NLP - MODULE 3 - CHAPTER 3

Syntax Analysis

Part Of Speech Tagging

- process of marking up a word in a text as corresponding to a particular POS.
- process of converting a sentence to a form (word, pos tag)

POS Categories

- Open class :
  NOUN, VERB, ADJECTIF, ADVERB
- Closed class :
  PREPOSITION, DETERMENER, CONJUNCTION PRONOUN, PARTICIPLES

Use of POS Tagging

1) Text to Speech
2) Parser
3) Search

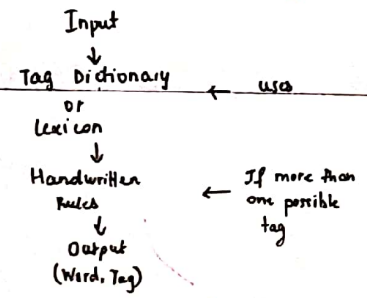
Challenges

- Ambiguity

Methods of POS tagging

1) Rule Based POS Tagging
2) Stochastic POS Tagging
   - Word Frequency Approach
   - Tag Sequence Probabilities

Rule Based POS Tagging



Properties of Rule Based POS Tagging

- Knowledge driven taggers
- information is coded in form of rules
- Rules are built manually
- Approximately 1000 rules
- Smoothing and language modelling is defined explicitly

Stochastic POS Tagging

- model that includes frequency or probability can be called stochastic

1) Word Frequency Approach

- tag encountered most frequently in training set is the one assigned to the ambiguous instance of word

2) Tag Sequence Probabilities

- the best tag for given word is determined by probability that it occurs with the n previous tag
- also called as N-Gram approach

Properties of Stochastic POS Tagging

- Based on probability of tag occuring
- Requires training corpus
- It uses different testing purposes (other than training purposes)
- There would no probability for the words that do not exist in the corpus
- Simplest POS tagging because it choose most frequent tag associated with a word in training corpus

Issues with POS Tagging

- multiple tags & multiple words
- unknow words

CFG

- also called as phrase structure grammar
- Formalism = BNF (Backus-Naur-form)
- CFG consists of
  - → terminals
  - → non terminals
  - → rules

Parsing

- Process of taking a string & grammar and returning a parse tree for that string
- can be viewed as search problem

Types of Parsing

1) Top down parsing
2) Bottom up parsing

Conditional Random Fields (CRF)

- it is a condition probabilistic model for sequence labelling
- CRF are built on logistic Regression classifier