

NLP Module 3 - Syntax Analysis

Part of Speech Tagging

- It is a process of marking up a word in a text as corresponding to a particular part of speech.
- It is a process of converting a sentence to forms.

↓
List of words, List of tuples,
(words, pos tags)

- POS are divided into 2 categories

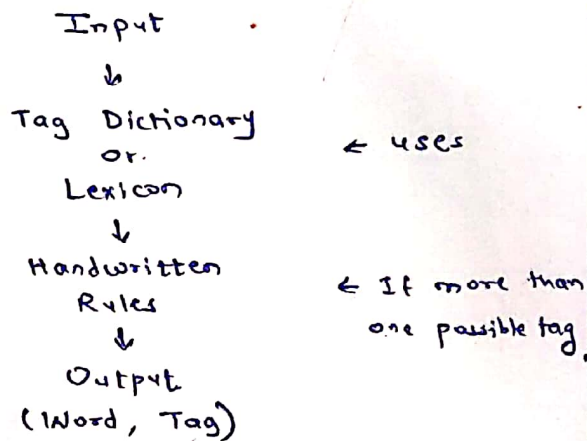
- ① Open class → Noun, verb, adjective, adverb.
- ② Closed class → Preposition, Determiner, Conjunction, Pronouns, Participles

- POS Tagging is majorly used in

- ① Text to speech
- ② Parser
- ③ Search bot

- Challenge → Ambiguity

Rule-based POS Tagging



Properties of Rule-based POS Tagging

- ① These Taggers are knowledge-driven taggers.
- ② Rules are built manually
- ③ Information is coded in the form of rules
- ④ We have approximately 1000 rules
- ⑤ Smoothing and language modeling is defined explicitly in rule-based taggers.

Stochastic POS Tagging

- The model that includes frequency or probability (statistic) can be called stochastic.

- 2 approaches:

① Word Frequency Approach

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

② Tag Sequence Probabilities

- The best tag for a given word is determined by the probability that it occurs with the n previous tags.
- It is also called as n -gram approach.

Properties of Stochastic POS Tagging

- ① The POS tagging is based on the probability of tag occurring
- ② It requires training corpus.
- ③ There would be no probability for the words that do not exist in the corpus
- ④ It uses different testing corpus (other than training corpus).
- ⑤ It is the simplest POS Tagging because it chooses most frequent tags associated with a word in training corpus.