CS 4395, Assignment 0 - Overview of NLP

Spring 2023, Section 001

Aditya Rathod (NetID: AGR190000)

Due January 28, 2022 at 11:59pm

Introduction

Natural Language Processing is a field that focuses on teaching computers to understand and manipulate human languages. NLP is considered a subfield of artificial intelligence, and uses rule-based and ML-based techniques to achieve its goal. Natural language understanding is all about processing existing text and comprehending the text, and natural language generation is focused on creation of text. They go hand-in-hand, with NLU being able to understand the output of NLG.

Applications

Some modern NLP applications include ChatGPT, Siri, and Google Translate. ChatGPT is an application that utilizes a very large neural network to generate text based on prompts in a conversational fashion. Siri is a virtual assistant that analyzes user speech to produce results for. Google Translate is a machine translation tool that utilizes community ratings for translations to improve its performance.

Approaches

The three main ways to approach the problem of NLP include rules-based, statistical and probabilistic, and deep learning.

Rules-based

Rules-based approaches tend to be older, and dealt with using regular expressions and mapping out sentences. Tasks like removing plurals from words tended to require regular expressions with lists of exceptions. Sentences were mapped using context-free grammars that created production rules for sentences, which in turn could be used for NLG or NLU (either forming grammatically correct sentences or checking for the form of sentences). The problem with this approach tends to be the scaling of the rules to the complexity of the language; since nobody could possibly encode all the intricacies of English down to to a bunch of rules. An example of a rules-based system is ELIZA.

Statistical and probabilistic

Statistical and probabilistic methods came after rules-based approaches, and reduced language down to numerical values. Researchers determined that via simple word counts and finding the probabilities of words and sequences was significantly more effective in modeling language in a way that is conducive to usages such as machine translation and predictive text. Classical machine learning algorithms fall into this bucket, including regression, SVMs, and even multi-layer perceptrons. These methods typically learned from a dataset called a *corpus*. An example of a system that could use this technique are the typeahead autosuggestion features of mobile keyboards.

Deep learning

After the increase in computing power and data came advances in machine learning, increasing the size and structure of neural networks to the point where it was somewhat its own field. With neural network types such as RNNs, CNNs, LSTMs, and transformers, deep learning offers many different methods for extending the previously described probabilistic approaches. Some of the most advanced language models today use hundreds of billions of parameters (thanks to the famous Google paper, "Attention is All You Need"). An example of a system that uses this technique would be ChatGPT.

Personal interest statement

Machine learning has interested me ever since I completed the famous Andrew Ng Coursera course as a junior in high school. But after that, I still felt inadequately informed about deep learning, and thus began my casual study of deep learning on the side. But I lacked an appealing application area to motivate me to continue learning, and while I retained the knowledge I gained, I stepped away from ML and deep learning until right before my college career.

Before my freshman year, I worked as a Clark Scholar at UT Dallas, where I learned the basics of deep learning and was encouraged to pick a topic to explore as a final project. That's when I realized that models focusing on natural language of were partiular interest to me, and I set off to learn some rudimentary NLP techniques in my attempt to create my own model for news article summarization. As part of that project, I learned about things like tokenization, word vectors, and sequence modeling. While that project was ultimately not as successful as I had hoped, I walked away with a keen interest in the intersection of ML and language.

With my newfound experience, I began informally teaching what I knew to underclassmen as a founding project lead for the ACM Research program. This was a great way to keep my knowledge fresh, and kept me reading about the latest advances in deep learning. In fact, in my third semester leading a group for the program, the team I led created a mobile keyboard model trained on Twitter data that was able to pick up on emerging language trends and suggest them to users. While working on that project, I realized that I lacked a clear first-principles understanding of NLP, and was determined to learn the complete picture before I graduated. Unfortunately, due to the structuring of my degree, I was only able to take this class much later, my last semester here.

I hope to gain a deeper understanding and appreciation of modern NLP systems, and wish to apply these techniques in projects of my own, such as chatbots.