Part-of-Speech Tagging

Parts of Speech (POS)

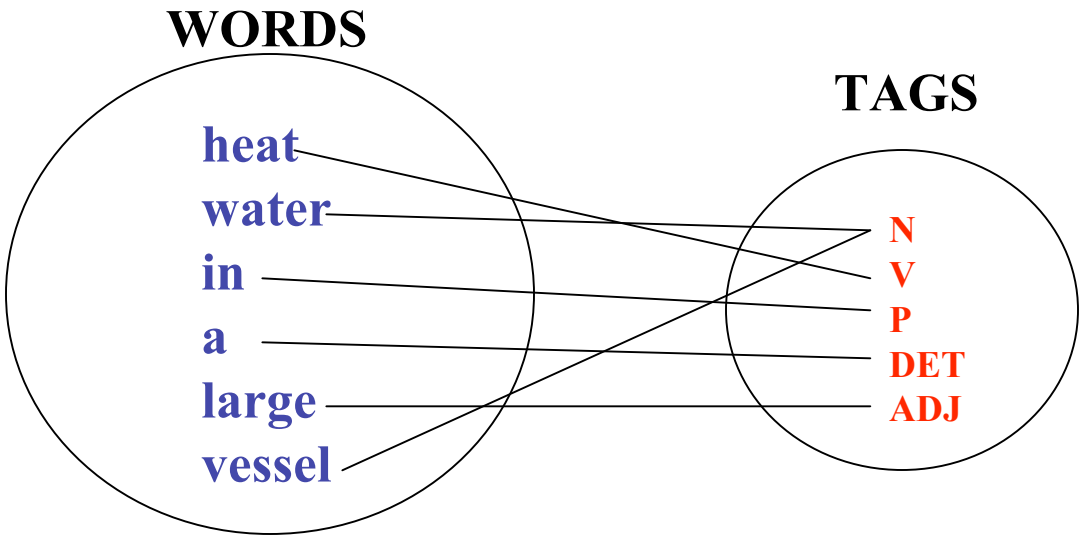
- Categories to which words are assigned according to their function.
  - Noun, verb, adjective, preposition, adverb, article, pronoun, conjunction, etc.
  - This idea has been around for over 2000 years (Dionysius Thrax of Alexandria, 100 B.C.)
  - Called: parts-of-speech, lexical category, word classes, morphological classes, lexical tags, POS

POS examples

• N	noun	chair, bandwidth, pacing
• V	verb	study, debate, munch
• ADJ	adjective	purple, tall, ridiculous
• ADV	adverb	unfortunately, slowly,
• P	preposition	of, by, to
• PRO	pronoun	I, me, mine
• DET	determiner	the, a, that, those

POS Tagging

- The process of assigning a part-of-speech to each word in a sentence



Example

Word	**Tag**
heat	verb (noun)
water	noun (verb)
in	prep (noun, adv)
a	det (noun)
large	adj (noun)
vessel	noun

What is POS tagging good for?

- Useful in
  - Information Retrieval
  - Text to Speech: OBject(N) vs. obJECT(V);
    DIScount(N) vs. disCOUNT(V)
  - Word Sense Disambiguation

- Useful as a preprocessing step of parsing
  - Unique tag to each word reduces the number of parses

Choosing a tagset

- Need to choose a standard set of tags to do POS tagging
  - One tag for each parts-of-speech
- Could pick very coarse tagset
  - N, V, Adj, Adv, Prep.
- More commonly used set is finer-grained
  - E.g., the *UPeen TreeBank* tagset, 45 tags
    - PRP, PRP$, VBG,VBD, JJR, JJS …
- Even more finely-grained tagsets exist

Why is POS tagging hard?

- Ambiguity
  - Plants/N need light and water.
  - Each one plant/V one.

- The number of tags used by different systems varies a lot. Some systems use < 20 tags while others use >70.

Methods for POS tagging

- Rule-Based POS tagging
  - ENGTWOL [Voutilainen, 1995]
- Stochastic (Probabilistic) tagging
  - TNT [Brants, 2000]
- Transformation-based tagging
  - Brill's tagger [Brill, 1995]

Stochastic Tagging

- Based on probability of certain tag occurring, given various possibilities
  - Necessitates a *training corpus*
    - A collection of sentences that have already been tagged
  - Several such corpora exist

Approach 1

- Assign each word its most likely POS tag
  - Success: 91% for English

- Example

heat :: noun/89, verb/5

Approach 2

- Given: sequence of words

$$W = w_1, w_2, \ldots, w_n \text{ (a sentence)}$$

- Assign tags:

$$T = t_1, t_2, \ldots, t_n$$

- Find T that maximizes P(T | W)

Practical Statistical Tagger

- By Bayes' Rule,

$$P(T \mid W) = P(W|T)\ P(T) / P(W) = \alpha\ P(W|T)\ P(T)$$

- So find T that maximizes P(T | W) P(T)

$$P(T) = P(t_1)\ P(t_2 \mid t_1)\ P(t3 \mid t_1, t_2)\ \ldots\ P(t_n \mid t1, t_2, \ldots\ t_{n-1})$$

  - As an approximation, use

$$P(T) \approx P(t_1)\ P(t_2 \mid t_1)\ P(t_3 \mid t_2)\ \ldots\ P(t_n \mid t_{n-1})$$

- Assume each word is dependent only on its own POS tag: given its POS tag, it is independent of the other words around it. Then

$$P(W|T) = P(w_1 \mid t_1)\ P(w_2 \mid t_2)\ \ldots\ P(w_n \mid t_n)$$

- So

$$P(T)\ P(W|T) \approx P(t_1)\ P(t_2|t_1)\ \ldots\ P(t_n|t_{n-1})\ \ P(w_1|t_1)\ P(w_2|t_2)\ \ldots\ P(w_n|t_n)$$

Getting the Conditional Probabilties

- How do we compute $P(w_i|t_i)$?
  - $c(w_i,t_i)/c(t_i)$
- How do we compute $P(t_i|t_{i-1})$?
  - $c(t_{i-1},t_i)/c(t_{i-1})$

Example

- Secretariat/NNP is/VBZ expected/VBN to/TO **race**/VB tomorrow/NN
  - to/TO race/???
- People/NNS continue/VBP to/TO inquire/VB the DT reason/NN for/IN the/DT **race**/NN for/IN outer/JJ space/NN
  - the/DT race/???

- $t_i = \text{argmax}_j \, P(t_j|t_{i-1})P(w_i|t_j)$
  - i = number of word in sequence, j = number among possible tags
  - max[P(VB|TO)P(race|VB) , P(NN|TO)P(race|NN)]

- From the Brown corpus
  - P(NN|TO) = .021 P(race|NN) = .00041
  - P(VB|TO) = .34 P(race|VB) = .00003
- So
  - P(NN|TO) P(race|NN) = .021 × .00041 = .000007
  - P(VB|TO) P(race|VB) = .34 × .00003 = .00001