Charles Wallis
1/26/2023

Natural Language Processing (NLP) is the branch of study that makes computers able to understand human language. It is a subcategory of artificial intelligence, a discipline created by the subdivision of AI that analyzes, interprets, and processes human language after attempts to create GAI in the 1960s. This is a field of research in language engineering, computer science, AI, and computational linguistics, and is also related to interpreting natural languages with computers, analyzing and understanding their meanings, and automatically generating them. Information extraction, automatic calibration, dialogue systems, and machine translations are all part of NLP. One of the famous methods of testing NLP systems is the Turing test, which tests the ability of the machine or a program's ability to be completely indistinguishable from a human.

Although research on natural languages has been going on for a long time, computers still do not understand natural languages like humans do. Instead, the technology for processing large amounts of information using superficial probabilities and statistics without a deep understanding of the language has advanced a lot. Recent research and advancement of machine learning and AI has improved the performance of NLP significantly. Best example is Google for their search engine, and the recent emergence of chatGPT.

Natural Language Understanding (NLU) is the computer's ability to understand and interpret human language, through analyzing text or speech to extract the meaning, emotion, and purpose. Natural Language Generation (NLG) is the computer's ability to generate human language text (or speech), through using AI algorithms and models to generate a natural and humanlike text or speech. This is used in chatbots (dialogue systems) like siri or chatGPT.

Information Retrieval and Information Extraction
IR is the study of creating search engines. It is the fastest growing and richest sector in the last 10 years. In a broad sense, it refers to the process of quickly finding documents that users want among numerous documents. The search term (query) by the user and the index is used for quick search, by dividing all documents on the web into words, then groups them into unigram, bigram, trigram, etc., and stores a link to which documents the gram is included. In addition, there are content-based IR or multimedia IR that uses images, voices, and music as queries to search for similar results. IE is largely divided into techniques for extracting normalized information from unstructured documents, object named entity recognition (NER) and relation extraction.

Speech Recognition / Speech-to-text
This branch of study makes computers understand human speech language. Siri, Alexa, Google home, and bixby are all great examples of speech recognition technologies. SR and STT technologies are slightly differentiated in this field, as SR aims for computers to understand human speech languages while STT only aims to convert human speech languages into text. Automatic captions are an example of STT.

Sentiment Analysis / Opinion Mining
This is used to classify what intentions or feelings a sentence represents, and are used on user reviews and comments that may be left on product reviews. Recent advance in artificial neural network techniques has allowed for a significant performance improvement in this field.

Dialogue Systems
Recently the dialogue system chatGPT became popular, this is a study in which a computer generates the most suitable sentence to the user based on previous input (sometimes speech), to simulate a conversation. HMM based machine learning technologies are mainly used for this.

Machine Translation

This research area aims to translate one language to another. Inter-lingual translation with high systematic similarity shows great performance, but languages with low systematic similarity lack performance yet. Recent advancement in artificial neural networks and recurrent neural networks are allowing significant improvements, but because this requires vast amounts of linguistic data, it takes a very long time to compute.

Three main approaches to NLP are rules-based, statistical and probabilistic, and deep learning. Rules-based approaches are the oldest technique. They are used to generate syntactically correct sentences, or to check if sentences are grammatically correct. One famous approach was to use a regular expression to echo talking points back to the user, until it couldn't form an answer. Because human language is so complex, it can't be fully encapsulated in rules.

Next approach is the statistical and probabilistic approach, which understands text through large amounts of data to train models and then identify patterns and make predictions. Machine translation and sentiment analysis is an example of this, because they learn by statistical and probabilistic methods.

Deep learning is the final approach, using neural networks and artificial intelligence to analyze and understand text. This is a lot more flexible and robust since it has a lot of 'knowledge' but it requires a lot of data to train the model and it takes a lot of time. chatGPT used this approach and it was built on top of GPT-3. As the processing power increases through advancing GPUs and cloud computing, with more data and time, technologies like RNN and CNN, LSTMs can show incredible results on NLP.

I grew up bilingual and I worked as a translator for a few years, and with the help of machine translation I was able to quickly learn more about both languages and their systems. My personal goal lies in creating my own translator where I can select the category for and the translator will know in what context it exactly lies in to output a proper, more natural, and colloquially correct grammar and style for each context/category.