Overview of NLP

NLP, or natural language processing, is a field of study that attempts to process human language using AI and Machine Learning techniques in a way that can produce useful insights and data regarding language-based datasets.

NLP, as stated earlier, uses techniques within the realm of AI to process natural language, since it is a complex and dynamic task that warrants the use of AI for processing.

Natural language understanding and natural language generation are both subsets of natural language processing, and both require a thorough understanding of human language in the context of their specific task. However, whereas natural language understanding only needs to understand and make predictions/classifications regarding language data, natural language generation is the task creating human-like text/speech based on the data it was trained on.

NLP applications vary from usecase to usecase. One prominent subset of NLP is that of sentiment analysis, something that businesses value very highly because of the vast amount of customer review data that exists outside of specialized review pages. The power that is held by a company that's able to parse Twitter and figure out how many people love/hate their product automatically using an sentiment analyzer is valued very highly, and recently researchers at the Microsoft Research Labs at Washington were able to predict which women were at risk of postnatal depression just by similarly analyzing their Twitter posts. A more consumer-focused usecase is that of speech recognition, which can be seen in Siri, Cortana, Google voice assistant, and Alexa.

There have been 3 major approaches to NLP throughout the years (ever since the 1960s). The first and weakest of the approaches is the Rules-base approach. A rules-based approach attempted to boil down human language to a set of rules (like grammar) in order to determine if a section of text was valid or not. This was very limited, which is why its use cases are also simplistic: Spell check, context-free grammar, and a basic chatbot that was only able to

output valid responses to a user's input but not any helpful responses that actually took context into account.

The second approach was a statistical/probabilistic approach. Using probabilities based on certain word and Part of speech frequencies, the models (given a moderate amount of data and processing power) could more reliably respond like a human would. This is because this "probabilistic model" is based upon traditional Machine Learning algorithms (such as the Baum-Welch algorithm, which can use observed data to find maximum likelihood parameters for the HMM that will then model the problem. This was originally used in 1975 by James K Baker in the field of speech recognition, and later also used in speech synthesis). More examples with HMMs include Part of speech tagging, a task that is very important even today, since determining nouns vs verbs can easily make use of probabilities built upon vast amounts of data to create an HMM that can classify a word based on the previous word.

The final and most recent approach is that of deep learning. Deep learning utilizes deep neural networks in order to better model the complexity of human language, since the vast number of layers and neurons allows the model to better understand the intricacies of the underlying structure of language. This allows it to perform better than other models in usecases like language generation, translation, and even unlock usable results in specialized subtasks of language understanding like sarcasm detection. The only issue is that most deep learning models require a large amount of data for each task, exceptions include the BERT model which does need a massive amount of data for pretraining (a form of training that allows the model to understand language in general, usually a large Wikipedia or in more recent times even Twitter based corpus of text, which are both very easy to obtain large amounts of) however afterwards can use a relatively small amount of data to “fine tune” the model on its specific task.

I have a great interest in the idea of removing menial work from the workplace, and developments in NLP in order to analyze vast amounts of human data automatically is a huge part of that. I recently did a research project under Dr. Vincent Ng here at UTD studying sarcasm detection specifically for tweets, which struck me as a particularly interesting topic simply because the sooner we’re able to understand message on a human level using ML, the

sooner things like automatic moderation/flagging can be done, and Twitter is just one of many platforms in dire need of some moderation. I don't personally support the idea of companies using all forms of text data in order to "personalize" ad content even more, but rather I'm interested in the use of NLP to finally moderate and clean the internet, something that has always been hand-waved despite being harmful simply because no amount of people could ever keep track of everything from doxing to actionable threats and hate speech. Now it's within sight to be able to detect all those things and make the internet (or at least private platforms who wish to enforce their own kind of moderation, things like PBS kids or YouTube kids comment sections being allowed because automatic monitoring would keep it sanitary) something that can be safe for everyone.