Abstract of Contribution 256

Abstract
Topics: Linguistics, others
Keywords: statistical models, natural language, non-randomness, term clustering

Statistical Models of Non-Randomness in Natural Language

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

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

Accurate estimates for the occurrence probabilities of words and other linguistic phenomena play an important role in computational modelling of the human language faculty, as well as in natural-language processing applications. Researchers have long been aware of the fact that the random sampling assumption made by standard significance tests does not hold for language data. One particularly serious problem is term clustering, i.e. the tendency of topical words and expressions to be used repeatedly in a single text. Our goal is to model such term clustering effects using generalized linear models (GLM), by treating each text as an autocorrelated time series of word occurrences.

In our model, we map texts to binary sequences of indicator variables with respect to a given word or expression W. Each indicator variable shows whether W occurs in a specific position in the text. We then predict the occurrence probability of W at each position as a conditional probability based on its previous usage and other parameters (such as overall position in the text, position relative to the current sentence, text genre, domain, etc.). Our model is therefore strictly causal, meaning that only past events are considered to explain the occurrence patterns of W. To estimate its parameters we use a generalized linear model (GLM). We implement the conditioning on temporal structure, i.e. clustering, based on cubic spline basis functions. Goodness-of-fit is assessed by cross-validation using deviance and the time-rescaling theorem to evaluate predictability of temporal structure.

To identify important components of the model we employ L1 regularized generalized linear models in a second step. L1 regularized solutions are sparse. We use this sparsity to devise a reduced model using only components with large parameters that indicate greater importance.