NLP - Master Notes

June 11, 2021

1 Language Models

1.1 N-Gram Models

- Language (prediction) models which make the *Markov assumption* for an $(n-1)^{th}$ order Markov Model; i.e. that only the previous n-1 words have a probabilistic dependence on the current word.
- Probability of words 1 to n: $P(w_1^n) = \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1})$
- Pre-process the input to include n-1 $< s >$ symbols and a $< /s >$ symbol at the start and end of each sentence respectively.
- *example:*

  Text:
  One cat sat. Three **cats sat**. Eight **cats sat**. The **cats** had nine lives.

- Sentence Generation: until you produce a $< /s >$ symbol, continually generate words using: $argmax_(w_k) \frac{C(w_{k-n+1},...,w_k)}{C(w_{k-n+1},...,w_{k-1})}$
- Perplexity:
- Smoothing:
  - Laplace (add-one): $P(W_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w)+1}{c(w_{n-N+1}^{n-1})+V}$

  General steps for creating an n-gram model:

  1. choose a vocabulary
  2. replace unknown words in the training corpus with ¡UNK¿
  3.