“Natural Language Processing”

by

Alber Moied

for

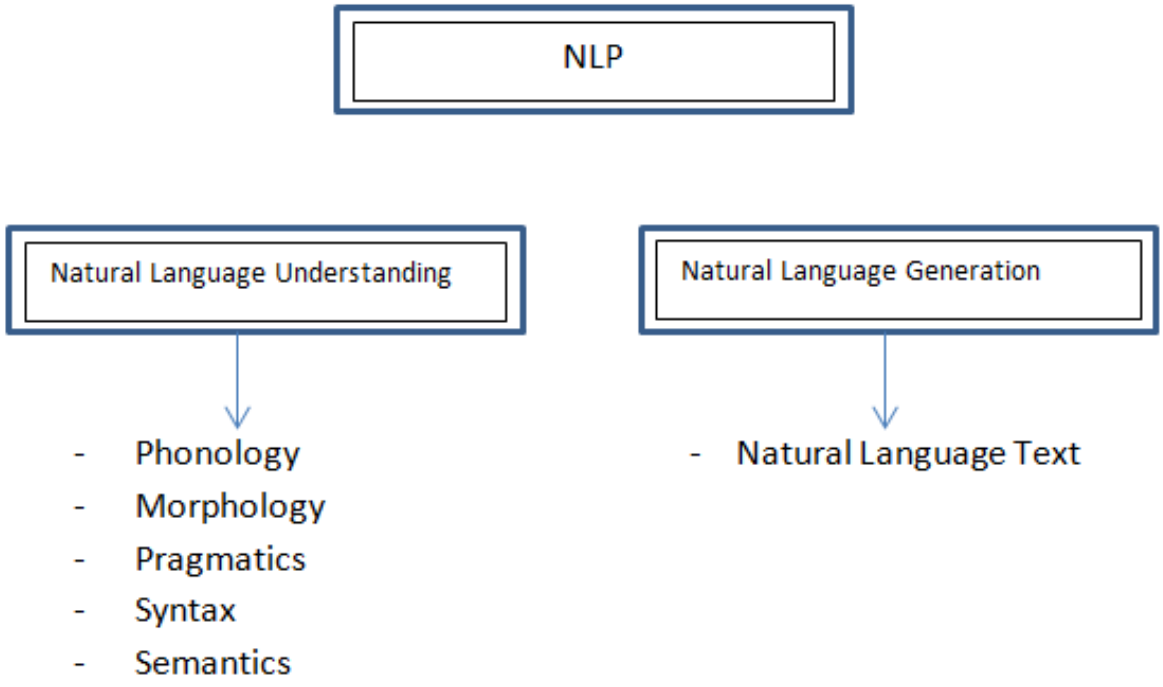
COMP840 Fall 2019

1. Introduction

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. Most NLP techniques rely on machine learning to derive meaning from human languages.

It is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human natural languages, in particular how to program computers to process and analyze large amounts of natural language data. It came into existence to ease the user's work and to satisfy the wish to communicate with the computer in natural language. Since all the users may not be well-versed in machine specific language, NLP caters those users who do not have enough time to learn new languages or get perfection in it. NLP is classified into two areas:

- Natural language understanding
- Natural language generation

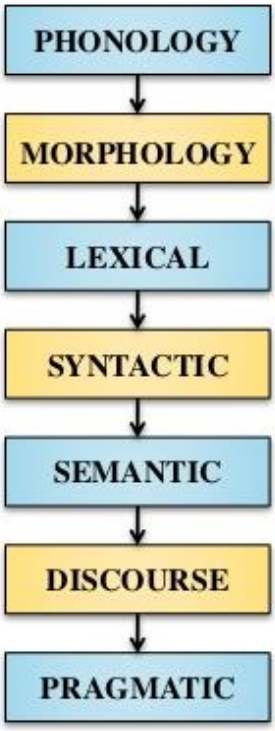


Broad classification of NLP

Linguistics is the science of language which includes Phonology that refers to sound, Morphology word formation, Syntax sentence structure, Semantics syntax and Pragmatics which refers to understanding.

