Matt McCormack

Insight 2

Professor Frazier

9/17/2020

One complex yet incredibly helpful application of data science is natural language processing, also known as NLP, which is the analysis of natural language data using machine learning and artificial intelligence. NLP is generally used with mainstream languages, like English – so, I thought it would be interesting to look into how NLP may be used with indigenous languages that are unfamiliar for many corpuses and methods currently used with common languages. A research project entitled *Challenges of Language Technologies for the Indigenous Languages of the Americas* goes into detail about how NLP is being used with indigenous languages in the Americas.

The project explains how native languages are incredibly diverse, which I think would logically make it more difficult to maintain overarching themes present in a corpus; each language has its own unique set of characteristics that set it apart from other languages. For many NLP methods to really be used in a meaningful way, there must be a large collection of pre-existing data (in the form of a corpus) that the model can learn from. So when it comes to these niche languages, it can be hard to amass enough natural language data to perform analysis. Machine translation is one advanced method of potentially getting from an indigenous language to a more well-known one; you can implement rule-based, neural, or statistical methods. Out of these approaches, it seems that the best one for indigenous languages would be rule-based machine translation. Rule-based machine translation is effectively an algorithm that uses grammar and dictionary rules to translate from the given language to a desired one. This method can function with relatively little data when compared to the statistical and neural methods. Both latter approaches require large quantities of training data to function well; these approaches rely on detecting patterns and translating based on those patterns, rather than dictionary relationships. As we discussed, getting such corpora can be somewhat difficult for indigenous languages.

In the context of human development in data science, NLP has a lot of potential - especially when it comes to these indigenous languages. If used properly, NLP (and particularly machine translation) can unlock barriers between people who use these languages and people who do not, whether that be speaking, writing, or reading. This research in NLP could allow the world to be more connected even in niche parts where the language is spoken by only a handful of people. While I already knew a decent amount about NLP processes, I found it very interesting to delve deeper into NLP outside of the English language. The challenges of having limited natural language data to feed into statistical/neural machine translation approaches proved to be more complicated than I thought because these methods rely on having parallel corpora system. I did end up learning a lot about rule-based, statistical, and neural machine translation systems, all of which have different caveats and applications. I did not know too much about these methods before reading this report, and learning about them in the context of indigenous language gave me a better idea of their capabilities and limitations.