Assignment 4

L545

Due Monday, February 25

1. Using your knowledge of n-grams and perhaps being inspired by papers such this one, *The Unreasonable Effectiveness of Data* (Halevy, Norvig, & Pereira (2009), `http://www.csee.wvu.edu/~gidoretto/courses/2011-fall-cp/reading/TheUnreasonable%20EffectivenessofData_IEEE_IS2009.pdf`), explain to a layman what n-grams models are, what they do, what they don't do well, what smoothing is, and one of the ideas behind doing smoothing well. You should address whether linguistic knowledge is necessary in NLP. You are writing not to me, but to a reasonably educated person who knows nothing about NLP. Write no more than a page.

2. Do question 5.2 on page 171 in Jurafsky & Martin, i.e., annotate some selected sentences by hand with POS labels from the Penn Treebank. (The tagset is located on page 131, and the guidelines are referenced in the next question.)

3. (a) Using the guidelines for the Penn Treebank tagging found at: `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz` and your own linguistic intuitions, hand-write some rules to distinguish the following tags:
   i. VBD vs. VBN
   ii. JJ vs. VBN

   (b) The following are some rules learned by a Brill tagger. Explain what they do and why they work.
   i. `VB VBP PREVTAG NNS`
   ii. `IN WDT NEXT1OR2TAG VB`
   iii. `JJ NNP NEXTTAG NNP`

4. (a) The tagging systems we talked about in class use the preceding context (e.g., the preceding tag) to disambiguate a tag for a given word. Why? And would you expect this to work for all languages? Why or why not?

   (b) Some tagging decisions seem to depend upon the following word; for example, *I can/VBP tuna* vs. *I can/MD help*. HMM taggers condition on the previous tags, e.g., $P(w_i|w_{i-2}w_{i-1})$. How is it, then, that HMM taggers can often tag these cases accurately? In other words, how do they capture cases where the tag of w_i depends upon w_{i+1}?