MIS4311
Machine Learning Applications

Spring 2025

Lecture #10

Natural Language Processing

Processing of Natural Language is required when you want an intelligent system like robot to perform as per your instructions, when you want to **hear decision from a dialogue** based clinical expert system, etc.

The input and output of an NLP system can be

- Speech
- Written Text

Components of NLP

There are two components of NLP :

Natural Language Understanding (NLU)

- Mapping the given input in natural language into useful representations.
- Analyzing different aspects of the language.

Natural Language Generation (NLG)

It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation. It involves:

- **Text planning** − This includes retrieving the relevant content from the knowledge base.
- **Sentence planning** − This includes choosing the required words, forming meaningful phrases, setting tone of the sentence.
- **Text Realization** − This is mapping sentence plan into sentence structure.

Difficulties in NLU

The NLU is very rich in form and structure; however, it is ambiguous (uncertain or indefinite). There can be different levels of ambiguity:

Lexical ambiguity:

It is at a very primitive level such as the word-level. For example, treating the word “board” as noun or verb?

Syntax level ambiguity

A sentence can be parsed in different ways. For example, “He lifted the beetle with red cap.” − Did he use cap to lift the beetle or he lifted a beetle that had red cap?

Referential ambiguity

Referring to something using pronouns. For example, Rima went to Gauri. She said, “I am tired.” − Exactly who is tired?

NLP Terminology

Phonology − It is study of organizing sound systematically.
Morphology − It is a study of construction of words from primitive meaningful units.
Morpheme − It is a primitive unit of meaning in a language.
Syntax − It refers to arranging words to make a sentence. It also involves determining the structural role of words in the sentence and in phrases.
Semantics − It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.
Pragmatics − It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected.
Discourse − It deals with how the immediately preceding sentence can affect the interpretation of the next sentence.
World Knowledge − It includes the general knowledge about the world.

Steps in NLP

Lexical Analysis:

It involves identifying and **analyzing the structure of words**. Lexicon of a language **means the collection of words and phrases in a language**. Lexical analysis is **dividing the whole chunk of text** into paragraphs, sentences, and words.

Syntactic Analysis (Parsing):

It involves analysis of words in the sentence **for grammar and arranging words** in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

Semantic Analysis:

It 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. The semantic analyzer disregards sentence such as “hot ice-cream”.

Steps in NLP

Discourse Integration:

The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

Pragmatic Analysis:

During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

NLP with Python

To build such applications we will use the Python package called NLTK (Natural Language Toolkit Package).

conda install -c anaconda nltk

>>> import nltk

Downloading NLTK's Data

>>> nltk.download()

Installing Other Necessary Packages

pip install genism # It is a robust semantic modeling library

pip install pattern # It is used to make **genism** package work properly

Concept of Tokenization, Stemming, and Lemmatization

Tokenization:

It may be defined as the process of breaking the given text i.e. the character sequence into smaller units called tokens. The tokens may be the words, numbers or punctuation marks.

Divide the input text into **sentences** by

from nltk.tokenize import sent_tokenize

sample="This is a sample sentence. This is second sample sentence."

Output: ['This is a sample sentence.', 'This is second sample sentence.']

Divide the input text into **words** by

from nltk.tokenize import word_tokenize

Output: ['This', 'is', 'a', 'sample', 'sentence', '.', 'This', 'is', 'second', 'sample', 'sentence', '.']

Concept of Tokenization, Stemming, and Lemmatization

Stemming : stemming is the heuristic process of extracting the base forms of the words.
For example: ***democracy***, ***democratic***, and ***democratization***

In the Python NLTK module, we have different packages related to stemming. These packages can be used to get the base forms of word.

from nltk.stem.porter import ***PorterStemmer***

from nltk.stem.lancaster import ***LancasterStemmer***

from nltk.stem.snowball import ***SnowballStemmer***

Concept of Tokenization, Stemming, and Lemmatization

Lemmatization: It extracts the base form of words by lemmatization. It basically does this task with the **use of a vocabulary and morphological analysis** of words, normally aiming to remove inflectional endings only. This kind of base form of any word is called lemma.

WordNetLemmatizer package

This Python package will extract the base form of the word depending upon whether it is used as a noun or as a verb.

*from nltk.stem import **WordNetLemmatizer***

rocks : rock

corpora : corpus

better : good

larger : large

worst : bad

Natural Language Processing

Bag of Word (BoW) Model : It is used to extract the features from text so that the text can be used in modeling such that in machine learning algorithms.

The conversion of text data into numeric data is called **feature extraction** or feature encoding.

- **Sentence 1:** We are using the Bag of Words model.
- **Sentence 2:** Bag of Words model is used for extracting the features.

- **Sentence 1:** [1,1,1,1,1,1,1,1,0,0,0,0,0]
- **Sentence 2:** [0,0,0,1,1,1,1,1,1,1,1,1,1]

Natural Language Processing

Gender Finder Application:

In this problem statement, a classifier would be trained to find the gender (male or female) by providing the names.

Next Week

- **Genetic Algorithms**

Thank you for your participation ☺