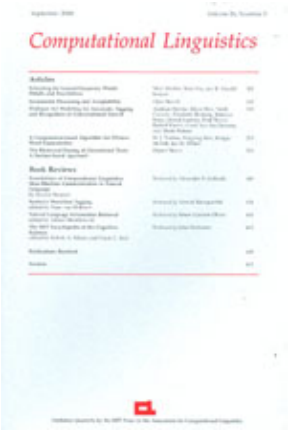

Quarterly (March, June, September, December)
160pp. per issue
6 3/4 x 10
Founded: 1974
2018 Impact Factor: 1.319
2018 Google Scholar h5-index: 32
ISSN: 0891-2017
E-ISSN: 1530-9312

Extracting the Lowest-Frequency Words: Pitfalls and Possibilities

Marc Weeber, Rein Vos and R. Harald Baayen

Posted Online March 13, 2006
https://doi.org/10.1162/089120100561719

Computational Linguistics
Volume 26 | Issue 3 | September 2000
p.301-317

Abstract Authors

In a medical information extraction system, we use common word association techniques to extract side-effect-related terms. Many of these terms have a frequency of less than five. Standard word-association-based applications disregard the lowest-frequency words, and hence disregard useful information. We therefore devised an extraction system for the full word frequency range. This system computes the

significance of association by the log-likelihood ratio and Fisher's exact test. The output of the system shows a recurrent, corpus-independent pattern in both recall and the number of significant words. We will explain these patterns by the statistical behavior of the lowest-frequency words. We used Dutch verb-particle combinations as a second and independent collocation extraction application to illustrate the generality of the observed phenomena. We will conclude that a) word-association-based extraction systems can be enhanced by also considering the lowest-frequency words, b) significance levels should not be fixed but adjusted for the optimal window size, c) hapax legomena, words occurring only once, should be disregarded a priori in the statistical analysis, and d) the distribution of the targets to extract should be considered in combination with the extraction method.

Lexicon-Based Methods for Sentiment Analysis (436 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

A Systematic Comparison of Various Statistical Alignment Models (174 times)
Franz Josef Och et al.
Computational Linguistics
Volume: 29, Issue: 1, pp. 19-51

Opinion Word Expansion and Target Extraction through Double Propagation (147 times)
Guang Qiu et al.
Computational Linguistics
Volume: 37, Issue: 1, pp. 9-27

(Note that the Most Cited numbers are based on Crossref's Cited-by service and reflect citation information for the past 24 months.)

Site Help