Practice Problems for Midterm

1 Finite State Transducers

1. State the difference between inflectional and derivational morphology, noting their relation with word class

2. What is the difference between a FSA and a FST?

3. Write a transducer or transducers for the consonant doubling spelling rule in English.
   The simplified consonant doubling rules:
   Rule: Words ending with a Consonant-Vowel-Consonant Pattern
   ED = If the word ends in a CVC pattern, it gets a double consonant + ED. * note
   ING = If the word ends in a CVC pattern, it gets a double consonant + ING. * note
   Examples:
   ED = RUB → rubbed, STOP → stopped
   ING = HOP → hopping, SIT → sitting
   Otherwise, do not double.
   (This is Exercise 3.4 from the textbook)
2 Naive Bayes Classifier

Suppose that we are using a Naive Bayes Classifier to tell if the topic of a document is Pittsburgh. Assume that all the training set is included in the table. All the words used as features are included in the table. Can you tell if the test set document is about Pittsburgh? Show how you get the result.

<table>
<thead>
<tr>
<th>Keywords in Document</th>
<th>If topic is Pittsburgh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td></td>
</tr>
<tr>
<td>Pittsburgh Pirates</td>
<td>True</td>
</tr>
<tr>
<td>Pittsburgh Pittsburgh CMU</td>
<td>True</td>
</tr>
<tr>
<td>Pittsburgh University</td>
<td>True</td>
</tr>
<tr>
<td>Pittsburgh Cleveland Philadelphia</td>
<td>False</td>
</tr>
<tr>
<td>Test set</td>
<td>Pittsburgh Pittsburgh Pittsburgh CMU</td>
</tr>
</tbody>
</table>

Table 1: Naive Bayes Classifier
3 Noisy Channel

Noisy Channel is widely used for Part of Speech (POS) Tagging, in which a sequence of tokens are marked correspondingly as a sequence of tags, based on both tokens definition and context.

1. If we have $Y \rightarrow \text{Channel} \rightarrow X$, then what is $Y$ and $X$ in the context of POS Tagging?

2. What probability expression do we need to maximize in the context of POS Tagging?

3. How would Noisy Channel Model be applied to machine translation?

4. Give another in the field of NLP that could be solved using Noisy Channel.
4 N-gram Model and Smoothing

In a landmark 1996 paper, Saffran et al. demonstrated that infants can distinguish likely transitions between syllables from unlikely ones. For example, in the phrase “pretty baby”, the babies can tell that words are likely to contain a transition from “pre-” to “-ty”, but they do not usually contain a transition from “-ty” to “ba-”.

To simplify slightly, the experimenters played the babies several minutes of “training data” - simulated speech consisting solely of four randomly repeated fake words: pabiku, tibudo, golatu, and daropi. They then compared the babies’ reactions to words from this list and to other fake words generated from the same syllables that had not been played (e.g., tudaro and pigola).

1. Assume that the training data consisted solely of the following sequence (with dashes separating syllables and spaces separating words):
   Assuming the babies have learned a bigram model of syllables, what would be their estimates of $P(ku \mid bi)$, $P(do \mid bu)$, $P(da \mid ku)$, and $P(ti \mid pi)$? (You can leave these as fractions.)

2. Now assume the babies apply add-one smoothing to their bigram model. What would be their smoothed estimates of $P(ku \mid bi)$ and $P(da \mid ku)$?

3. Why wouldn’t the babies want to use add-one smoothing with this amount of training data? What simple modification could you make to the smoothing to make it work better?

4. Would the babies want to use smoothing if they heard three hours of speech? Why or why not?

5. Consider a complete bigram model for the syllables in these words. How many parameters would this model have? In other words, how many entries (including entries that are zero) will be in the conditional probability table?

6. When we speak, we don’t typically pause between words, leaving it to the listener to distinguish the word boundaries. How would you use a bigram model like this one to help you segment speech into words? What preprocessing would your system have to do to the sound data before it could apply this model?
5 POS tagging and HMM

Recall that the Church (1988) tagger is not an HMM tagger since it incorporates the probability of the tag given the word:

\[ P(\text{tag} \mid \text{word}) \times P(\text{tag} \mid \text{previous } n \text{ tags}) \]

rather than using the likelihood of the word given the tag, as an HMM tagger does:

\[ P(\text{word} \mid \text{tag}) \times P(\text{tag} \mid \text{previous } n \text{ tags}) \]

As a thought-experiment, construct a sentence, a set of tag transition probabilities, and a set of lexical tag probabilities that demonstrate a way in which the HMM tagger can produce a better answer than the Church tagger, and another example in which the Church tagger is better.

(This is the Exercise 5.7 from the textbook)
6 Syntax and Parsing

Attached is a chart showing one way of visualizing the CYK (CKY) parsing algorithm. Note that other ways of drawing such charts are used in different sources. It is up to you to figure out how the operation of CYK is illustrated in this chart. Based upon the chart and your knowledge of the CYK algorithm, answer the following questions:

1. What properties must all the rules in the grammar being employed in this case have? What is the name for the way of writing CFGs so that they have these properties?

2. What rules can you infer to be part of this grammar based upon this chart?

3. What is the relationship between the position of a symbol in the chart and the associated span of terminals?

4. Give a complete parse tree corresponding to the parse shown in the chart.

5. How does the CYK algorithm, when it is used for parsing, interact with syntactic ambiguities (relative to the grammar)? How might the CYK algorithm be modified so that all possible parses are preserved?