How Google's SyntaxNet attempts to solve NLP's Ambiguity Problem

By: Zoheb Satta

Natural Language Processing, or NLP, can be defined as the manipulation of a natural language, such as speech or text, by software. Computer Scientists have been working on the problem of NLP for over 50 years and we still have not solved it completely.

NLP is difficult for computers for many reasons, mainly that natural languages were designed for humans to communicate efficiently, thus leading to "shortcuts" in the language such as:

- often, we omit common sense knowledge which we assume the reader/listener already possesses
- there is a lot of ambiguity in our speech that we also assume the reader/listener can discern

Both issues stem from the same problem: the fact that our language requires the reader/listener to already have a working base of knowledge and experience with the language to intuit the subtleties. This is almost impossible for a simple program to do, so we have relegated the job to artificial intelligence using machine learning. As one of the leading companies in the artificial intelligence field, Google has developed SyntaxNet, a neural network that acts as a foundation for a Natural Language Understanding (NLU) system.

One of the problems that NLP faces is with syntactic ambiguity, to demonstrate: "A man saw a boy with a telescope" could be interpreted in two ways: the first, is that the man was the one using the telescope and saw the boy through it, the other is that the boy had a telescope and the man saw him. Humans are easily able to infer the correct meaning of this sentence using the surrounding context of the text, but computers have more difficulty with it. This ambiguity problem is what one of the problems that SyntaxNet focuses on.

SyntaxNet processes an input sentence from left to right, with dependencies between words being incrementally added as each word in the sentence is considered. Each point in the processing has many possible outcomes, due to the ambiguity, and the neural network scores each one based on their plausibility. SyntaxNet utilizes a “beam search” in its model so that instead of immediately taking the first-best decision at each

point it allows for multiple partial hypotheses at each step, which are only discarded when multiple new hypotheses are ranked higher.[1]

At its foundation, SyntaxNet is a neural network. A neural net is a collection of algorithms that learns by analyzing huge quantities of data, SyntaxNet works by analyzing millions of sentences which have been labelled by human users, and once it has learned those sentences, it can apply itself to identify similar properties on other sentences. To make SyntaxNet work even better, it works in conjunction with another Google technology: "Parsey McParseface", which is their own pre-built NLP service. According to Google, SyntaxNet with Parsey McParseface is almost 94 percent accurate in identifying the relation of a word to the sentence it is in, which they believe is close to the performance of a human at approximately 96-97 percent.[2]

Currently however, the SyntaxNet-Parsey McParsface relationship isn't the perfect NLU system as the datasets are limited to a collection of old newspaper stories, and the language is different than what is being spoken today on the internet, so when SyntaxNet is applied with more modern internet articles the accuracy drops closer to 90 percent.

SyntaxNet already has multiple applications with its current accuracy levels but Google is working on making SyntaxNet even better in multiple areas, including: increasing its accuracy with a more modern dataset as well as the fact that currently, SyntaxNet only works with the English language, although Google hopes to expand it to encompass all languages eventually.

---

[1] Petrov, Slav. “Announcing SyntaxNet: The World’s Most Accurate Parser Goes Open Source” ai.googleblog.com https://ai.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

[2] Metz, Cade “Google Has Open Sourced SyntaxNet, Its AI for Understanding Language” www.wired.com https://www.wired.com/2016/05/google-open-sourced-syntaxnet-ai-natural-language/