Home | Computational Linguistics | List of Issues | Volume 32 , No. 3 | Orthographic Errors in Web Pages: Toward Cleaner Web Corpora

Article navigation

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

Author Resources

Orthographic Errors in Web Pages: Toward Cleaner Web Corpora

Christoph Ringlstetter, Klaus U. Schulz and Stoyan Mihov

Posted Online August 24, 2006

https://doi.org/10.1162/coli.2006.32.3.295

Computational Linguistics

Volume 32 | Issue 3 | September 2006

p.295-340

Download Options

Abstract Authors

Since the Web by far represents the largest public repository of natural language texts, recent experiments, methods, and tools in the area of corpus linguistics often use the Web as a corpus. For applications where high accuracy is crucial, the problem has to be faced that a non-negligible number of orthographic and grammatical errors occur in Web documents. In this article we investigate the distribution of orthographic errors of various types in Web pages. As a by-product, methods are developed

for efficiently detecting erroneous pages and for marking orthographic errors in acceptable Web documents, reducing thus the number of errors in corpora and linguistic knowledge bases automatically retrieved from the Web.

Forthcoming

Most Read

See More

Lexicon-Based Methods for Sentiment Analysis (14087 times)
Maite Taboada et al.
Computational Linguistics Volume: 37, Issue: 2, pp. 267-307

Computational Linguistics and Deep Learning (10542 times)
Christopher D. Manning
Computational Linguistics Volume: 41, Issue: 4, pp. 701-707

Near-Synonymy and Lexical Choice (3675 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

(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

Submission Guidelines
Publication Agreement
Author Reprints

Reader Resources

Rights and Permissions
Most Read
Most Cited

More About Computational Linguistics

Metrics

14

14	Total citations
1	Recent citation
4.16	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 globe. All articles are

© 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