ECE467 – Pset 1

Jonathan Lam

February 27, 2021

1. Consider the following sentence: "*Engineering is a complex endeavor.*" Assume that an HMM POS tagger is used to choose between these two tag-sequences (involving Penn Treebank POS tags) for the given sentence:

   (a) `Engineering/NN is/VBZ a/DT complex/NN endeavor/NN ./.`

   (b) `Engineering/NN is/VBZ a/DT complex/JJ endeavor/NN ./.`

   In other words, the system is only considering a single tag for most of the words, but the word "complex" can be tagged as either a singular noun or an adjective. Without stating any actual numbers, show which tag transition probabilities and word observation likelihoods (a.k.a. emission probabilities) will determine the choice for this tag. In terms of these probabilities, express an equation that must hold for the JJ tag to be chosen.

   In general, an HMM tries to fit the tag sequences following the equation (assuming a bigram approximation):

   $$\hat{t}_1^n \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} [P(w_i \mid t_i) P(t_i \mid t_{i-1})]$$

   JJ would be chosen as the tag for "complex" if:

   $$P(\text{"complex"} \mid \text{JJ}) P(\text{JJ} \mid \text{DT}) P(\text{NN} \mid \text{JJ}) > P(\text{"complex"} \mid \text{NN}) P(\text{NN} \mid \text{DT}) P(\text{NN} \mid \text{NN})$$

   where $P(\text{"complex"} \mid \text{JJ})$ represents an emission probability and $P(\text{JJ} \mid \text{DT})$ represents a transition probability. (These are the only terms that are different in the approximation.)

2. Consider all words representing the sound a cow makes to be the following infinite set: "moo!", "mooo!", "moooo!", etc. Show a regular expression that would accept any line in which the sound a cow makes occurs at least two times (they don't have to be consecutive or identical). Each instance must be preceded by a non-letter or occur at the start of a line. It does not matter what occurs after the explanation points.

   Using standard PCRE syntax, with non-capturing groups:

   ```
   (?:(?:^|[^a-zA-Z])moo+!.*){2,}
   ```

3. Assume a bigram model and a trigram model are both trained on a large corpus of English that includes many technical, AI-oriented documents. Consider the bigram probability associated with the phrase "language processing" and the trigram probability associated with the phrase "natural language processing". Which do you expect would be larger? Explain your answer.

   The bigram probability is $P(\text{"processing"} \mid \text{"language"})$. The trigram probability is $P(\text{"processing"} \mid \text{"natural"}, \text{"language"})$. The trigram probability is likely higher because when you say "natural language" you are typically talking in a scientific context and are likely talking about NLP (especially in an AI-related corpus), whereas the word "language" can be followed by many different words and is more likely not related to a phrase on language processing.

4. What is the major linguistic distinction between languages such as Mohawk and languages such as English?

   Mohawk is a "polysynthetic language," which means that "verbs must include some expression of each of the main participants in the event described by the verb (the subject, object, and indirect object)." English does not have this parameter.

5. Consider an information retrieval system that relies on a vector space model and TF*IDF weights. A straight-forward system would have to loop through all documents in the collection to compute the similarity between the query and each document. Briefly explain how an inverted index can be used to make the IR system much more efficient.

   A full document-term matrix would be very large and sparse. An inverted-index data structure would map each term to which documents it occurs in and how many times it occurs in those documents, which is sufficient to calculate the TF*IDF metric while saving space and only looping though documents that contain the term.