Empirical Research on Association Measures
The UCS Toolkit

Stefan Evert

University of Osnabrück – Germany

Keywords: association measures, mathematical properties, evaluation, simulation, visualization

1. Introduction

Quantitative measures of statistical association are an essential tool for the identification of recurrent word combinations and the extraction of lexical collocations from text corpora (based on the intuition that recurrence is a good indicator of collocativity). While one might expect statistical association to be a well-defined mathematical concept, a bewildering multitude of association measures is available, each one resulting in a different set of "recurrent combinations" when applied to corpus frequency data. Theoretical discussions such as Pedersen (1996) have failed to identify an "ideal" AM. In fact, decades of controversial discussion in mathematical statistics show that there are many ways of measuring association and its significance, all of which are equally plausible (cf. the overview given by Yates (1984)). Therefore, empirical studies are necessary in order to gain further insights into the characteristic properties of different association measures and make an appropriate choice for a specific application.

2. The UCS toolkit

The UCS toolkit is a collection of libraries and tools based on the Perl programming language (http://www.perl.com/) and the statistical computing environment and language GNU R (http://www.r-project.org/). Using Perl as a general-purpose scripting language and R as a back-end for statistical calculations and graphical display, UCS provides a comprehensive set of reference implementations for well-known and less well-known association measures (see http://www.collocations.de/AM/ for a complete list). The toolkit supports three important types of empirical research, as detailed in the following sections.

2.1. Simulation

Different association measures can be applied to real or invented frequency data. A comparison of the computed association scores gives a first glimpse of the differences between the measures (e.g., some of them might favour low-frequency candidates or candidates whose marginal frequencies are highly skewed). Measures that are based on the same line of statistical reasoning can be compared directly, and in some cases also with their theoretical limiting distribution (cf. Dunning 1998), indicating which of them over- or underestimate the association of particular word combinations.

2.2. Evaluation

While simulation experiments may highlight some desirable (or undesirable) general properties of association measures, they cannot help in choosing between fundamentally different approaches such as measures based on statistical hypothesis tests vs. direct or conservative estimates for the strength of association. This choice is primarily determined by the intended application of the respective measure, such as the extraction of recurrent word combinations or a specific type of lexical collocations. Consequently, it is essential to arrive at a better understanding of the relation between the linguistic features of word combinations and the corresponding association scores. This can be accomplished in a direct way by annotating the extracted candidate word pairs according to linguistic criteria and counting how many specimens of each type of collocation are found among the highest-scoring candidates. The manual annotation work involved in such experiments can be substantially reduced when only a random sample of the full set of candidates is evaluated (Evert and Krenn to appear).

2.3. Visualization

The most detailed and intuitive insights into the nature of different association measures are to be gained from a graphical visualization of frequency data and association scores. Association measures can be represented by surfaces in a three-dimensional parameter space, whose comparison immediately reveals why a particular candidate is accepted by one measure but rejected by another (Evert 2004: Sec. 3.3). When the candidates have been annotated manually (cf. Section 2.2), it is also possible to identify typical “frequency profiles” that correspond to certain linguistic features. Association measures that match such frequency profiles should be particularly well suited for extracting a specific subtype of collocations.

3. Availability

The UCS toolkit is free software, released under the same terms as Perl (the Artistic License). It can be downloaded from its homepage at http://collocations.sf.net/. A detailed description of the kinds of experiments that are possible with the UCS toolkit as well as their mathematical background is given by Evert (2004). The software distribution includes demonstration scripts and data sets, so that all examples and graphs presented there can be reproduced.

References

Dunning, T. E. (1998). *Finding Structure in Text, Genome and Other Symbolic Sequences.* Ph.D. thesis, Department of Computer Science, University of Sheffield.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. thesis, IMS, University of Stuttgart. [manuscript available from http://www.collocations.de/EK/]

Evert, S. and B. Krenn (to appear). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language.*

Pedersen, T. (1996). Fishing for exactness. In: *Proceedings of the South-Central SAS Users Group Conference,* Austin, TX.

Yates, F. (1984). Tests of significance for 2x2 contingency tables. *Journal of the Royal Statistical Society, Series A*, **147**(3), 426 – 493.