Fall Semester 2016
Iowa State University

ENGL 520 - Computational Analysis of English

Problem Set 4
Language Models and POS Tagging
(ungraded)

1. Take any ready to use POS tagger available online and do the following analysis: pick up some samples of non-canonical language such as spoken language, child language (example source: CHILDES database for English), tweets, historical English etc., Run the POS tagger on these sample sentences and study how robust the POS tagger is.

2. Do problem 5.6 in Jurafsky and Martin textbook, 2nd Edition, Chapter 5.

3. Identify specific patterns of English learner errors that can affect a POS tagger performance, and program a set of additional rules that works on a incorrect POS-tagged output and corrects it.

4. Program a simple tagger, that just assigns the most likely tag for each word in the sentence, and assigns NN for any unknown word. Use the training data given for Assignment 3 to build the most likely tag frequencies list.

5. Take a non-English language you can speak, and explore if there are any POS tagging support for that language.

6. Based on the output of a POS tagger and regular expressions, write a program to identify names of educational organizations (Iowa State University, Ames Middle School etc) in a input sentence.

7. Write a tagger that tags date and money expressions in a sentence, using regular expressions.

8. Use the tagger you developed for Assignment 3, and test it using all the sentences in the training set as input (without the tags!). Now, calculate the accuracy of the tagger by calculating the percentage of tags your tagger identified correctly (compare the output of the tagger with the actual tags of words). Please note: This is not a useful measure in real-life. We usually test using some data that was not used during the training process.

9. Read Chapter 5 of NLTK book[1], to understand different varieties of POS taggers and their differences. Python users: Try to follow the examples. Perl users: Look for perl modules that implement POS taggers and compile a list - learn to use them if possible.

10. Assuming that we get a POS tagged sentence as input, write a program (using regular expressions), to identify dependent clauses in a sentence.

[1] `http://www.nltk.org/book/ch05.html`