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 bar
- Challenge → Ambiguity

Rule-based POS Tagging

Input
↓
Tag Dictionary or Lexicon ← uses
↓
Handwritten Rules ← If more than one possible tag
↓
Output
(Word, Tag)

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 (statistics) 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.