Natural Language Processing(NLP)
(IT401)

Prepared by

Mr. Umesh B. Sangule
Assistant Professor

Department of Information Technology

Unit-I

INTRODUCTION TO NATURAL LANGUAGE PROCESSING (NLP)

Course Objectives : *To introduce the Natural Language Processing (NLP) basics and Basics of linguistics.*

Course Outcome(CO1) : *Understand concept and processing of NLP.*

Content

- Introduction to NLP
- Ambiguity in NLP
- Automata for NLP
- Stages of NLP
- Challenges and Issues in NLP
- Basics of Text Processing

Introduction

- Language is used to shape the thoughts; it has structure and also it carries a meaning. By using our language, we naturally learn the new concepts and hardly realise how we process this natural language.

- Natural Language Processing is the process of computer analysis of input provided in a human language, and conversion of this input into a useful form of representation.

- Natural language processing is concerned with the development of computational models of aspects of human language processing.

Cont....

- The following are the two main reasons for such developments.

  1) Develop automatic tool for natural language processing.

  2) Gain better understanding of human communication.

- When we build computational models by using human language, we need processing abilities where this processing abilities and incorporates how human collects, Store and process the language. It also needed a knowledge of the world and of language.

- The input and output of the NLP is text or speech.

Cont....

- Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that deals with the interaction between computers and human languages.

- NLP is used to analyze, understand and generate natural language text and speech.

- The goal of NLP is to enable computers to understand and interpret human language in a way that is similar to how humans process language.

Cont....

Natural Language Processing (NLP) techniques are used in wide ranges of applications:

- *Speech Recognition and Transcription*: NLP techniques are used to convert speech to text, which is useful for tasks such as dictation and voice-controlled assistants.

- *Language Translation*: NLP techniques are used to translate text from one language to another, which is useful for tasks such as global communication and e-commerce.

- *Text Summarization*: NLP techniques are used to summarize long text documents into shorter versions, which is useful for tasks such as news summarization and document indexing.

Cont....

- *Sentiment Analysis*: NLP techniques are used to determine the sentiment or emotion expressed in text, which is useful for tasks such as customer feedback analysis and social media monitoring

- *Question Answering*: NLP techniques are used to answer questions asked in natural language, which is useful for tasks such as chatbots and virtual assistants.

Ambiguity in NLP

- NLP is hard because Ambiguity and Uncertainty exist in the language.
- Ambiguity means not having well defined solution.
- Any sentence in a language with a large-enough grammar can have another interpretation.
- There are various forms of ambiguity related to natural language and they are:

1. Lexical Ambiguity

2. Syntactic Ambiguity

3. Semantic Ambiguity

4.Metonymy Ambiguity

Ambiguity in NLP

she won two silver medals.
she made a silver speech.

1) Lexical Ambiguity:

- When words have multiple assertion then it is known as lexical ambiguity.

- The word “back” can be a noun or an adjective.

  Noun: back stage

  Adjective: back door

Ambiguity in NLP

2) Syntactic Ambiguity: I saw a girl on the beach with telescope.

- Syntactic ambiguity means sentences are parsed in multiple syntactical forms or A sentence can be parsed in different ways.

- For example:

"I saw the boy on the beach with binoculars"

"I saw the boy on the beach with my binoculars"

In this sentence, confusion in meaning is created. The phrase with my binoculars could modify the verb, saw or the noun, boy.

Ambiguity in NLP

The meaning is not clear from this sentence

Ram loves his mother and shreya does too.

3) Semantic Ambiguity:

- Semantic ambiguity is related to the sentence interpretation.

- For example:

"I saw the boy on the beach with my binoculars"

The sentence means that I saw a boy through my binoculars or the boy had my binoculars with him.

Ambiguity in NLP

4) Metonymy Ambiguity:

- Metonymy ambiguity is related to phrases in which literal meaning is different from the figurative assertion.
- It is most difficult ambiguity.

