Harry Wada <hcw170000@utdallas.edu>   Jan 28, 2023
CS 4395.001

Overview of NLP

a. Define NLP in your own words

NLP, referring to natural language processing, is the study and pursuit of allowing computers to process human language. This has the obvious practical benefit of bridging the communicative rift between humans and computers; no longer would it be necessary for the user to understand code or rely on more rigid interfaces. Instead, the computer would be capable of grokking the basic meaning of language without the aid of human intervention, presumably interpreting and translating this into some other form that proves expedient for the machine's designed purpose.

b. Describe the relationship between AI and NLP

As language is a uniquely human phenomenon, NLP naturally has a close association with other disciplines that relate closely to Man and the mind, perhaps most saliently artificial intelligence. In fact, as a specialization of AI, NLP benefits from many of the same fruits that are bore by efforts in more general AI research. In this way they operate and build off each other, making up, among the other AI sub-fields, a pool of closely related ideas and methodologies.

c. Write a sentence or two comparing and contrasting natural language understanding and natural language generation

Natural language generation and processing are in some sense two sides of the same coin. To wit, in the case of generation one has meaning but not the serialized representation, while with understanding one has the representation but not the meaning.

d. List some examples of modern NLP applications

Various applications include improving search engine results, more dynamic customer support, and automating quotidian tasks to no longer require direct human-computer engagement.

e. Write three paragraphs describing each of the three main approaches to NLP, and list examples of each approach

The first attempt at NLP came in the form of pre-constructed rules, often regular expressions, by which to parse or generate text. This involves recognizing the general forms that a language tends to take and attempting to conform (in the case of parsing) or project (in the case of generation) that into novel input. Attempts to apply these principle include pluralizing words and verifying grammatically correct sentences.

Another approach came in the form of utilizing probabilistic and statistical models. Given a sufficiently large training set, it's possible for a system to analyze and generate models based on the frequency and sequencing of words, making it more dynamic than the previous rule-based techniques. Some of these models were the likes of naïve Bayes, logistic regression, SVM, small neural networks, and others, and found practical use in problems such as translation and predicting future text from past evidence.

The approach that currently has its stranglehold on modern times is deep learning, an extension of neural networks incorporating extremely large amounts of data, capitalizing on advances in GPU performance. Models of this ilk include recurrent neural networks, convolutional neural networks, LSTMs, among others. Each of these are optimized with different goals in mind, but achieve their results using the same fundamental principles that derive from models of the brain.

f. Write a paragraph describing your personal interest in NLP and whether/how you would like to learn more about NLP for personal projects and/or professional application

Personally, I have little interest in NLP or even AI more generally; I take this class only to fulfill a requirement and graduate. My view is that there are fundamental differences between Man and machine, and that AI cannot even approximate the mind without crippling technical debt and intractability. Neural networks are already conspicuously nebulous, with most performance gains only seen through blind tinkering of hyperparameters. There's also the issue of ethics surrounding collection and potential generation of copyrighted material, the imminent threat of legislation toward which places the whole field in a very precarious position. Without large, unrestricted training data as have been used by many of the modern state-of-the-art technologies, I doubt any truly compelling results can be produced.