Overview of NLP

Natural Language Processing, or NLP for short, is a rapidly developing field that involves the examination and generation of natural human languages. These human languages differ from other types of languages, like programming languages, which generally have stricter syntax and can be more easily examined with rules-based approaches. Natural languages are incredibly complex and expansive, and while rules-based methods have their merits, the more common methods for language processing involve statistical models and machine learning. This, of course, ties advancements in the field directly to advances in machine learning and artificial intelligence; one of the most popular and effective approaches to natural language processing is the use of deep learning, which involves large-scale neural networks, most of which require an immense amount of data and processing power.

As stated before, NLP covers both the understanding and generation of human language. Understanding language requires taking human input and examining it to find meaning or provide some service. Search engines are a common example of language understanding without the need for generation, as the search engine's response to a request is usually a collection of preexisting content. Language generation is the creation and output of brand new dialogue. Chatbots, as well as the more theoretical general artificial intelligence, must be able to both understand language and generate it in response, thus resulting in conversation. With technology giants having easy access to brilliant programmers and heaps of language data from their users, NLP has advanced tremendously. New models like GPT-3 can be trained on user text to emulate a certain style, and the ChatGPT service can effectively answer questions and solve trivial problems.

The earliest instances of NLP used rules to examine language. These approaches are not as sophisticated as machine learning, but they still find use as a subset of today's methodologies. Context-free grammars are one such rules-based approach; these grammars are a collection of production rules, where specific nonterminal symbols produce other sets of symbols. In this way, one can start off with a collection of nonterminal symbols and end up with language. The characters of any given language, including punctuation, make up the terminal symbols, which have no production rules of their own.

As time progressed, statistical and probabilistic methods became the driving force behind NLP. Using statistics, it is possible to examine the likelihood of encountering or requiring specific words or symbols. Unlike a set of rules, these methods require training data to determine the probabilities behind a given language, which are not predetermined. Examples of these approaches include standard machine learning methods like logistic regression and decision trees. Small neural networks also fall under statistical methods, but neural networks expanded greatly and formed a category of their own.

Deep learning has taken the world by storm, and NLP is no exception. Complex neural networks are trained with thousands, if not millions of data instances in order to form the statistical basis for certain languages. Rather than start out with a set of rules, deep learning models can learn the rules through propagation. Most users do not have the time, money, or information to develop deep learning models for themselves, but a variety of pre-trained models exist for private development in artificial intelligence and NLP. GPT-3 is a pre-trained language model; most of the training was already done by people with the necessary resources, and new users can simply replace and train the final layer for more specific results. In modern NLP models, there usually is not a "one size fits all" approach. Rather than using just one of these methods, cutting-edge NLP technology generally uses a combination of rules, statistics, and deep learning to achieve the best results.

I personally do not have much interest in NLP technology; I took this course primarily because I enjoyed my Intro to Machine Learning course last semester and loved Dr. Mazidi's teaching style. I am, however, concerned by some of the ethical ramifications of new machine learning developments, including those in the NLP field. Human language generation can be used to easily cheat on school assignments, and in the professional space it could lead to the replacement of human beings in more trivial jobs. Written and artistic works are also being stolen with products like AI art and GitHub Copilot. In the future, I would like to look into ways to combat the downsides of machine learning without limiting the positive uses of such technology.