Week 3 Research Stay

Context

This week's focus was on Language modeling (natural language processing techniques that seek to predict words based on probability), N-grams (represent text or speech as a collection of n-length word sequence) and the advances and problems that these techniques represent.

Search Metodology

Since what we have researched so far is based on English, I decided to look for an article that mentioned how these techniques are applied (how different they are) in other languages, so I searched through the article's citations that mentioned something related to this.

Week 3 Research Stay

Preliminary Terms

- Language Modeling (LM): Probability based word prediction
- N-grams: Represent text or speech as a collection of n-length word sequences.
  "I like to play football" ↳ 2-grams (bigrams): ["I like", "like to", "to play", "play football"].
  ↳ 3-grams (trigrams): ["I like to", "like to play", "to play football"].
- Neural language models (LMs): Statistical language models based on neural networks
- Statistical language models (LMs): learns the probability of word occurrence based on examples of text
- OOV (Out-of-Vocabulary): terms that are not part of the normal lexicon found in a NLP environment
- UNK token: Unknown word, a word that doesn't exist the the vocabulary set

Week 3 Research Stay

Comparison [1]

In the first reading, it is mentioned that n-gram-based models can give relatively poor results. In the second, it is mentioned that, with the right tools, an n-gram-based model can be as good as one based on neural networks, which are much more expensive and may also be slower.

Now, how effective are n-grams when we consider other languages?

Week 3 Research Stay

Comparison [2]

Since I started reading about NLP, many of the techniques mentioned seemed to me to be specifically useful for English, but might not be as effective in other languages.

According to the article I found (“Are All Languages Equally Hard to Language-Model?”), factors such as grammatical complexity, sentence structure, lexical richness, and the frequency of certain linguistic patterns affect the difficulty of modeling. Languages with a more flexible grammatical structure or rich morphology tend to be more challenging for language models. This adds a new level of difficulty to NLP

References

- Beyond n-grams: Can linguistic sophistication improve language modeling?
- Show some love to your n-grams: A bit of progress and stronger n-gram language modeling baselines
- Building Wikipedia N-grams with Apache Spark

- Are All Languages Equally Hard to Language-Model?