Laboratory 2: Running the Part of Speech Taggers

This page provides more information to help you run the part-of-speech taggers for Laboratory 2.

Tokenizing the text

Note that you must "tokenize" the input to both the taggers. In particular, you must perform the following substitutions:

- Split punctuation from adjoining words
- Convert double quotes (") to doubled single forward and backward quotes (`` and '')
- Split verb contractions and possessive 's from the component morphemes:
  - `children's -> children 's`
  - `parents -> parents '`
  - `won't -> wo n't`
  - `gonna -> gon na`
  - `I'm -> I 'm`

The texts provided have been tokenized for you. If you want to try tagging any other text, please make sure that it is properly tokenized.

Running the taggers

Brill tagger

The Brill tagger is described in the textbook, and can be downloaded from Eric Brill's home page at: http://research.microsoft.com/~brill/. His paper, entitled "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging," describes the tagger in more detail. In particular, you should familiarize yourself with the `-i <filename>` output option of the tagger, which you will need to answer the lab questions. This makes the tagger write out its first, unamended guesses for tags to a separate file as well as its intermediate files and the rules it is using to change tags. If you use the default arguments as given below you are using a full system trained on the entire Brown corpus of one million words and 5 million words from the Wall Street Journal. (The README files discuss this in much greater detail.)