repeated words get the same token.

- In target language, each OOV word, that is aligned to some OOV word in the source language, is assigned the same token as the word in the source language.
- The OOV word in the target language, which has no alignment or is aligned with a known word in the source language. is assigned the null token.
- Pros
  - Very straightforward
- Cons
  - Misses out on words which are not labelled as OOV in the source language.

PosAll - Positional All Model

- All OOV words in the source language are assigned a single *unk* token.
- All words in the target sentences are assigned positional tokens which denote that the *jth* word in the target sentence is aligned to the *ith* word in the source sentence.
- Aligned words that are too far apart, or are unaligned, are assigned a null token.
- Pros
  - Captures complete alignment between source and target sentences.
- Cons
  - It doubles the length of target sentences.

PosUnk - Positional Unknown Model

- All OOV words in the source language are assigned a single *unk* token.
- All OOV words in the target sentences are assigned *unk* token with the position which gives the relative position of the word in the target language with respect to its aligned source word.
- Pros:
  - Faster than PosAll model.
- Cons
  - Does not capture alignment for all words.

Experiments

- Dataset
  - Subset of WMT'14 dataset
- Alignment computed using the Berkeley Aligner
- Used architecture from Sequence to Sequence Learning with Neural Networks paper.

Results

- All the 3 approaches (more specifically the PosUnk approach) improve the performance of existing NMTs in the order PosUnk > PosAll > Copy.
- Ensemble models benefit more than individual models as the ensemble of NMT models works better at aligning the OOV words.
- Performance gains are more when using smaller vocabulary.
- Rare word analysis shows that performance gains are more when proposition of OOV words is higher.