Assignment 7
L245
Due Monday, April 17

1. [**Moved from previous assignment:**] (Based on a previous assignment by Jason Baldridge):

   Assume you have a corpus of 4200 tweets about broccoli that have been labeled for whether they are positive or negative in their sentiment. 3247 of them are negative. We find the word *hate* in 643 negative messages and in 28 positive messages.

   (a) What is $P(negative)$ according to the above dataset?

   (b) What is $P(hate|negative)$ according to the above dataset?

   (c) What is $P(positive)$ according to the above dataset?

   (d) What is $P(hate|positive)$ according to the above dataset?

   (e) Using the ham-to-spam ratio as a model, what is the positive-to-negative ratio for the word *hate*?

2. Do question #5 from chapter 5 of the textbook (p. 152).

   - Section 5.5.1 has a miscalculation (`http://purl.org/lang-and-comp/errata/`), whereas the calculations on our slides 29–30 are correct.

3. Do question #1 from chapter 7 of the textbook (p. 210).

4. Do question #2 from chapter 7 of the textbook (p. 210).

5. The goal of this question is to determine how much semantic ambiguity is a problem in machine translation.

   (a) Come up with an English word that you find to be semantically ambiguous. To verify its ambiguity, go to WordNet (`http://wordnet.princeton.edu`, search interface: `http://wordnetweb.princeton.edu/perl/webwn`) and confirm that it has at least 5 different meanings. Keep this page open.

   (b) Now, go to the OPUS parallel corpus (`http://opus.lingfil.uu.se`). Click on *OPUS multilingual search interface* (or go directly to: `http://opus.lingfil.uu.se/bin/opuscqp.pl`). Select a corpus or corpora of your own choosing and then select *en*, since you'll be searching for an English word.

   (c) Search for your ambiguous word and click that you want to see (sentence) alignments with at least one other language (preferably one you do not know, but this is not crucial).

   (d) Based on these sentence alignments, write down your best guesses as to how translate your word. It is almost a guarantee that you will have several possible translations. Note that: a) there may be some sentence alignment errors; b) in some languages, a single word may translate to a subword (i.e., morpheme); and c) some languages are written in writing systems you may not know (which could be fun to try or could be a headache you don't want to deal with).

   (e) Based on your guesses and on the WordNet meanings, do you see a connection between a word's meaning and its translation? This question is rather open-ended, and the best responses will discuss a range of examples showing challenges you encountered.