The MIT Press

Journals

Sign In / Register

Books   **Journals**   **Digital**   **Resources**   **About**   **Contact**

Home | Computational Linguistics | List Article navigation of Issues | Volume 37 , No. 4 |
Levenshtein Distances Fail to Identify Language Relationships Accurately

Computational Linguistics

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

Journal Resources

Editorial Info
Abstracting and Indexing
Release Schedule
Advertising Info

Author Resources

Submission Guidelines
Publication Agreement

Levenshtein Distances Fail to Identify Language Relationships Accurately

Simon J. Greenhill

Download Options ⟩

Abstract   **Authors**

The Levenshtein distance is a simple distance metric derived from the number of edit operations needed to transform one string into another. This metric has received recent attention as a means of automatically classifying languages into genealogical subgroups. In this article I test the performance of the Levenshtein distance for classifying languages by

Reader Resources

Rights and Permissions
Most Read
Most Cited

More About Computational Linguistics ⌄

Metrics ⌄

11

11 Total citations

3 Recent citations

4.58 Field Citation Ratio

n/a Relative Citation Ratio

Open Access ⌄

Computational Linguistics Computational Linguistics is Open Access. All content is freely available in electronic format (Full text HTML, PDF, and PDF Plus) to readers across the

subsampling three language subsets from a large database of Austronesian languages. Comparing the classification proposed by the Levenshtein distance to that of the comparative method shows that the Levenshtein classification is correct only 40% of time. Standardizing the orthography increases the performance, but only to a maximum of 65% accuracy within language subgroups. The accuracy of the Levenshtein classification decreases rapidly with phylogenetic distance, failing to discriminate homology and chance similarity across distantly related languages.This poor performance suggests the need for more linguistically nuanced methods for automated language classification tasks.

Forthcoming

Most Read

See More

Lexicon-Based Methods for Sentiment Analysis (14019 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

Computational Linguistics and Deep Learning (10513 times)
Christopher D. Manning
Computational Linguistics
Volume: 41, Issue: 4, pp. 701-707

Near-Synonymy and Lexical Choice (3658 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

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

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

(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 🔓
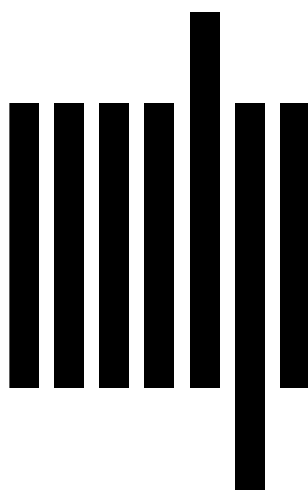
Journals

Books
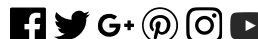
Terms & Conditions

Privacy Statement

Contact Us

US

One Rogers Street Cambridge MA 02142-1209

UK

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

Connect

© 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. Site Help