2. Levels of NLP

The ‘levels of language’ are one of the most explanatory method for representing the Natural Language processing which helps to generate the NLP text by realizing Content Planning, Sentence Planning and Surface Realization phases. The various important levels of Natural Language Processing are:



1. Phonology

Phonology is the part of Linguistics which refers to the systematic arrangement of sound. The term phonology comes from Ancient Greek and the term phono- which means voice or sound, and the suffix –logy refers to word or speech. Phonology include semantic use of sound to encode meaning of any Human language.

2. Morphology

The different parts of the word represent the smallest units of meaning known as Morphemes. Morphology which comprise of Nature of words, are initiated by morphemes. An example of Morpheme could be, the word precancellation can be morphologically scrutinized into three separate morphemes: the prefix pre, the root cancella, and the suffix -tion. The interpretation of morpheme stays same across all the words, just to understand the meaning humans can break any unknown word into morphemes. For example, adding the suffix –ed to a verb, conveys that the action of the verb took place in the past. The words that cannot be divided and have meaning by themselves are called Lexical morpheme (e.g.: table, chair) The words (e.g. -ed, -ing, -est, -ly, -ful) that are combined with the lexical morpheme are known as Grammatical morphemes (eg. Worked, Consulting, Smallest, Likely, Use). Those grammatical morphemes

that occurs in combination called bound morphemes (eg. -ed, -ing). Grammatical morphemes can be divided into bound morphemes and derivational morphemes.

3. Lexical

In Lexical, humans, as well as NLP systems, interpret the meaning of individual words. Various types of processing present to word-level understanding – the first of these being a part-of-speech tag to each word. In this processing, words that can act as more than one part of-speech are assigned the most probable part-of speech tag based on the context in which they occur. At the lexical level, Semantic representations can be replaced by the words that have one meaning. In NLP system, the nature of the representation varies according to the semantic theory deployed.

4. Syntactic

This level emphasis to scrutinize the words in a sentence so as to uncover the grammatical structure of the sentence. Both grammar and parser are required in this level. The output of this level of processing is representation of the sentence that reveal the structural dependency relationships between the words. Syntax conveys meaning in most languages because order and dependency contribute to meaning. For example, the two sentences: 'The cat chased the mouse.' and 'The mouse chased the cat.' differ only in terms of syntax yet convey quite different meanings.

5. Semantic

In semantic most people think that meaning is determined, however, this is not it is all the levels that bestow to meaning. Semantic processing determines the possible meanings of a sentence by pivoting on the interactions among word-level meanings in the sentence. This level of processing can incorporate the semantic disambiguation of words with multiple senses; in a similar way to how syntactic disambiguation of words that can run as multiple parts-of-speech is practiced at the syntactic level. For example, amongst other meanings, 'file' as a noun can mean either a binder for gathering papers, or a tool to form one's fingernails, or a line of individuals in a queue. The semantic level scrutinizes words for their dictionary clarification, but also for the explanation they derive from the meaning of the sentence.

6. Discourse

While syntax and semantics goes side by side with sentence, the discourse level of NLP travail with units of text longer than a sentence i.e, it does not interpret multi sentence texts as just sequence sentences, apiece of which can be explained individually. Rather, discourse focuses on the properties of the text as a whole that convey meaning by making connections between component sentences. The two of the most common levels are Anaphora Resolution - Anaphora resolution is the replacing of words such as pronouns, which are semantically stranded, with the relevant entity to which they refer. Discourse/Text Structure Recognition - Discourse/text structure recognition sways the functions of sentences in the text, which, in turn, adds to the meaningful representation of the text.

7. Pragmatic:

Pragmatic is concerned with the firm use of language in situations and utilizes core over and above the crux of the text for understanding the goal and to explain how extra meaning is read

into texts without literally being encoded in them. This requisite much world knowledge, including the understanding of intentions, plans, and goals. For example, the following two sentences need aspiration of the anaphoric term ‘they’, but this aspiration requires pragmatic or world knowledge.

