In search of the key word: Bursts of certain words within a text are what make them keywords

July 17 2012

This image is a symbolic representation of how language can be depicted in a binary sequence and thus makes it possible to draw conclusions about the content of the text. Credit: Gianluca Costantini

Human beings have the ability to convert complex phenomena into a one-dimensional sequence of letters and put it down in writing. In this process, keywords serve to convey the content of the text. How letters and words correlate with the subject of a text is something Eduardo Altmann and his colleagues from the Max Planck Institute for the

Physics of Complex Systems have studied with the help of statistical methods. They discovered that what denotes keywords is not the fact that they appear very frequently in a given text. It is that they are found in greater numbers only at certain points in the text. They also discovered that relationships exist between sections of text which are distant from each other, in the sense that they preferentially use the same words and letters.

The Dresden-based scientists mathematically studied the semantic properties of texts by translating ten different English texts into various codes. One of the chosen texts was the English edition of Leo Tolstoy's "War and Peace".

One example of what the scientists did was translate letters in a text into a binary sequence. They replaced all vowels with 1 and all consonants with 0. By employing additional mathematical functions, the scientists examined different levels of the text – both individual vowels and letters, as well as whole words – which had been translated into various codes. In so doing, it was possible to identify repeating patterns within the text as a whole. Such correlation within a text is referred to as long-range correlation. This indicates whether two letters located at arbitrarily distant points in the text are connected with each other. For example, when we find a letter "W" at a certain point, there is a measurably higher probability that we will find the letter "W" again a few pages later.

"Understandably enough, if a certain point in the book talks about war, there is a high probability that the word war will also appear a few pages later. What is surprising is that we also find this higher probability at the level of individual letters," says Altmann.

Keywords are more frequent in certain passages of text

The scientists found this long-range correlation not only between letters, but also within higher linguistic levels, such as words. Within individual levels, the correlation remains when looking at different texts. "What we find much more interesting is to examine how the correlation changes between the levels," says Altmann. Long-range correlation enables the scientists to draw conclusions about the extent to which certain words are connected to a topic. "Even the connection between a word and the letters it is composed of can be analysed in this way," explains Altmann.

Furthermore, the scientists also studied what is known as "burstiness", which describes whether increased occurrence of a pattern of characters is present in a passage of text. It shows, for instance, whether a word comes up at increased frequency in a certain text section. The more frequently a certain word is used in a passage, the more likely it is that that word is representative of a certain subject.

The scientists demonstrated that certain words come up repeatedly throughout a text, are however not present in bursts in a given text passage. Although these words do exhibit long-range correlation, they are not closely related to the topic at hand. "Articles are the best examples of these. They come up very frequently in every text, but they are not crucial in conveying a given topic," says Altmann.

Statistical text analysis works irrespective of language

Whereas both letters and words exhibit long-range correlation, it is rare for letters to appear in bursts at certain points in a text. "It is, in fact, very rare for a letter to be as closely connected with a topic as the word it forms a part of. In a manner of speaking, letters can be used more flexibly," explains Altmann. An "a", for example, can be a part of a great many words that have no connection with one and the same topic.

The scientists employed statistical text analysis as an easy way of

identifying the defining words of a given text. "By so doing, it is absolutely irrelevant which language the text is written in. The only thing that matters is the story and not language-specific rules," says Altmann. Their findings could be used in future to improve Internet search engines, and they could also help to analyse texts and identify plagiarism.

More information: Eduardo G. Altmann, Giampaolo Cristadoro and Mirko Degli Esposti, On the origin of long-range correlations in texts, *PNAS*, July 2, 2012, doi: 10.1073/pnas.1117723109

Provided by Max Planck Society