NLP

Module 2 - Word Level Analysis

Morphology Analysis

- Morph → Form / Shape
  ology → Study of
- Morphology is the study of the way words are formed from morphemes.
- Morphemes → smallest / minimal meaning-bearing units.
- Example: Cats
  morpheme cat and
  morpheme -s

	Root	Morphological variant.
①	Walk	Walks, Walked, Walking
②	Noise	Noisy
③	atom	Atomic
④	order	reorder, orderly

① Morphemes

- Stem [root word]
- Affix
  - → Prefix
  - → Infix
  - → Suffix

Example
Unwell
Passersby
Killer

② Free and Bound morphemes.

→ Free morphemes are independent words. such as Camera, pen, boat.

③ Free morpheme

(a) Lexical morpheme ← Picture words, ex → Yellow
↳ Unlimited

(b) Grammatical morpheme
↳ Limited

④ Bound Morpheme

- Bound morphemes does not have meaning unless attached with free morphemes.
  eg. -ing
- 2 Types.

(a) Inflection morpheme

Cat + s = Cats
Noun → Noun

It could be Suffix or Infix. [No Prefix]

(b) Derivation morpheme

danger + ous = dangerous
Noun → Adjective

FST properties

① Union ② Intersection ③ Composition

Inflection v/s Derivation

	Inflection	Derivation
①	It is a morphological process that adopts existing words so that they function effectively in sentences without changing pos of base morpheme.	It is concerned with the way morphemes are connected to existing lexical forms as affixes.
②	They close off the word. eg. plays.	They never close off the word. eg. Playful.
③	They cannot be found in dictionaries	They can be found in dictionaries.
④	Inflection is relevant to syntax.	Derivation is irrelevant to syntax.
⑤	It is obligatory	It is optional
⑥	Express the same concept as base.	Expresses a new concept.
⑦	Inflection is semantically regular	Derivation is semantically irregular
⑧	It is expressed closed to the root	It is expressed at the node of words.
⑨	Meanings are less relevant to the meaning of the base	Meanings are relevant to the meaning of the base.
⑩	It expresses a relatively abstract meaning.	Meanings are relatively Concrete.
⑪	Can only be suffix or infix and not prefix	Can be prefix and suffix.
⑫	Example: cat + s = cats (Noun → Noun)	Example: danger + ous = dangerous (Noun → Adjective)

- FST is a multi-function device

① Translator
- It reads one string on one tape and outputs another string.

② Recognizer
- It takes a pair of strings as two tapes and accepts/rejects based on their match

③ Generator
- It outputs a pair of strings on two tapes, along with yes/no result based on matching.

④ Relator
- It computes the relationship between two sets of strings available on two tapes

Stemming

1. Stemming is faster because it chops words without knowing the context the word is given sentences
2. It is a rule based approach.
3. Accuracy is less
4. When we convert any word into root from then stemming may create the non-existence meaning of word.
5. Stemming is preferred when the meaning of the word is not important of analysis.
   Example: Spam Detection.
6. Example:
   "studies" ==> "studi"

Lemmatization

1. Lemmatization is slower as compared to stemming but is knows the context of the word before proceeding
2. It is a dictionary based approach
3. Accuracy is more
4. Lemmatization always gives the dictionary meaning word while converting into root form.
5. Lemmatization would be recommended when the meaning of the word is important for analysis
   Example: Question Answer.
6. Example:
   "studies" => "study"

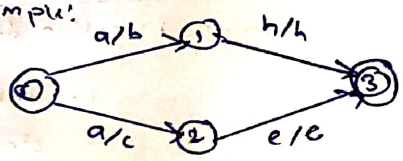
Regular Expression

- It is a language used for specifying text search string.
- Also called as Regex.
- Example:
  1. /[abcd]/ - will match a, b, c & d.
  2. /[3-6]/ - Specifies any one of the digit 3, 4, 5, 6.
  3. /[^A-Z]/ - Not an upper case letter.
     [^] - Not.
  4. * or + - Occurences

Finite Automata / Finite State Automaton

Finite state Transducers.

- It is finite state machines with two steps.
  1. Input tape
  2. Output tape
- Unlike finite state automaton which only has one tape.
- Example:



- FSA represents a set of strings.
  eg. { walk, walks, walked }
  FST represents a set of pairs of strings. (input, output pair).
- FST is a multifunction device
  1. Translator
  2. Recognizer
  3. Generator
  4. Relator

N-Gram

- It is a continuous sequence of n items from a given sample of text or speech.
- The items can be letters, words,
- N-grams are collected from a text or speech corpus.

N-Gram language Model.

- It predicts the probability of a given N-gram within any sequence of words in the language.
- A good N-gram model can predict the next word in the sentence.
- Example:
  1. Unigram
     ("This", "article", "is", "on", "NLP")
  2. Bigram
     ('The article', 'article is', 'is on', 'on NLP')

N-gram for spelling correction.

- N-gram is used without a dictionary, this way employs to find in which position in the incorrect word the error occurs.
- If there is a special way to change the incorrect word so that it contains only correct n-gram, there is as correction.