Chapter 4 - Introduction to NLP

Natural language processing

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written referred to as natural language. It is a component of artificial intelligence (AI).

NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors such as ears to hear and eyes to see computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.

NLP Tasks

The following are the tasks:

- ***Speech recognition***, also called speech-to-text, is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions. What makes speech recognition especially challenging is the way people talk—quickly, slurring words together, with varying emphasis and intonation, in different accents, and often using incorrect grammar.
- ***Part of speech tagging***, also called grammatical tagging, is the process of determining the part of speech of a particular word or piece of text based on its use and context. Part of speech identifies ‘make’ as a verb in ‘I can make a paper plane,’ and as a noun in ‘What make of car do you own?’
- ***Word sense disambiguation*** is the selection of the meaning of a word with multiple meanings through a process of semantic analysis that determine the word that makes the most sense in the given context. For example, word sense disambiguation helps distinguish the meaning of the verb 'make' in ‘make the grade’ (achieve) vs. ‘make a bet’ (place).
- ***Named entity recognition***, identifies words or phrases as useful entities. Named entity recognition identifies ‘Kentucky’ as a location or ‘Fred’ as a man's name.
- ***Co-reference resolution*** is the task of identifying if and when two words refer to the same entity. The most common example is determining the person or object to which a certain pronoun refers (e.g., ‘she’ = ‘Mary’), but it can also involve identifying a metaphor or an idiom in the text (e.g., an instance in which 'bear' isn't an animal but a large hairy person).
- ***Sentiment analysis*** attempts to extract subjective qualities - attitudes, emotions, sarcasm, confusion, suspicion from text.

- ***Natural language generation*** is sometimes described as the opposite of speech recognition or speech-to-text; it's the task of putting structured information into human language.

How does natural language processing work?

There are two main phases to natural language processing: data preprocessing and algorithm development.

Data preprocessing involves preparing and "cleaning" text data for machines to be able to analyze it. Preprocessing puts data in workable form and highlights features in the text that an algorithm can work with. There are several ways this can be done, including:

- ***Tokenization*****:** This is when text is broken down into smaller units to work with.
- ***Stop word removal*****:** This is when common words are removed from text so unique words that offer the most information about the text remain.
- ***Lemmatization and stemming*****:** This is when words are reduced to their root forms to process.
- ***Part of speech tagging*****:** This is when words are marked based on the part-of speech they are such as nouns, verbs and adjectives.

Once the data has been preprocessed, an algorithm is developed to process it. There are many different natural language processing algorithms, but two main types are commonly used:

- ***Rules-based system*****:** This system uses carefully designed linguistic rules. This approach was used early on in the development of natural language processing, and is still used.
- ***Machine learning-based system*****:** Machine learning algorithms use statistical methods. They learn to perform tasks based on training data they are fed, and adjust their methods as more data is processed. Using a combination of machine learning, deep learning and neural networks, natural language processing algorithms hone their own rules through repeated processing and learning.

Techniques and methods of natural language processing

Syntax and semantic analysis are two main techniques used with natural language processing.

Syntax is the arrangement of words in a sentence to make grammatical sense. NLP uses syntax to assess meaning from a language based on grammatical rules. Syntax techniques include:

- ***Parsing*****:** This is the grammatical analysis of a sentence. Example: A natural language processing algorithm is fed the sentence, "The dog barked." Parsing involves breaking this sentence into parts of speech -- i.e., dog = noun, barked = verb. This is useful for more complex downstream processing tasks.
- ***Word segmentation*****:** This is the act of taking a string of text and deriving word forms from it. Example: A person scans a handwritten document into a computer. The algorithm would be able to analyze the page and recognize that the words are divided by white spaces.

- ***Sentence breaking*:** This places sentence boundaries in large texts. Example: A natural language processing algorithm is fed the text, "The dog barked. I woke up." The algorithm can recognize the period that splits up the sentences using sentence breaking.
- ***Morphological segmentation*:** This divides words into smaller parts called morphemes. Example: The word untestablely would be broken into [[un[[test]able]]ly], where the algorithm recognizes "un," "test," "able" and "ly" as morphemes. This is especially useful in machine translation and speech recognition.
- ***Stemming*:** This divides words with inflection in them to root forms. Example: In the sentence, "The dog barked," the algorithm would be able to recognize the root of the word "barked" is "bark." This would be useful if a user was analyzing a text for all instances of the word bark, as well as all of its conjugations. The algorithm can see that they are essentially the same word even though the letters are different.

Semantics involves the use of and meaning behind words. Natural language processing applies algorithms to understand the meaning and structure of sentences. Semantics techniques include:

- ***Word sense disambiguation*:** This derives the meaning of a word based on context. Example: Consider the sentence, "The pig is in the pen." The word pen has different meanings. An algorithm using this method can understand that the use of the word pen here refers to a fenced-in area, not a writing implement.
- ***Named entity recognition*:** This determines words that can be categorized into groups. Example: An algorithm using this method could analyze a news article and identify all mentions of a certain company or product. Using the semantics of the text, it would be able to differentiate between entities that are visually the same. For instance, in the sentence, "Daniel McDonald's son went to McDonald's and ordered a Happy Meal," the algorithm could recognize the two instances of "McDonald's" as two separate entities -- one a restaurant and one a person.
- ***Natural language generation*:** This uses a database to determine semantics behind words and generate new text. Example: An algorithm could automatically write a summary of findings from a business intelligence platform, mapping certain words and phrases to features of the data in the BI platform. Another example would be automatically generating news articles or tweets based on a certain body of text used for training.

What is natural language processing used for?

Some of the main functions that natural language processing algorithms perform are:

- ***Text classification*:** This involves assigning tags to texts to put them in categories. This can be useful for sentiment analysis, which helps the natural language processing algorithm determine the sentiment, or emotion behind a text. For example, when brand A is mentioned in X number of texts, the algorithm can determine how many of those mentions were positive and how many were negative. It can also be useful for intent detection, which helps predict what the speaker or writer may do based on the text they are producing.

- ***Text extraction*:** This involves automatically summarizing text and finding important pieces of data. One example of this is keyword extraction, which pulls the most important words from the text, which can be useful for search engine optimization. Doing this with natural language processing requires some programming it is not completely automated. However, there are plenty of simple keyword extraction tools that automate most of the process -- the user just has to set parameters within the program. For example, a tool might pull out the most frequently used words in the text. Another example is named entity recognition, which extracts the names of people, places and other entities from text.
- ***Machine translation*:** This is the process by which a computer translates text from one language, such as English, to another language, such as French, without human intervention.
- ***Natural language generation*:** This involves using natural language processing algorithms to analyze unstructured data and automatically produce content based on that data. One example of this is in language models such as GPT3, which are able to analyze an unstructured text and then generate believable articles based on the text.