Natural Language Processing (NLP)

Regarding Course

Instructor: Sourav Kumar Dandapat(Sourav@iitp.ac.in)

Teaching Assistant:

- Arpan Phukan (arpan_2121cs33@iitp.ac.in)
- Sudhir Kumar (sudhir_2221cs14@iitp.ac.in)
- Pankaj Kumar Paswan (pankaj_2411ai49@iitp.ac.in)
- Suman Hazra (suman_2411ai06@iitp.ac.in)
- Tanmay Pawar (tanmay_2411ai07@iitp.ac.in)
- Aman Kumar (aman_2411ai53@iitp.ac.in)

Class Timing:

- Wednesday (12 pm-12.55 pm)
- Thursday (9 am-9.55 am)
- Friday (10 am-10.55 am)

Course Page: 10.22.10.100/~sourav/nlp_autumn_2025/

Evaluation Plan:

- 30% class evaluation
- 30% Mid sem
- 40% End sem

Tentative Dates for Quizes

- 2 quizzes before midsem (19/08, 16/09) [best of 2]
- 2 quizzes/presentation/projects after midsem [will be announced]

Text Book

- Speech and Language Processing (Daniel Jurafsky)

Introduction

- NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence.
- Human communicate through some form of language either by text or speech.
- To make interactions between computers and humans, computers need to understand natural languages used by humans.
- Natural language processing is all about making computers learn, understand, analyse, manipulate and interpret natural(human) languages.
- 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 ability of machines to interpret human language is now at the core of many applications that we use every day - chatbots, Email classification and spam filters, search engines, grammar checkers, voice assistants, and social language translators.
- The input and output of an NLP system can be Speech or Written Text

Why Natural Language Processing (NLP)

Natural Language Processing (NLP) is one of the hottest areas of artificial intelligence (AI) thanks to applications like

text generators that compose coherent essays,

chatbots that fool people into thinking they're scientiest,

and text-to-image programs that produce photorealistic images of anything you can describe.

Recent years have brought a revolution in the ability of computers to understand human languages, programming languages, and even biological and chemical sequences, such as DNA and protein structures, that resemble language.

The latest AI models are unlocking these areas to analyze the meanings of input text and generate meaningful, expressive output.

The process of computer analysis of input provided in a human language (natural language), and conversion of this input into a useful form of representation.

The field of NLP is primarily concerned with getting computers to perform useful and interesting tasks with human languages.

Some salient points

- Makes human-computer interaction more natural
- Powers translation, search engines, chatbots, and more
- Helps analyze massive amounts of text

Real-World Applications

- Chatbots & virtual assistants (e.g., Siri, Alexa)
- Machine translation (e.g., Google Translate)
- Sentiment analysis (e.g., social media monitoring)
- Text summarization

Forms of Natural Language

The input/output of a NLP system can be:

- **written text**
- **speech**

We will mostly concerned with written text (not speech).

To process written text, we need:

- **lexical, syntactic, semantic knowledge about the language**
- **discourse information, real world knowledge**

To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.

Components of NLP

Natural Language Understanding

- Mapping the given input in the natural language into a useful representation.
- Different level of analysis required:

  morphological analysis,
  syntactic analysis,
  semantic analysis,
  discourse analysis, …

Natural Language Generation

- Producing output in the natural language from some internal representation.
- Different level of synthesis required:

  deep planning (what to say),
  syntactic generation

Why NL Understanding is hard?

Natural language is extremely rich in form and structure, and **very ambiguous**.

- How to represent meaning

One input can mean many different things. Ambiguity can be at different levels.

- Lexical (word level) ambiguity -- different meanings of words
- Syntactic ambiguity -- different ways to parse the sentence
- Interpreting partial information -- how to interpret pronouns
- Contextual information -- context of the sentence may affect the meaning of that sentence.

Many input can mean the same thing.

Interaction among components of the input is not clear.

Knowledge of Language

Phonology – concerns how words are related to the sounds that realize them.

Morphology – concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive meaning bearing unit of a language.

Syntax – concerns how words can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.

Semantics – concerns what words mean and how these meaning combine in sentences to form a meaningful sentence. The study of context-independent meaning.

Pragmatics – concerns how sentences are used in different situations/contexts and how use affects the interpretation of the sentence.

