Intelligent Machines

Data mining reveals fundamental pattern of human thinking

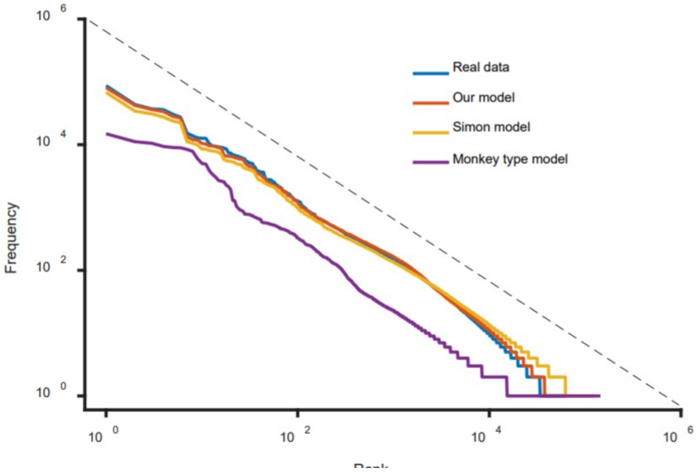
Word frequency patterns show that humans process common and uncommon words in different ways, with important consequences for natural-language processing.

by Emerging Technology from the arXiv July 16, 2018

Back in 1935, the American linguist George Zipf made a remarkable discovery. Zipf was curious about the relationship between common words and less common ones. So he counted how often words occur in ordinary language and then ordered them according to their frequency.

This revealed a remarkable regularity. Zipf found that the frequency of a word is inversely proportional to its place in the rankings. So a word that is second in the ranking appears half as often as the most common word. The third-ranked word appears one-third as often and so on.

In English, the most popular word is *the,* which makes up about 7 percent of all words, followed by *and*, which occurs 3.5 percent of the time, and so on. Indeed, about 135 words account for half of all word appearances. So a few words appear often, while most hardly ever appear.



But why? One intriguing possibility is that the brain processes common words differently and that studying Zipf's distribution should reveal important insights into this brain process.

There is a problem, though. Linguists do not all agree that the statistical distribution of word frequency is the result of cognitive processes. Instead, some say the distribution is the result of statistical errors associated with low-frequency words, which can produce similar distributions.

What's needed, of course, is a bigger study across a wide range of languages. Such a large-scale study would be more statistically powerful and so able to tease these possibilities apart.

Today, we get just such a study thanks to the work of Shuiyuan Yu and colleagues at the Communication University of China in Beijing. These guys have found Zipf's Law in 50 languages taken from a wide range of linguistic classes, including Indo-European, Uralic, Altaic, Caucasian, Sino-Tibetan, Dravidian, Afro-Asiatic, and so on.

Yu and co say the word frequencies in these languages share a common structure that differs from the one that statistical errors would produce. What's more, they say this structure suggests that the brain processes common words differently from uncommon ones, an idea that has important consequences for natural-language processing and the automatic generation of text.

Yu and co's method is straightforward. They begin with two large collections of text called the British National Corpus and the Leipzig Corpus. These include samples from 50 different languages, each sample containing at least 30,000 sentences and up to 43 million words.

The researchers found that the word frequencies in all the languages follow a modified Zipf's Law in which the distribution can be divided into three segments. "The statistical results show that Zipf's laws in 50 languages all share a three-segment structural pattern, with each segment demonstrating distinctive linguistic properties," they say Yu.

This structure is interesting. Yu and co have tried to simulate it using a number of models for creating words. One model is the monkey-at-a-typewriter model, which generates random letters that form words whenever a space occurs.

This process generates a power-law distribution like Zipf's Law. However, it cannot generate the three-segment structure that Yu and co have found. Neither can this structure be generated by errors associated with low-frequency words.

However, Yu and co are able to reproduce this structure using a model of the way the brain works called the dual-process theory. This is the idea that the brain works in two different ways.

The first is fast intuitive thinking that requires little or no reasoning. This type of thinking is thought to have evolved to allow humans to react quickly in threatening situations. It generally provides good solutions to difficult problems, such as pattern recognition, but can easily be tricked by non-intuitive situations.

However, humans are capable of much more rational thinking. This second type of thinking is slower, more calculating, and deliberate. It is this kind of thinking that allows us to solve complex problems like mathematical puzzles and so on.

The dual-process theory suggests that common words like *the, and, if* and so on are processed by fast, intuitive thinking and so are used more often. These words form a kind of backbone for sentences.

However, less common words and phrases like *hypothesis* and *Zipf's Law* require much more careful thought. And because of this they occur less often.

Indeed, when Yu and co simulate this dual process, it leads to the same three-segment structure in the word frequency distribution that they measured in 50 different languages.

The first segment reflects the distribution of common words, the last segment reflects the distribution of uncommon words, and the middle segment is the result of the crossover of these two regimes. "These results show that Zipf's Law in languages is motivated by cognitive mechanisms like dual-processing that govern human verbal behaviors," say Yu and co.

That's interesting work. The idea that the human brain processes information in two different ways has gained considerable momentum in recent years, not least because of the book *Thinking, Fast and Slow* by the Nobel Prize–winning psychologist Daniel Kahneman, who has studied this idea in detail.

A well-known problem used to trigger fast and slow thinking is this:

"A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?"

The answer, of course, is 5 cents. But almost everyone has the initial inclination to think 10 cents. That's because 10 cents feels about right.

It's the right order of magnitude and is suggested by the framing of the problem. That answer comes from the fast, intuitive side of your brain.

But it's wrong. The right answer requires the slower, more calculating part of your brain.

Yu and co say the same two processes are involved in generating sentences. **The** fast-thinking part **of your** brain creates **the** basic structure **of the** sentence (**the** words **here** marked **in** bold). **The** other words require **the** slower, **more** calculating part **of your** brain.

It is this dual process that leads to the three-segmented Zipf's Law.

That should have interesting consequences for computer scientists working on natural language processing. This field has benefited from huge advances in recent years. These have come from machine-learning algorithms but also from large databases of text gathered by companies like Google.

But generating natural language is still hard. You don't have to chat with Siri, Cortana, or the Google Assistant for long to come up against their conversational limits.

So a better understanding of how humans generate sentences could help significantly. Zipf surely would have been fascinated.

Ref: arxiv.org/abs/1807.01855: Zipf's Law in 50 Languages: Its Structural Pattern, Linguistic Interpretation, and Cognitive Motivation

More from Intelligent Machines

Artificial intelligence and robots are transforming how we work and live.

01 **How to tell if you're talking to a bot**

The five best ways to detect fake social-media accounts.

by Will Knight

02 **Evolutionary algorithm outperforms deep-learning machines at video games**

Neural networks have garnered all the headlines, but a much more powerful approach is waiting in the wings.

by Emerging Technology from the arXiv

03 **From the age of perplexity to the era of opportunities**

Finance for Growth

by Francisco Gonzalez

More from Intelligent Machines