Subword Models

Shubham Gupta

February 28, 2020

1 Phonetics and phonology

- Phonetics is the sounds steam
- *Phonemes*: Small sets of units used commonly by humans to communicate with each other
- *Morphenes* Smallest sematic unit. Helps break down words into smaller units that have distinct meanings
- Difficult to work with morphenes though. Hence use character n-grams.
- Need to handle large, open vocab for word based models
- Transliteration: Translate names using how they sound in different language

2 Character level models

- Word embeddings composed from character embeddings
  - Generate embeddings for unknown words
  - Similar spellings =¿ Similar embeddings
  - Solves OOV problem
- OR processed language as characters
- Both methods work well
- Char level models give strong results via deep conv stack

3 Purely char-level NMT models

- Decoder only models worked well
- **Problem**: Sequences are too long. Models train/predict VERY slow.

4 Two trends of sub-word models

- Use smaller units: *word pieces*
- *Hybrid* arch:
  - Main models has words. Something else for chars

5 Byte Pair Encoding

- Originally compression algo
- Collection of stuff with bytes. Look for seq of 2 bytes. Add to dict of possible values
- Replace bytes with character ngrams
- Use as normal words in the NMT
- VERY good results. Top places in WMT 2016. Still used now as well
- Google: Use both wordpiece and sentencepiece
- Wordpiece model tokenizes inside words
- Sentence works with raw text
- Whitespace is retaied as
- Reverse things at end by joining pieces and recoding them to spaces
- BERT uses a variation for BPE
- Use CNN or BiLSTM to build word representations

6 Hybrid NMT

- Translate at word level
- Go to char level when needed
- Word level bean search

7 FastText

- Next efficient w2v. Better for rare words and longuages with lots of morphology
- w2v skip-gram with character n-grams
- Word as n-grams augmented with boundary sumbols and as a whole word
- where = ¡wh, whe, her, ere, re¿, ¡where¿
- Word is sum of these representations
- Some hashing trick to reduce vocab size