New framework brings accuracy, efficiency to identifying stop words

December 2 2019, by Alex Gerage

Credit: CC0 Public Domain

A research team led by Northwestern Engineering's Luis Amaral has developed an algorithmic approach for data analysis that automatically recognizes uninformative words—known as stop words—in a large

collection of text. The findings could dramatically save time during natural language processing as well as reduce its energy footprint.

"One of the challenges in machine learning and artificial intelligence approaches is that you don't know which data is useful to an algorithm and which data is unhelpful," said Amaral, Erastus Otis Haven Professor of Chemical and Biological Engineering at the McCormick School of Engineering. "Using information theory, we created a framework that reveals which words are uninformative for the task at hand."

The trouble with stop words

One of the most common techniques data scientists use in natural language processing is the bag-of-words model, which analyzes the words in a given text without considering the order in which they appear. To streamline the process, researchers filter out stop words, those adding no context to the data analysis. Many stop word lists are manually curated by researchers, making them time consuming to develop and maintain as well as difficult to generalize across languages and disciplines.

"Imagine you analyze millions of blog posts and want to learn what topic each post addresses," said Amaral, who codirects the Northwestern Institute on Complex Systems. "You would typically filter out common words like 'the' and 'you,' which don't provide any background about the topic."

However, the majority of words that are not useful for that specific task depend on the language and the blog's particular subject area. "For a collection of blogs on electronics, for example, there are many words that could not enable an algorithm to determine whether a blog post is about quantum computing or semiconductors," he added.

An information theoretic framework

The research team used information theory to develop a model that more accurately and efficiently identifies stop words. Central to the model is a 'conditional entropy' metric that quantifies a given word's certainty of being informative. The more informative the word, the lower its conditional entropy. By comparing the observed and the expected values of conditional entropy, the researchers could measure the information content of specific words.

To test the model, the researchers compared its performance to common topic modelling approaches, which infers the words most related to a given topic by comparing them to other text in the data set. This framework produced improved accuracy and reproducibility across the texts studied, while also being more applicable to other languages in a straightforward manner. Additionally, the system achieved optimal performance using significantly less data.

"Using our approach, we could filter 80 percent or more of the data and actually increase the performance of existing algorithms for topic classification of text corpora," Amaral said. "In addition, by filtering so much of the data, we are able to dramatically reduce the amount of computational resources needed."

Beyond saving time, the filtering system could lead to long-term energy savings, combating the negative impact large-scale computing has on climate change.

A paper describing the work was published December 2 in the journal *Nature Machine Intelligence*. Amaral was a co-corresponding author on the paper along with Martin Gerlach, a postdoctoral fellow in Amaral's lab.

While the researchers' analysis was restricted to bag-of-words approaches, Amaral is confident that his system could be expanded to account for additional structural features of language, including sentences and paragraphs.

In addition, since information theory provides a general framework for the analysis of any sequence of symbols, the researchers' system could be applicable beyond text analysis, supporting pre-processing methods for analyzing audio, images—even genes.

"We have begun applying this approach to the analysis of data from experiments measuring gene-specific RNA-molecules in individual cells as a way to automatically identify different cell types," Gerlach said. "Filtering uninformative genes—think of them as "stop genes"—is particularly promising to increase accuracy. Those measurements are much more difficult compared to texts and current heuristics are not nearly as well developed."

More information: Martin Gerlach et al, A universal information theoretic approach to the identification of stopwords, *Nature Machine Intelligence* (2019). DOI: 10.1038/s42256-019-0112-6

Provided by Northwestern University