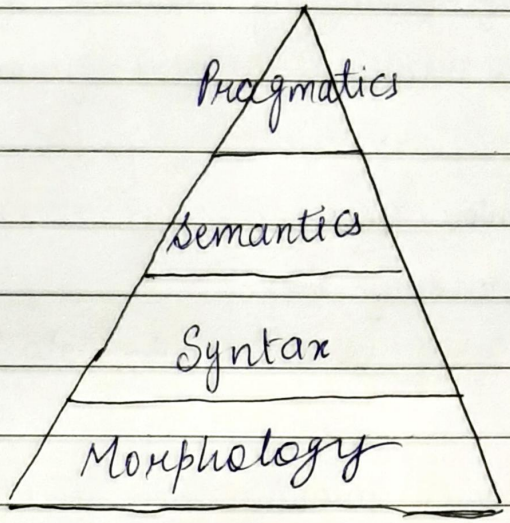
Assignment - 1

1. Describe NLP Pyramids in detail. What are the steps for NLP?

Ans



Natural Language Processing

NLP can also be viewed as a pyramid. The most common NLP tasks build one upon another.

Morphology

It analyses how words are formed, what is their origin, how does their form change depending on the content. In NLP you'll mostly deal with

1. Prefixes / suffixes
2. singularization / pluralization
3. Gender detection
4. word inflection
5. lemmatization
6. Spell checking.

Syntax

Syntax cares about what proper word constructions are. Determining the underlying structure of a sentence or building valid sentences is what syntax is all about.

1. POS tagging.
2. Building Syntax Trees
3. Building Dependency Trees.

Semantics

Semantics derives meaning from text.

1. Named Entity Extraction
2. Relation extraction
3. Semantic Role Labelling
4. Word sense Disambiguation.

Pragmatics

Pragmatics analyses the text as a whole.

1. Coreference / Anaphora resolution.
2. Topic segmentation
3. Lexical chains
4. Summarization

2. What are the different types of ambiguities?

Ans i) Lexical ambiguity :- words have multiple meanings

2) Syntactic ambiguity :- A sentence has multiple parse Trees. Common sources of ambiguity in English are Phrase attachment, Conjunction, Noun group structure

3.) Semantic ambiguity :- Even after the syntax and the meanings of the individual words have been resolved, there are two ways of reading the sentence.

4) Anaphoric ambiguity :- A phrase or words refers to something previously mentioned, but there is more than one possibility.

3. Differentiate b/w Formal and natural language.

Ans.

Natural Language	Formal Language
→ A language that evolved amongst people.	→ A language invented to serve a purpose with well defined syntax & semantics.
→ Such as Mandarin, Chinese, English, Hindi & Spanish.	→ It is in the context of a science, art, & Industry.
~~Business emails~~	
→ Some business correspondence.	→ Business emails.
→ Social media	→ Presentations.
→ Advertising	→ Reports
→ Talking to superiors	→ Talking to peers
→ Team meetings	→ Job interviews

4. What is regular expression, how it helps in NLP?

Ans It is a language for specifying text search strings. It helps us to match or extract other strings or sets of strings, with the help of a specialized syntax present in a pattern.

Regular expressions are used in various tasks such as :-

- Data pre-processing
- Rule-based information mining systems.
- Pattern matching
- Text features engineering
- Web scraping

In NLP, we can use Regular expressions at many places such as,

1. To validate data fields.
2. To filter a particular text from the whole corpus.
3. To identify particular strings in a text.
4. To convert the output of once processing component into the format required for a second component.

Q.5 Write regular expressions for the following languages :-

a) The set of all alphabetic strings.

Ans /[a-z A-z]+/

b) the set of all lowercase alphabetic strings ending in a b.

/[a-z]*b/

c) The set of all strings with two consecutive repeated words.

Ans /([a-z A-Z]+)␣\1/

d) The set of all strings from the alphabet a,b such that each a is immediately preceeded and immediately followed by a,b.

Ans /b+(ab+)+/

e) write a pattern which places the first word of an English sentence in a register.

Ans /^[^a-z A-Z]* ([a-z A-Z]+)/

Assignment - 2

1. What is Tokenization?

Ans Tokenization is the proces of breaking down a piece of text into small units called tokens. A token may be a word, part of a word or just characters like punchation. It is one of the most foundational NLP task and a difficult one, because every language has its own grammatical constructs, which are often difficult to write down as rules. It defines what our NLP models can express. Even though tokenization is super important, it's not always top of mind. A token is a string with a known meaning.

