The MIT Press

Journals

Books   Journals   Digital   Resources   About   Contact

Home | Computational Linguistics | List Article navigation of Issues | Volume 39 , No. 4 |
Generation of Compound Words in Statistical Machine Translation into Compounding Languages

Journal Resources

Editorial Info
Abstracting and Indexing
Release Schedule
Advertising Info

Author Resources

Submission Guidelines

Generation of Compound Words in Statistical Machine Translation into Compounding Languages

Sara Stymne, Nicola Cancedda and Lars Ahrenberg

◎ **Download Options** ⟩

Abstract   Full Text   Authors

Reader Resources

More About
Computational
Linguistics   ∨

Metrics   ∨

2   Total
citations

1   Recent
citation

0.8   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

In this article we investigate statistical machine translation (SMT) into Germanic languages, with a focus on compound processing. Our main goal is to enable the generation of novel compounds that have not been seen in the training data. We adopt a split-merge strategy, where compounds are split before training the SMT system, and merged after the translation step. This approach reduces sparsity in the training data, but runs the risk of placing translations of compound parts in non-consecutive positions. It also requires a postprocessing step of compound merging, where compounds are reconstructed in the translation output. We present a method for increasing the chances that components that should be merged are translated into contiguous positions and in the right order and show that it can lead to improvements both by direct inspection and in terms of standard translation evaluation metrics. We also propose several new methods for compound merging, based on heuristics and machine learning, which outperform previously suggested algorithms. These methods can produce novel compounds and a translation with at least the same overall quality as the baseline. For all subtasks we show that it is useful to include part-of-speech based information in the translation process, in order to handle compounds.

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.)
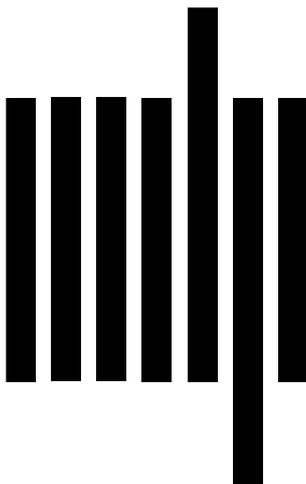
◉ Download Options ⟩

Favorite ♡

Sign up for Alerts 🔔

Download Citation ↓

RSS TOC 📶

RSS Citation 📶

Submit your article

Support OA at MITP 🔓

Journals

Books

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

Trademark Office.