

Home | Computational Linguistics | List of Issues | Volume 32 , No. 4 | Unsupervised Multilingual Sentence Boundary Detection



Unsupervised Multilingual Sentence Boundary Detection

[Tibor Kiss](#) and [Jan Strunk](#)

Posted Online November 21, 2006

<https://doi.org/10.1162/coli.2006.32.4.485>

© 2006 Massachusetts Institute of Technology

Computational Linguistics
Volume 32 | Issue 4 | December 2006
p.485-525



Download Options



Abstract Authors

In this article, we present a language-independent, unsupervised approach to sentence boundary detection. It is based on the assumption that a large number of ambiguities in the determination of sentence boundaries can be eliminated once abbreviations have been identified. Instead of relying on orthographic clues, the proposed system is able to detect abbreviations with high accuracy using three criteria that only require information about the candidate type itself and are independent of context: Abbreviations can be defined as a very tight collocation consisting of a truncated word and a final period, abbreviations are usually short, and

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

Submission Guidelines
Publication Agreement
Author Reprints

Reader Resources

Rights and Permissions
Most Read
Most Cited

More About Computational Linguistics

Metrics



Twitter (5)

Mendeley (215)
CiteULike (7)



73 Total citations
16 Recent citations
24 Field Citation Ratio
n/a Relative Citation Ratio

Open Access

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

abbreviations sometimes contain internal periods. We also show the potential of collocational evidence for two other important subtasks of sentence boundary disambiguation, namely, the detection of initials and ordinal numbers. The proposed system has been tested extensively on eleven different languages and on different text genres. It achieves good results without any further amendments or language-specific resources. We evaluate its performance against three different baselines and compare it to other systems for sentence boundary detection proposed in the literature.

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

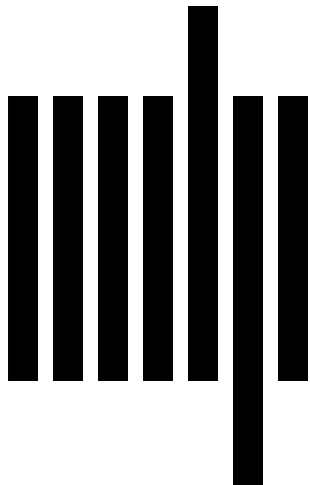
Sign up for
Favorite Alerts

in 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](#)

Download Citation  [RSS TOC](#) 

[RSS Citation](#)  [Submit your article](#)

[Support OA at MITP](#) 



[Journals](#)

[Books](#)

[US](#)

One Rogers Street
Cambridge MA 02142-1209

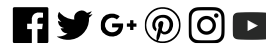
[Terms & Conditions](#)

[UK](#)

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

[Privacy Statement](#)

[Connect](#)



[Contact Us](#)

© 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](#)