Home | Computational Linguistics | List of Issues | Volume 41 , No. 3 | Large Linguistic Corpus Reduction with SCP Algorithms

Article navigation

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

ISSN: 0891-2017

E-ISSN: 1530-9312

Journal Resources

Editorial Info
Abstracting and Indexing
Release Schedule
Advertising Info

Author Resources

Submission Guidelines
Publication Agreement

Large Linguistic Corpus Reduction with SCP Algorithms

Nelly Barbot, Olivier Boëffard, Jonathan Chevelu and Arnaud Delhay

Posted Online September 08, 2015
https://doi.org/10.1162/COLI_a_00225

Computational Linguistics
Volume 41 | Issue 3 | September 2015
p.355-383

Download Options

Abstract Full Text Authors

Linguistic corpus design is a critical concern for building rich annotated corpora useful in different domains of applications. For example, speech technologies such as ASR (Automatic Speech Recognition) or TTS (Text-to-Speech) need a huge amount of speech data to train data-driven models or to produce synthetic speech. Collecting data is always related to costs (recording speech, verifying annotations, etc.),

Author Reprints

Reader Resources

Rights and Permissions
Most Read
Most Cited

More About Computational Linguistics

Metrics

Open Access

and as a rule of thumb, the more data you gather, the more costly your application will be. Within this context, we present in this article solutions to reduce the amount of linguistic text content while maintaining a sufficient level of linguistic richness required by a model or an application. This problem can be formalized as a Set Covering Problem (SCP) and we evaluate two algorithmic heuristics applied to design large text corpora in English and French for covering phonological information or POS labels. The first considered algorithm is a standard greedy solution with an agglomerative/spitting strategy and we propose a second algorithm based on Lagrangian relaxation. The latter approach provides a lower bound to the cost of each covering solution. This lower bound can be used as a metric to evaluate the quality of a reduced corpus whatever the algorithm applied. Experiments show that a suboptimal algorithm like a greedy algorithm achieves good results; the cost of its solutions is not so far from the lower bound (about 4.35% for 3-phoneme coverings). Usually, constraints in SCP are binary; we proposed here a generalization where the constraints on each covering feature can be multi-valued.

Forthcoming

Most Read

See More

Lexicon-Based Methods for Sentiment Analysis (13965 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

Computational Linguistics and Deep Learning (10500 times)
Christopher D. Manning
Computational Linguistics
Volume: 41, Issue: 4, pp. 701-707

Near-Synonymy and Lexical Choice (3653 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

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

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