Discourse – concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, interpreting pronouns and interpreting the temporal aspects of the information.

World Knowledge – includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

The process of text analytics involves three stages as given below:

Lexical processing: In this stage, we do basic text pre-processing and text cleaning such as tokenization, stemming, lemmatization, correcting spellings, etc.

Syntactic processing: In this step, we extract more meaning from the sentence, by using its syntax this time. Instead of just blindly looking at the words, we here look at the syntactic structures, i.e., the grammar of the language to understand the meaning.

Semantic processing: Lexical and syntactic processing do not suffice when it comes to building advanced NLP applications such as language translation, chatbots, etc. After performing lexical and syntactic processing, we will still be incapable of understanding the meaning of each word. Here, we try and extract the hidden meaning behind the words which is also the most difficult part for computers.

Lexical Processing

Lexicon describes the vocabulary that makes up a language. Lexical analysis deciphers and segments language into units or lexemes such as paragraphs, sentences, phrases, and words. A few of the techniques involved in Lexical Processing are:

- Word Frequencies and Stop Words
- Stop words removal
- Bag-of-Words and TF-IDF Representation
- Tokenization
- Stemming
- Lemmatization

Syntactic Processing

It is about analysing the syntax or the grammatical structure of sentences.

Following are some of the popular techniques performed for the syntactic processing of textual data:

- POS tagging techniques
- Constituency and Dependency parsing

Let’s start with an example to understand Syntactic Processing and consider two sentences "Canberra is the capital of Australia." and "Is Canberra the of Australia capital.”

Both sentences have the same set of words

However, only the first one is syntactically correct and comprehensible.

Lexical processing techniques wouldn't be able to tell this difference.

Therefore, more sophisticated syntactic processing techniques are required to understand the relationship between individual words in the sentence.

Semantic Processing

Lexical and syntactic processing doesn't suffice when it comes to building advanced NLP applications such as language translation, chatbots, etc.

Semantic processing is about understanding the meaning of a given piece of text.

It is probably the most challenging area in the field of NLP, partly because the concept of 'meaning' itself is quite wide, and it is a genuinely hard problem to make machines understand the text the same way as we humans do

Such as inferring the intent of a statement, meanings of ambiguous words, dealing with synonyms, detecting sarcasm and so on.

Semantic text processing focuses on teaching machines to process meaning of the text in similar ways. There are various semantics techniques used such as:

Word Sense Disambiguation: Identifying the intended meaning of an ambiguous word.

Distributional Semantics: The technique helps to arrange semantically similar words together as compared to other words.

Topic modelling: Identifying topics being talked about in documents.

Ambiguity

I made her duck.

- How many different interpretations does this sentence have?
- What are the reasons for the ambiguity?
- The categories of knowledge of language can be thought of as ambiguity resolving components.
- How can each ambiguous piece be resolved?
- Does speech input make the sentence even more ambiguous?

Some interpretations of : **`I made her duck.`**

1. I cooked *duck* for her.
2. I cooked *duck* belonging to her.
3. I created a toy duck which she owns.
4. I caused her to quickly lower her head or body.
5. I used magic and turned her into a *duck*.

Brief History of NLP

1940s –1950s: Foundations
- Development of formal language theory (Chomsky, Backus, Naur, Kleene)
- Probabilities and information theory (Shannon)

1957 – 1970s:
- Use of formal grammars as basis for natural language processing (Chomsky, Kaplan)
- Use of logic and logic based programming (Minsky, Winograd, Colmerauer, Kay)

1970s – 1983:
- Probabilistic methods for speech recognition (Jelinek, Mercer)
- Discourse modeling (Grosz, Sidner, Hobbs)

1983 – 1993:
- Finite state models (morphology) (Kaplan, Kay)

1993 – present:
- Strong integration of different techniques, different areas.

Topics to be covered:

Regular Expression

Lexical Analysis

Edit Distance

N-Gram Language Model

Naïve Bayes

Logistic Regression

Vector Semantics

Neural Network

RNN And LSTM

Transformer

Large Language Model

Masked Language Model

Prompting

Machine Translation

Question Answering

Dialogue Management

Part Of Speech Tagging

Constituency Parsing

Dependency Parsing

Named Entity Recognition