



The MIT Press

Journals

[Sign In / Register](#)
[Books](#)
[Journals](#)
[Digital](#)
[Resources](#)
[About](#)
[Contact](#)


Home | Computational Linguistics | List Article navigation
of Issues | Volume 26 , No. 3 |
Extracting the Lowest-Frequency Words:
Pitfalls and Possibilities



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

Journal

Resources

Editorial Info

Abstracting and
Indexing

Release Schedule

Advertising Info

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>

© 2000 Association for Computational Linguistics

Computational Linguistics

Volume 26 | Issue 3 | September 2000

p.301-317

 **Download Options** >

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

Author Resources

Submission Guidelines
Publication Agreement
Author Reprints

Reader Resources

Rights and Permissions
Most Read
Most Cited

More About Computational Linguistics ▼

Metrics ▼



13 Total citations

0 Recent citations

1.78 Field Citation Ratio

n/a Relative Citation Ratio

Open Access ▼



Computational Linguistics Computational Linguistics is Open Access. All content is freely available in

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.

Forthcoming

Most Read

[See More](#)

Lexicon-Based Methods for Sentiment Analysis (14129 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

Computational Linguistics and Deep Learning (10558 times)
Christopher D. Manning
Computational Linguistics
Volume: 41, Issue: 4, pp. 701-707


Near-Synonymy and Lexical Choice (3688 times)
Philip Edmonds et al.
Computational Linguistics
Volume: 28, Issue: 2, pp. 105-144


(Note that the Most Read numbers are based on the number of full text downloads over the last 12 months.)


Most Cited

[See More](#)

electronic format (Full text HTML, PDF, and PDF Plus) to readers across the globe. All articles are published under a [CC BY-NC-ND 4.0 license](#). For more information on allowed uses, please view the [CC license](#). [Support OA at MITP](#)

 **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.)

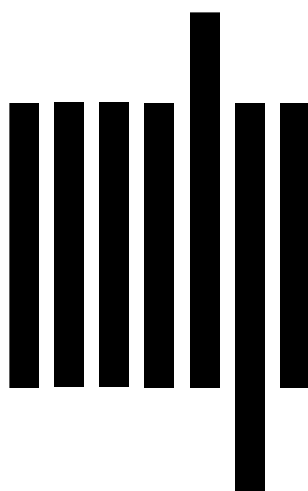
 **Download** >
Options

Favorite  Sign up for Alerts 

Download Citation  RSS TOC 

RSS Citation  Submit your article

[Support OA at MITP](#) 



Journals

Terms & Conditions

Privacy Statement

Contact Us

Books

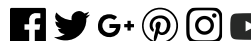
US

UK

Connect

One Rogers Street
Cambridge MA
02142-1209

Suite 2, 1 Duchess
Street London,
W1W 6AN, UK



© 2018 The MIT Press
Technology Partner:
[Atypon Systems, Inc.](#)
[CrossRef Member](#)
[COUNTER Member](#)
The MIT Press colophon is registered in the

U.S. Patent and Trademark Office.

