Your Name: _____

COMP 394 (NLP), Fall 2024

Practice Exam 1

Suhas Arehalli

Notes about this exam:

- Please write legibly and large enough so I can read your answers.
- You **may and should** ask me for clarification if you have any questions.
- You may use your handwritten notecard you have created.
- This is an individual assignment **DO NOT** discuss it with anyone.

Problem 1

Suppose you were given the following set of unigram counts from a corpus.

this: 17
then: 9
the: 43
withering: 14
rowing: 4
check: 7

(a) Begin running the BPE algorithm using the provided word counts and provide the first 3 merges the algorithm would compute. For each merge, indicate the two tokens that should be merged.

(b) Suppose a BPE tokenizer is trained on the above corpus and learns just those 3 merges. How would BPE tokenize the word *there*? Be sure to clearly indicate token boundaries.

Problem 2

Consider a Context-Free Grammar G with the following rules and the appropriate terminals and non-terminals. The start symbol is S.

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N} \\
\text{VP} &\rightarrow \text{V} \\
\text{VP} &\rightarrow \text{VP NP} \\
\text{N} &\rightarrow \textit{student} \\
\text{N} &\rightarrow \textit{students} \\
\text{N} &\rightarrow \textit{dog} \\
\text{N} &\rightarrow \textit{dogs} \\
\text{N} &\rightarrow \textit{ball} \\
\text{N} &\rightarrow \textit{balls} \\
\text{V} &\rightarrow \textit{throw} \\
\text{V} &\rightarrow \textit{throws} \\
\text{D} &\rightarrow \textit{a} \\
\text{D} &\rightarrow \textit{the}
\end{aligned}
$$

(a) Provide a parse tree that demonstrates that $\omega_1 = \textit{the students throw the dogs a ball}$ is grammatical with respect to G.

(b) Is the grammar above in Chomsky-Normal Form? If not, indicate how the grammar would need to be modified to convert it to an equivalent CNF Grammar.

(c) Observe that a sentence like $\omega_2 = the\ students\ throws\ the\ dogs\ a\ ball$, which violates subject-verb agreement, is also grammatical under G. Indicate how you might update the rules of this context-free grammar such that ω_2 is ungrammatical under G but ω_1 remains grammatical.

Problem 3

Consider the following small, tokenized corpus:

> EOS this is a test sentence . EOS this is n't because it is a part of a test set , but because it is on a practice test . EOS

(a) Suppose you trained an bigram model on this corpus with no smoothing. What is $p(on \mid is)$?

(b) What is $p(\text{this is n't a test EOS})$ under the bigram model?

(c) What is the perplexity of the sequence "EOS this is n't because it is a test . EOS" under the bigram model?

Problem 4

For each issue below, propose and briefly justify a strategy to solve the problem. Be sure to identify what parts of the description given may have lead to the problem and how you might test whether your strategy will work.

(a) You train a 3-gram model on a large corpus of articles from Wikipedia and estimate the parameters using Maximum Likelihood Estimation. You begin to evaluate the model on the validation set and end up computing infinite/undefined perplexities.

(b) You train a 3-gram model on a small corpus of articles from Wikipedia and estimate the parameters using Laplace Smoothing. You begin to generate from the model, and find that nonsensical continuations (i.e., those like "Dog the which food ate") are generated much more often than you'd like.

(c) You train an n-gram model on a large corpus of articles from Wikipedia that interpolates between MLE-estimated unigram, bigram, and trigram models. you find reasonable looking perplexities, but run targetted evaluations and find that your models fail to respect grammatical constraints like agreement for preambles like "the dog with the bone…"