Introduction to Natural Language Processing

Broadly, Natural Language Processing (or NLP for short) consists of developing a set of algorithms and tools so that machines can make sense of data available in natural (human) languages such as Hindi, English, German, French, etc.

Natural language: refers to the medium in which humans communicate with each other.

* Writing (text): emails, articles, news, blogs, bank documents, etc
* Speech: lectures, speeches, audio calls, etc.

→ There are trillions of web pages full of natural text, so imagine the scale of data available today.

**Applications of NLP:**

* Document Classification deals with classifying textual documents and assigning it one or multiple categories.

Example:

* classifying news articles into categories like sports, politics, business, technology, etc
* segregating different types of invoices and sales deeds in a large company.
* Document Clustering is used to find similar documents and segregate them to form groups. Documents that are closely related will be part of the same group.

Example:

* finding similar questions that have already posted in a forum
* finding new published medical research related to a patient's symptoms.
* Sentiment Analysis is used to classify text for different sentiments ranging from negative to neutral to positive

Example:

* commonly used to understand customer opinions from product reviews or their posts on social media.
* Document Summarization helps to extract the most important and central ideas in a document.

Example:

* one could train a model to summarize a 3000-word article to 200 words.
* Question Answering systems are intelligent systems that generate responses to the questions being asked by the user.

Example:

* Many conversational AI and personal assistant solutions (for example Amazon Alexa) are able to perform question answering.
* Machine Translation is the task of automatically translating from one natural language to another.

Example:

* Google translate is performing when you visit a website that is written in a language you do not understand.

**Components of NLP**

There are 2 main components of natural language processing:

* Natural Language Understanding (NLU)

NLU enables machines to understand the intent or the meaning of the text.

The different levels of analysis that NLU requires are as follows:

* **Morphological Analysis** is the analysis of the structure of individual words. A morpheme is defined as the "minimal unit of meaning"
* **Syntactic Analysis** (also called parsing) involves analysis of words in the sentence for grammar.
* **Semantic Analysis** uses morphological and syntactic knowledge to understand the meaning, intent, and purpose of the text as a whole.
* **Discourse Analysis** is a more advanced stage of NLU where we perform syntactic or semantic analysis in a longer piece of text.
* Natural Language Generation (NLG)

Once the machine understands the natural language, NLG is used to respond in natural language, or to produce written text.

Applications include chat-bots and personal assistants like Alexa and Siri.

* **Content Determination** involves deciding what information we need to convey in the generated text.
* **Planning / Micro-planning** involves finding, mixing and merging different sentence representations into more concise representation.
* **Deep Learning** which has proven very successful in applications that require language generation, such as translation and question answering.

NOTE: Whereas the above two approaches involve lots of hand-crafted rules, deep learning can be used in an 'end-to-end' fashion.