NLP - MODULE 2 - CHAPTER 2

Word Level Analysis

Morphology Analysis

- study of the way words are formed from morphemes
- morphemes → smaller / minimal meaning - bearing units
- morph → form / shape
- ology → study of

Classes of Morphemes

- stem [root word]
- affixes
  - → prefix
  - → infix
  - → suffix

ex: unwell, passed by, killer

Types of Word Formation

1) Inflection
2) Derivation
3) Compounding

Regular Expression

- language used for specifying text search string
- also called as regex

Finite Automata / Finite State Automata

- represented by 5-tuple $(Q, \Sigma, \delta, q_0, F)$

Q = Finite set of states
Σ = Finite set of symbols
δ = Transition Func
Q_0 = Initial State
F = Final State

Types of Finite Automata

- Deterministic FA (DFA)
- $\delta : Q \times \Sigma \rightarrow Q$
- DFA can be represented by diagraphs (state diagrams)
- Non Deterministic FA (NDFA)
- $\delta : Q \times \Sigma \rightarrow 2^Q$

Finite State Transducer

- finite state machine with 2 tapes (i/p tape & o/p tap) unlike FSA which only has one tape
- FSA represents a set of strings
- Ex: {walk, walks, walked}
- FST represents a set of pair of strings (i/p, o/p pairs)
- multi-function device

1) Translator:
   - reads one string on one tape & outputs another string on another tape
2) Recognizer:
   - takes a pair of string as 2 tapes & accepts or rejects based on their matching
3) Generator:
   - outputs a pair of strings on two tapes along with yes or No result based on their matching.
4) Relator:
   - computes the relationship between 2 sets of strings available on 2 tapes

- FST properties
  1) union 2) inversion 3) composition

N-Gram

- continuous sequence of n-items from a given sample of text or speech
- items can be letters, words, paragraphs...
- collected from text or speech corpus

N-Gram Language Model

- predicts probability of given N-Gram within any sequence of words in the language
- a good N-Gram Model can predict the next word in the sentence
- Ex:
  Unigram: {"My", "Name", "is", "Filly"}
  Bigram: {"My name", "Name is", "is Filly"}

N-Gram For Spelling Correction

- N-Gram is without a dictionary → this way employs to find in which position in the correct word the error occured.
- if there is a special way to change incorrect word so that it contains only correct N-Grams, there is as corrections

Inflection

- morphological process that adopts existing words so they function effectively in sentences without changing POS of bat morphemes
- they close off the word, ex: plays
- ∉ in dictionary
- relevant to syntax

- obligatory
- express same concept as base
- semantically regular
- expressed closed to the root.
- meanings are less relevant to meaning of base
- express abstract meaning.

- can only be suffix or infix and not prefix
- ex:
  cat + s = Cats
  N → N

Derivation

- concerned with the way morphemes are connected to existing lexical forms as affixes
- they never close off the word, ex: playful
- ∈ in dictionary
- irrelevant to syntax

- optional
- express a new concept
- semantically irregular
- expressed at the node of word,
- meanings are relevant to meaning of base
- meanings are relatively concrete

- can be prefix and suffix
- ex:
  danger + ous = dangerous
  N → Adj.

Stemming

- Faster because it chops words without knowing the context in which the word is given
- Rule Based Approach
- Accuracy is less

- When we convert any word into root form, then stemming may create non existence meaning of words
- It is preferred when meaning of word is not important for analysis
- Ex: "studies" ⇒ "studi"

Lemmatization

- Slower as compared to stemming but it knows the context of word before proceeding
- Dictionary Based Approach
- Accuracy is more

- Always gives dictionary meaning word while converting into root form
- It is preferred when meaning of word is important for analysis. Ex: QA
- Ex: "studies" ⇒ "study"