3. History

The history of natural language processing (NLP) generally started in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence.

In early 1980s computational grammar theory became a very active area of research linked with logics for meaning and knowledge’s ability to deal with the user’s beliefs and intentions and with functions like emphasis and themes. By the end of the decade the powerful general-purpose sentence processors like SRI’s Core Language Engine and Discourse Representation Theory offered a means of tackling more extended discourse within the grammatic-logical framework. Practical resources, grammars, and tools and parsers became. The DARPA speech recognition and message understanding conferences were not only for the tasks they addressed but for the emphasis on heavy evaluation, starting a trend that became a major feature in 1990s. Some researches in NLP marked important topics for future like word sense disambiguation and probabilistic networks, statistically colored NLP, the work on the lexicon, also pointed in this direction. Statistical language processing was a major thing in 90s, because this does not only involve data analysts. Information extraction and automatic summarizing was also a point of focus.

Recent researches are mainly focused on unsupervised and semi-supervised learning algorithms. Such algorithms can learn from data that has not been hand-annotated with the desired answers or using a combination of annotated and non-annotated data. Generally, this task is much more difficult than supervised learning, and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available (including, among other things, the entire content of the World Wide Web), which can often make up for the inferior results if the algorithm used has a low enough time complexity to be practical.

In the 2010s, representation learning and deep neural network-style machine learning methods became widespread in natural language processing, due in part to a flurry of results showing that such techniques can achieve state-of-the-art results in many natural language tasks, for example in language modeling, parsing, and many others.

4. Applications of NLP

NLP has a wide spectrum of applicability. Only a tip of iceberg features has been explored and rest is still in progress. So far areas like Machine Translation, Email spam detection, Information Extraction, Summarization and question answering are some of the explored and worked areas.

- Machine Translation is very crucial as the entire world is present online and the task of data accessible to each individual is a huge challenge. Language barrier contributes most to the challenge, with every language associated is a multitude of structure and grammar.

- Spam filtering works using text categorization and in recent times various machine learning techniques have been applied to text categorization or anti-spam filtering just like Rule learning, Naïve Bayes models.
- Information extraction concern with identifying more relevant and correct textual data. There are many applications for whom, extracting entities such as names, places, dates and time is a powerful way of summarizing the relevant information as per the user's need is concerned.
- Summarization, as we are currently surrounded by data, which means our ability to understand it. Since data is on an ever-increasing trend and the ability to summarize it with exact meaning is high in demand. This gives us a better chance to manipulate data and also to take necessary decisions (which is what NLP is trying to do).
- Language translation applications such as Google Translate
- Word Processors such as Microsoft Word and Grammarly that employ NLP to check grammatical accuracy of texts.
- Interactive Voice Response (IVR) applications used in call centers to respond to certain users' requests.
- Personal assistant applications such as OK Google, Siri, Cortana, and Alexa.

5. Approaches

1. Hidden Markov Model (HMM)

An HMM is a system where a shifting takes place between several states, generating feasible output symbols with each switch. The sets of viable states and unique symbols may be large, but finite and known. Few of the problem could be solved are by Inference A certain sequence of output symbols, compute the probabilities of one or more candidate states with sequences. Pattern matching the state-switch sequence is realized are most likely to have generated a particular output- symbol sequence. Training the output-symbol chain data, reckon the state-switch/output probabilities that fit this data best. Hidden Markov Models are extensively used for speech recognition, where the output sequence is matched to the sequence of individual phonemes.

2. Naive Bayes Classifiers

Naive Bayes Classifier Algorithm is a family of probabilistic algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of a feature. The choice of area is wide ranging covering usual items like word segmentation and translation but also unusual areas like segmentation for infant learning and identifying documents for opinions and facts. Naive Bayes predict the tag of a text and calculates the probability of each tag for a given text and then output the tag with the highest one.

6. Future of NLP

Natural Language Processing plays a critical role in supporting machine-human interactions. As more research is being carried in this field, we expect to see more breakthroughs and models that will make machines smarter at recognizing and understanding the human language.