Abstract of Contribution 257

Abstract

Topics: Linguistics, Cognitive Neuroscience

Keywords: significance testing, Zipf's law, frequency counts, natural language, spike patterns

Probability Estimation of Rare Events in Linguistics and Computational Neuroscience

Stefan Evert[1], **Gordon Pipa**[1,2]

[1]Institute of Cognitive Science, University of Osnabrück, Germany; [2]Frankfurt Institute for Advanced Studies, Germany;
stefan.evert@uos.de

In many subdisciplines of Cognitive Science, researchers are confronted with frequency counts involving large numbers of rare event types, whose highly skewed frequency distribution adheres to the Zipf-Mandelbrot law. Notable examples include word frequencies and other linguistic phenomena, as well as neural spike patterns.

Standard hypothesis tests used to assess the statistical significance of frequency data are unsuitable for this situation. The primary reason is that they have been designed for single event types, but are now applied to the group of all rare events observed in a sample. In our talk, we show how the very large number of possible rare event types that could have occurred, their Zipfian probability distribution, and the discrete nature of frequency counts combine to inflate the risk of a false rejection (type I error) drastically.

We propose a posterior, group-adjusted significance test, which is not based on the probability of a type I error for an individual event type, but rather on the expected number of such type I errors in the group of all event types that occur with a certain frequency in the sample at hand. An important ingredient of this test is a model for the prior distribution of event type probabilities, derived from the power-law rank-frequency relationship stipulated by the Zipf-Mandelbrot law.

The new significance test is illustrated with applications to word frequency data and neural spike patterns. We compare the results to those of standard hypothesis tests and to adjusted probability estimates such as Good-Turing and Lidstone's law.