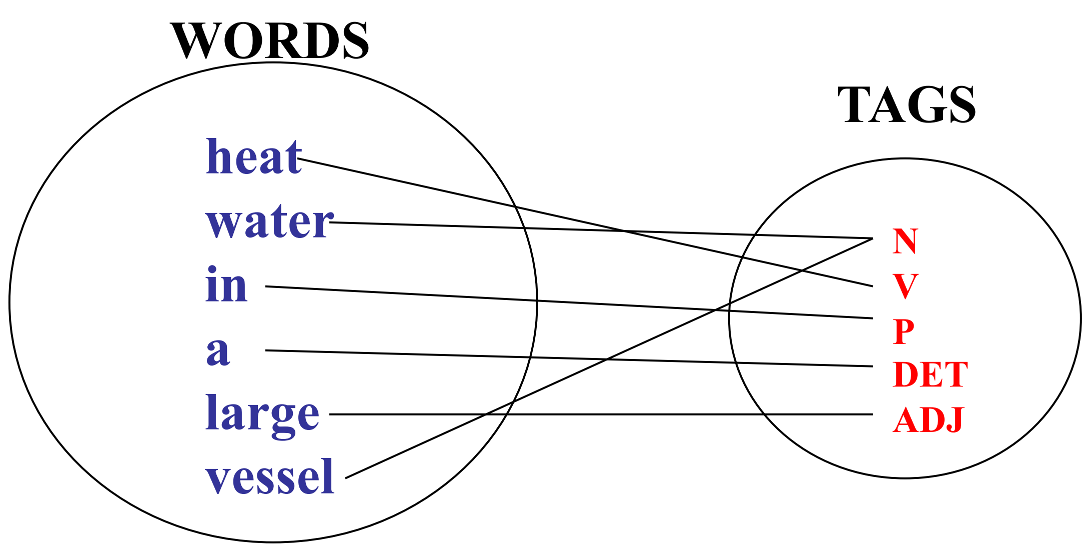
# Part-of-Speech Tagging

- 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: ***ob***ject(N) vs. ob***ject***(V);
    ***dis***count(N) vs. dis***count***(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 part of speech
- Could pick very coarse tagset
  - N, V, Adj, Adv, Prep.
- More commonly used set is finer-grained
  - E.g., the *UPenn TreeBank II* tagset has 36 word tags
    - PRP, PRP$, VBG,VBD, JJR, JJS …
  - (also has tags for phrases)
- Even more finely-grained tagsets exist

# Why is POS tagging hard?

- Ambiguity
  - “Plants/N need light and water.”
  - “Each one plant/V one.”
  - “Flies like a flower”
    - *Flies*: noun or verb?
    - *like*: preposition, adverb, conjunction, noun, or verb?
    - *a*: article, noun, or preposition?
    - *flower*: noun or verb?

# Methods for POS tagging

- Rule-Based POS tagging
  - e.g., ENGTWOL [ Voutilainen, 1995 ]
    - large collection (> 1000) of constraints on what sequences of tags are allowable
- Transformation-based tagging
  - e.g.,Brill's tagger [ Brill, 1995 ]
  - sorry, I don't know anything about this
- Stochastic (Probabilistic) tagging
  - e.g., TNT [ Brants, 2000 ]
  - I'll discuss this in a bit more detail

# 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
    - One of the best known is the Brown University Standard Corpus of Present-Day American English (or just the **Brown Corpus**)
    - about 1,000,000 words from a wide variety of sources
      - POS tags assigned to each

# Approach 1

- Assign each word its most likely POS tag
  - If w has tags $t_1, \ldots, t_k$, then can use

    $P(t_i \mid w) = c(w,t_i)/(c(w,t_1) + \ldots + c(w,t_k))$, where

    - $c(w,t_i) =$ number of times $w/t_i$ appears in the corpus
  - Success: 91% for English

- Example

  heat :: noun/89, verb/5

# Approach 2

- Given: sequence of words W

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

  - e.g., W = heat water in a large vessel
- Assign sequence of tags T:

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

- Find T that maximizes $P(T \mid W)$

# 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(W | T) P(T)
  - Chain rule:

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

  - 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 conditionally 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

- Want to compute
  - $P(T)\ P(W|T) \approx P(t_1)\ P(t_2|t_1) \ldots P(t_n|t_{n-1})\ \ P(w_1|t_1)\ P(w_2|t_2) \ldots P(w_n|t_n)$
- Let
  - $c(t_i)$ = frequency of $t_i$ in the corpus
  - $c(w_i,t_i)$ = frequency of $w_i/t_i$ in the corpus
  - $c(t_{i-1},t_i)$ = frequency of $t_{i-1}\ t_i$ in the corpus
- Then we can use
  - $P(t_i|t_{i-1}) = c(t_{i-1},t_i)/c(t_{i-1})$,
  - $P(w_i|t_i) = c(w_i,t_i)/c(t_i)$

# 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/???

- For each word $w_i$, $t_i = \text{argmax}_t\ P(t|t_{i-1})P(w_i|t)$
  - 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

# *UPenn TreeBank II* word tags

- CC - Coordinating conjunction
- CD - Cardinal number
- **DT** - Determiner
- EX - Existential there
- FW - Foreign word
- **IN** - Preposition or subordinating conjunction
- **JJ** - Adjective
- JJR - Adjective, comparative
- JJS - Adjective, superlative
- LS - List item marker
- MD - Modal
- **NN** - Noun, singular or mass
- NNS - Noun, plural
- **NNP** - Proper noun, singular
- NNPS - Proper noun, plural
- PDT - Predeterminer
- POS - Possessive ending
- PRP - Personal pronoun
- PRP$ - Possessive pronoun
- RB - Adverb
- RBR - Adverb, comparative
- RBS - Adverb, superlative
- RP - Particle
- SYM - Symbol
- **TO** - to
- UH - Interjection
- VB - Verb, base form
- VBD - Verb, past tense
- VBG - Verb, gerund or present participle
- **VBN** - Verb, past participle
- VBP - Verb, non-3rd person singular present
- **VBZ** - Verb, 3rd person singular present
- WDT - Wh-determiner
- WP - Wh-pronoun
- WP$ - Possessive wh-pronoun
- WRB - Wh-adverb