PARTS OF SPEECH TAGGING

INTRODUCTION

- Parts of speech (POS), word classes, morpho-logical classes, or lexical tags give information about a word and its neighbors

- Since the greeks 8 basic POS have been distinguished:
Noun, verb, pronoun, preposition, adverb, conjunction, adjective, and article

- Modern works use extended lists of POS: 45 in Penn Treebank corpus, 87 in Brown corpus

PARTS OF SPEECH TAGGING

Tagging is the process of assigning a tag to a word in a corpus

Used for different tasks

- Speech recognition. Pronunciation may change: DIScount noun, disCOUNT verb,

- Information retrieval- morphological affixes

- Lingusitic research- frequency of structures

PARTS OF SPEECH CATEGORIES

- **Closed class.** Function words: prepositions, pronouns, determiners,conjunctions, numerals, auxiliary verbs and particles (preposition or adverbs in phrasal verbs)
- **Open class**:
  - **Nouns:** people, place and things
    - proper nouns, common nouns,count nouns and mass nouns
  - **Verbs:** actions and processes. Main verbs, not auxiliaries
  - **Adjectives:** Properties
  - **Adverbs**

PARTS OF SPEECH TAGGING

PAVLOV **N NOM SG PROPER**
HAVE **V PAST VFIN**
SVO (verb with subject and object)
HAVE **PCP2(past participle) SVO**
Shown **SHOW PCP2 SVO SV**
SVOO (verb with subject and two complements)
that
ADV
PRON DEM SG
DET CENTRAL DEM SG
CS (subordinating conjunction)
salivation **N NOM SG**

PARTS OF SPEECH TAGGING

ADVERBIAL - *THAT* RULE
Given input: “that”
if

(+1 A/ADV/QUANT); /* if next word is adj, adverb, or quantifier */

(+2 SENT-LIM); /* and following is a sentence boundary, */

(NOT -1 SVOC/A); /* and the previous word is not a verb like */

/* ‘consider’ which allows adjs as object complements */

then eliminate non-ADV tags

else eliminate ADV tag

Ex: In the sentence *“I consider that odd “*, that will not be tagged as adverb (ADV)

Brill's set of templates

"Change tag ***a*** to tag ***b*** when: .."

a,b,z and **w** are part of speech tags

The preceding (following) word is tagged ***z***.
The word two before (after) is tagged ***z***.
One of the two preceding (following) words is tagged ***z***.
One of the three preceding (following) words is tagged ***z***.
The preceding word is tagged ***z*** and the following word is tagged ***w***.
The preceding (following) word is tagged ***z*** and the word
two before (after) is tagged ***w***.

STATISTICAL POS TAGGING

To find **the most probable tag sequence** given the observation sequence of n words w_1^n, $P(t_1^n|w_1^n)$ is highest.

But $P(t_1^n|w_1^n)$ is difficult to compute and Bayesian classification rule is used:

$$\mathbf{P(x|y) = P(x)\ P(y|x)\ /\ P(y)}$$

- When applied to the sequence of words, **the most probable tag sequence** would be

$$\mathbf{P(t_1^n)\ P(w_1^n|t_1^n)/P(w_1^n)}$$

- where $P(w_1^n)$ does not change and thus do not need to be calculated
- Thus, the **most probable tag sequence** is the product of two probabilites for each possible sequence:
  - **Prior probability of the tag sequence. $P(t_1^n)$ Only one previous tag is considered (bigrams)**
    - Ex. Probability of noun after determiner
  - **Likelihood of the word string. $P(w_1^n|t_1^n)$ Probability of the word given a tag (independent of other words)**
    - Ex. given the tag noun, probabilty of word *dog*