- For example:

"Nokia us screaming for new management"

Here it really doesn't mean that the company is literally screaming.

Programming Language Vs Natural Language

***Natural Language*:**

- The vocabulary of natural language is incredibly extensive.
- Humans are capable of understanding natural language.
- Natural Language is inherently ambiguous.

***Programming Language*:**

- Very few words are used in computer language.
- The machines can easily understand computer language.
- Computer Language is Unambiguous.

Programming Language Vs Natural Language

***Natural Language*:**

- Are open and allow combinations without risk of making mistakes.
- Human beings have ability to clarify meaning of expression.

***Programming Language*:**

- Are closed and Fixed to avoid confusion and mistakes.
- Computers are very precise about instructions they receive.

Finite Automata for NLP

- An Automaton having finite number of states is names as Finite Automata (FA) or finite State Automata(FSA).
- Finite Automata is use to identify the patterns.
- It takes strings of symbol as input and changes its state accordingly, when required symbol is found, then the transition happens.
- In Finite Automata either “Accept” or “Reject” state.

Finite Automata for NLP

- Automata can be represented by 5-tuple (Q, Σ, δ, q0, F)

Where,

Q : is a finite set of states

Σ : is a finite set of symbols, called the alphabet of the automaton

δ : is the transition function

q0 : is the initial state from where any input is processed

F : is a set of final state/states

Finite Automata for NLP

- There are two types of Finite Automata

    1. Deterministic Finite automation (DFA)
    2. Non-deterministic Finite Automation (NDFA)

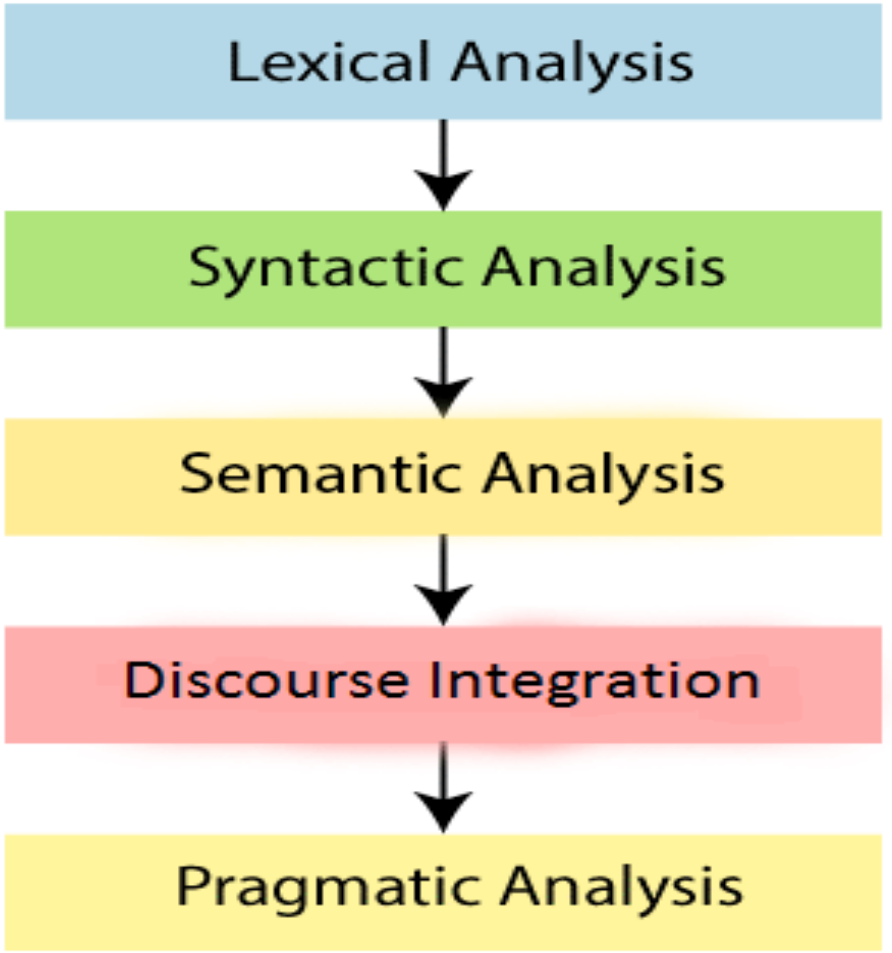
- ***Deterministic Finite Automation (DFA)***:
- Definition: It may be defined as the type of finite automation wherein, for every input symbol we can determine the state to which the machine will move. It has a finite number of states that is why the machine is called **Deterministic Finite Automation**(DFA)