2. Explain different types of morphology?

Ans Morphology is the study of morphemes, the smallest meaningful unit in a language. Morphemes can transform a word from one grammatical category to another, such as dance - a verb to a noun. Free lexical morphemes exist as independent words, such as zebra.

A prefix is a morphemic unit that is attached to the beginning of a base word to give it a different meaning, and there

are dozens of prefixes in English. The prefix un changes a meaning to its opposite, such as unavoidable, unforgiving & unfair.

Morphology calls morphemes that are fixed onto the ends of words suffixes. Like prefixes, they too alter the base of word's meaning.

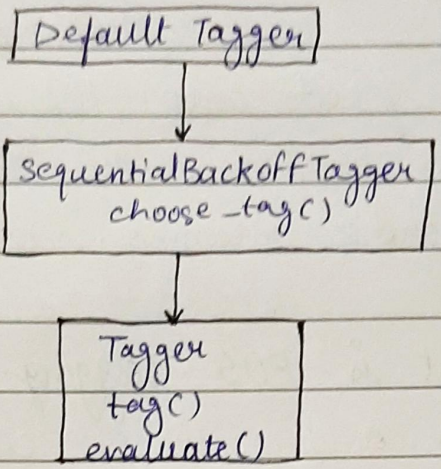
3. What is POS tagging? Explain.

Ans It is a process of converting a sentence to forms - list of words, list of tuples. The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb.

Part-of-speech	Tag
Noun	n
Verb	v
Adjective	a
Adverb	r

Default tagging is a basic step for the part-of-speech tagging. It is performed using the defaultTagger class. The defaultTagger class takes 'tag' as a single argument. NN is the tag for a singular Noun. It is most useful

when it gets to work with most common part-of-speech tag.



4. Describe finite state transducers.

Ans A finite-state transducer (FST) is a finite-state machine with two memory tapes, following the terminology for Turing machines: an input tape and an output tape. This contrasts with an ordinary finite-state automaton, which has a single tape. An FST is a type of finite-state automaton that maps between two sets of symbols. An FST is more general than an FSA. An FSA defines a formal language by defining a set of accepted strings, while an FST defines relations between set of strings.

An FST will read a set of strings on the input tape and generates a set of

relations on the output tape. An FST can be thought of as a translator or relater between strings in a set.

In morphological parsing, an example would be inputting a string of letters into the FST, the FST would then output a string of morphemes.

5. What is transformation based tagging?

Ans Transformation based tagging is also called Brill tagging. It is an instance of the transformation based learning, which is a rule-based algorithm for automatic tagging of POS to the given text. TBL, allows us to have linguistic knowledge in a readable form, transforms one state to another state by using transformation rules.

Working of TBL :-

Consider the following steps to understand the working of TBL -

- Start with the solution.
- Most beneficial transformation chosen.
- Apply to the problem.

The advantages of TBL are as follows - is much faster than marked - model Tagger.

The disadvantages of TBL is that does not provide tag probabilities.

6. What is need of speech Processing?

Ans Speech processing is a discipline of computer science that deals with designing computer systems that recognize spoken words. Speech processing and NLP allow intelligent devices, such as smartphones, to interact with users via verbal language.

7. Define Phondegical rules.

Ans A phonological rule is a formal way of expressing a systematic phonological or morphonological process or diachronic sound change in language.

Phonological rules are commonly used in generative phonology as a notation to capture sound-related operations and computations the human brain performs wh producing or comprehending spoken language

Rules are the way phonologists predict how a speech sound will change depending on its position in various speech environments.

8. Describe Bayesian methods of pronounciation variation.

Ans The Bayesian method for pronounciation: The bayesian algorithm can be used to solve pronounciation sub problem in speech recognition.

Pronounciation sub problem:

Given a series of phones, compute the most probable word that generated them.

Select a single word such that P(word/observation) is highest.

$$\hat{w} = \arg\max_{w \in V} \frac{\overbrace{P(O|w)}^{\text{likelihood}} \overbrace{P(w)}^{\text{prior}}}{P(O)}$$