Basic Text Processing: Sentence Segmentation

Importance of Punctuation

- So far we have discussed what words to keep and possible alternate forms of words, i.e. *word normalization*, through lower casing, stemming and lemmatization
- Note that for further steps of language processing, we need to keep all the punctuation as tokens.
- Punctuation determines the clauses of a sentence and can profoundly affect the meaning
  - From the book "Eats, Shoots and Leaves: The Zero Tolerance Approach to Punctuation" by Lynne Truss, a collection of quotes: http://www.goodreads.com/work/quotes/854886-eats-shoots-leaves-the-zero-tolerance-approach-to-punctuation
  - Another example seen on a t-shirt:

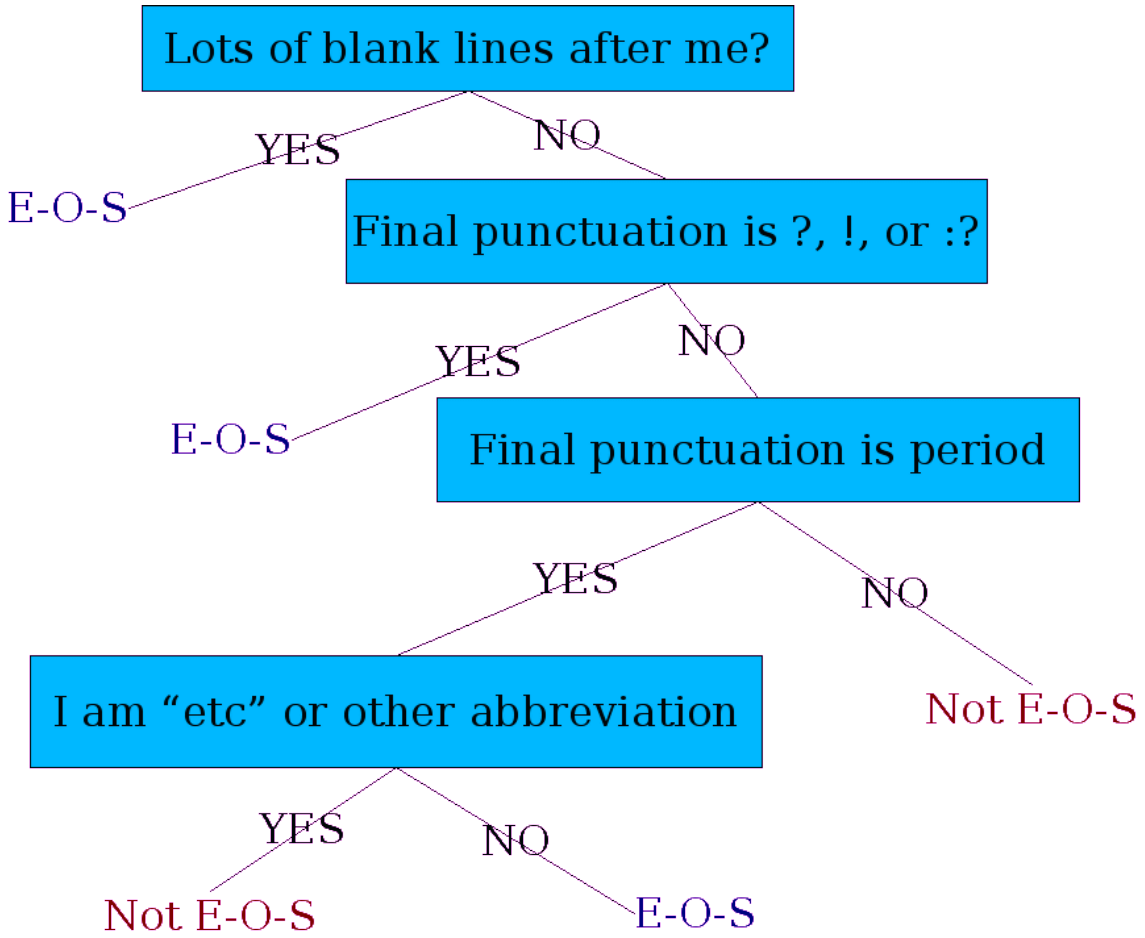
    Let's eat Grandma!
    Let's eat, Grandma!
    Commas save lives!

Sentence Segmentation

- Punctuation not only shows internal structure of sentences, but is crucial in determining the end of sentences.
- EndOfSentence determined by lots of white space or punctuation !, ?, .
  - !, ? are relatively unambiguous
  - Period "." is quite ambiguous
    - Sentence boundary
    - Abbreviations like Inc. or Dr.
    - Numbers like .02% or 4.3
- Treat this as a classification problem
  - Looks at a "." (or the word with the "." at the end)
  - Decides EndOfSentence/NotEndOfSentence
  - Classifiers: hand-written rules, regular expressions, or machine-learning

Classify whether a word is End-of-Sentence

- An example of one way to classify is a Decision Tree:



Slides in this section are from Dan Jurafsky

Classification Problem Features

- Each property used in the decision tree to decide which branch to take is called a feature of the word
- Features for end-of-sentence decision
  - Word with "." is on list of abbreviations
  - Word shape features
    - Case of word with ".": Upper, Lower, Cap, Number
    - Look at word after "." to see if it begins a new sentence:
      - Case of word after ".": Upper, Lower, Cap, Number
  - Numeric features
    - Length of word with "."
    - Probability(word with "." occurs at end-of-s)
    - Probability(word after "." occurs at beginning-of-s)