Finite Automata for NLP

- ***Non-Deterministic Finite Automation (NDFA)***:
- Definition: It may be defined as the type of finite automation where, for each input symbol we cannot determine the state to which the machine will move. (i.e. Machine can move to any combination of the states).
- It mean for each state there can be more than one transition or symbol.
- It has a finite number of states that is why the machine is called

Non-Deterministic Finite Automation(NDFA)

Stages of NLP



Stages of NLP

1) Lexical Analysis:

- Lexical Analysis is the first stage in NLP. It is also known as morphological analysis.
- At this stage the structure of the words is identified and analysed.
- Lexicon of a language means the collection of words and phrases in a language.
- Lexical analysis is dividing the whole portion of text into paragraphs, sentences, and words.

2) Syntactic Analysis:

- It involves analysis of words in the sentence for grammar and ordering words in a way that shows the relationship among the words.

- Example:

  The sentence such as “The school goes to girl” is rejected by English syntactic analyser.

3) ***Semantic Analysis:***

- Semantic analysis draws the exact meaning or the dictionary meaning from the text.
- The text is checked for meaningfulness.
- It is done by mapping syntactic structures and objects in the task domain.
- Example:

  The semantic analyser neglects sentence such as "hot ice-cream".

4) Discourse Integration:

- The meaning of any sentence depends upon the meaning of the sentence just before it.
- Furthermore, it also brings about the meaning of immediately following sentence.
- Example:

  “Meena is a girl, she goes to school” here "she" is a dependency pointing to Meena.

5) ***Pragmatic Analysis:***

- During this, what was said is re-interpreted on what it truly meant.
- It contains deriving those aspects of language which necessitate real world knowledge.
- Example:

  "John saw Mary in a garden with a cat"

  here we can't say that John is with cat or Mary is with cat

Steps to build NLP Pipeline

Step 1: Sentence Segmentation:

- Sentence Segment is the first step for building the NLP pipeline. It breaks the paragraph into separate sentences.

- Example: "*Independence Day is one of the important festivals for every Indian citizen. It is celebrated on the 15th of August each year ever since India got independence from the British rule.*"

- Sentence Segment Produces the Following Result:

  "*Independence Day is one of the important festivals for every Indian citizen.*"

  "*It is celebrated on the 15th of August each year ever since India got independence from the British rule.*"

Steps to build NLP Pipeline

Step 2: Tokenization:

- Word Tokenizer is used to break the sentence into separate words or tokens.

- Example:

  "*JavaTpoint offers Corporate Training, Summer Training, Online Training and Winter Training.*"

- Word Tokenizer generates the following result:

  "JavaTpoint", "offers", "Corporate", "Training", "Summer", "Training", "Online", "Training", "and", "Winter", "Training"

Steps to build NLP Pipeline

Step 3: Stemming:

- Stemming is used to normalize words into its base form or root form.
- For example: celebrates, celebrated and celebrating, all these words are originated with a single root word "celebrate."
- The big problem with stemming is that sometimes it produces the root word which may not have any meaning.
- Example:

  "Intelligence, intelligent and intelligently, all these words are originated with a single root word "intelligen." In English, the word "intelligen" do not have any meaning."

Step 4: Lemmatization:

- Lemmatization is quite similar to the Stemming. It is used to group different inflected forms of the word, called Lemma.

