Types of POS Tagging

- **1. Rules-based POS tagging**
- One of the oldest techniques of tagging is rule-based POS tagging.
- Rule-based taggers use dictionary or lexicon for getting possible tags for tagging each word.
- If the word has more than one possible tag, then rule-based taggers use hand-written rules to identify the correct tag.
- Disambiguation can also be performed in rule-based tagging by analyzing the linguistic features of a word along with its preceding as well as following words.
- For example, suppose if the preceding word of a word is article then word must be a noun.
- Example:
- I am reading a book (noun): consider book as noun rather than verb
- Book(verb) that flight.

- Rule-Based POS Tagging is two architecture:
- In the first Stage, it uses a dictionary to assign each word a list of potential parts-of-speech.
- In the second stage, it uses large lists of hand written disambiguation rules to sort down the list to a single part-of-speech for each word.

- **Properties of Rule-Based POS Tagging**
- Rule-based POS taggers possess the following properties −
- These taggers are knowledge-driven taggers.
- The rules in Rule-based POS tagging are built manually.
- The information is coded in the form of rules.
- We have some limited number of rules approximately around 1000.

- Advantages:
- 1. Small set of simple rules.
- 2. Less stored information.

- Disadvantages:
- 1. Generally less accurate as compared to stochastic taggers.
- 2. Strong language experts team is required.

- **2. Stochastic POS Tagging**
- Another technique of tagging is Stochastic POS Tagging.
- The model that includes frequency or probability (statistics) can be called stochastic.
- The simplest stochastic tagger applies the following approaches for POS tagging −
- Word Frequency Approach
- In this approach, the stochastic taggers disambiguate the words based on the probability that a word occurs with a particular tag.
- If a word can have multiple meanings, the model will assign it the meaning (or tag) that it most frequently had in the training data, assuming this is the most likely interpretation.
- The main issue with this approach is that it may yield inadmissible sequence of tags.

- Tag Sequence Probabilities
- It is another approach of stochastic tagging, where the tagger calculates the probability of a given sequence of tags occurring. It is also called n-gram approach. It is called so because the best tag for a given word is determined by the probability at which it occurs with the n previous tags.

- Properties of Stochastic POST Tagging
- Stochastic POS taggers possess the following properties −
- This POS tagging is based on the probability of tag occurring.
- It requires training corpus
- There would be no probability for the words that do not exist in the corpus.
- It uses different testing corpus (other than training corpus).
- It is the simplest POS tagging because it chooses most frequent tags associated with a word in training corpus.

- 3. Transformation-based Tagging
- This tagger is based on the concept of Transformation-Based Learning (TBL) approach.
- TBL uses supervised learning.
- It combines idea of the rule-based and stochastic taggers.
- Like the rule based taggers, TBL is based on rules that specify what tags should be assigned to what words.
- Like stochastic taggers, TBL is a machine learning technique, in which rules automatically induced from the data.
- It label the training set with most frequent tags.
- Example:
- The can was rusted: can/MD or can/NN
- It uses machine learning as well as grammar rules.
- 1. Modal verb is never preceded by determiner.
- 2. Also in corpus possibility of determiner modal verb pair is zero.
- Therefore can is replaced with Noun tag.