- The main difference between Stemming and lemmatization is that it produces the root word, which has a meaning.

- Example:

  "*In lemmatization, the words intelligence, intelligent and intelligently has a root word intelligent, which has a meaning.*"

Step 5: Identifying Stop Words:

- In English, there are a lot of words that appear very frequently like "is", "and", "the" and "a".
- NLP pipelines will flag these words as stop words. Stop words might be filtered out before doing any statistical analysis.
- Example:

  "*He is a good boy*."

Steps to build NLP Pipeline

Step 6: Dependency Parsing:

- Dependency Parsing is used to find that how all the words in the sentence are related to each other.

Step 7: POS Tags:

- POS stands for parts of speech, which includes Noun, verb, adverb and Adjective. It indicates that how a word functions with its meaning as well as grammatically within the sentences.
- Example:

  "Google" something on the Internet.

  In the above example, Google is used as a verb, although it is a proper noun.

Steps to build NLP Pipeline

Step 8: Named Entity Recognition (NER):

- Named Entity Recognition (NER) is the process of detecting the named entity such as person name, movie name organization name or location.

- Example:

  "Steve Jobs introduced iPhone at the Macworld Conference in San Francisco, California."

  In the above example, "Steve Jobs" is person name,

  "San Francisco" is Location,

Step 9: Chunking:

- Chunking is used to collect the individual piece of information and grouping them into bigger pieces of sentences

Challenges and Issues in NLP

1) Contextual words and phrases and homonyms:

- The same words and phrases can have **diverse meanings** according the context of a sentence and many words have the exact same pronunciation but completely different meanings

- Example:

  " I **ran** to the store because we **ran** out of milk."

  " The house is looking really **run** down."

- In the above sentences the meaning of the run is different according to the context

Challenges and Issues in NLP

- ***Homonyms*** means the pronunciation of two or more words is same but have different meaning.

- For example:

  “their” and “there”,
  “right” and “write”,
  “Know” and “No”

- This will create problem in question answering and speech-to-text applications.

Challenges and Issues in NLP

2) Synonyms:

- Synonyms can cause issues like contextual understanding since we use many different words to express the identical idea.

- Additionally, some of these words may convey exactly the same meaning, while some may be levels of complexity and different people use synonyms to denote slightly different meanings within their personal vocabulary.

- Example:

  " **small**, **little**, **tiny**, **minute** have same meaning"

Challenges and Issues in NLP

3) <u>Ambiguity</u>:

- Ambiguity in NLP refers to sentences and phrases that potentially have two or more possible interpretations.

- There is lexical, syntactic and semantic ambiguity.

Challenges and Issues in NLP

4) <u>Errors in text or speech</u>:

- Misspelled or misused words can generate problems for text analysis.
- Autocorrect and grammar correction applications can handle common mistakes, but do not at all times understand the writer's intention.
- With spoken language it is difficult for the machine to understand mispronunciations, different accents, stammers, etc.

Challenges and Issues in NLP

5) <u>Idioms and Slang</u>:

- Informal phrases, expressions, idioms, and culture-specific lingo present a number of problems for NLP especially for models intended for comprehensive use.

- Because as formal language, idioms may have no dictionary definition at all, and these expressions may ever have different meanings in different geographic areas.

- Cultural slang is continuously morphing and increasing, so new words arise every day.

Challenges and Issues in NLP

6) <u>Domain Specific Language</u>:

- Different businesses and industries often use very different language.

- Example:

  "An NLP processing model needed for healthcare would be very different than one used to process legal documents."

Challenges and Issues in NLP

7) <u>Low-resource languages</u>:

- Artificial Intelligence, machine learning and NLP applications have been mostly built for the most common, widely used languages.

- However, many languages especially those spoken by people with less access to technology often go overlooked and under processed.

- Example:

  "There are over 3,000 languages in Africa, alone.
  There simply isn't ample data on many of these languages."


THANK YOU