BUCC, 12th Workshop on Building and Using Comparable Corpora

Special topic: **Neural Networks for Building and Using Comparable Corpora**
Extended deadline: July 11, 2019
Co-located with RANLP 2019
Varna, Bulgaria
Sep. 5, 2019
Web site: https://comparable.limsi.fr/bucc2019/
Programme and proceedings
Workshop proceedings are now available

MOTIVATION

Research on comparable corpora is active but used to be scattered among many workshops and conferences. Hence this workshop series, which bundles this research and gives it a better platform. In the language engineering and the linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical NLP applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions.

Comparable corpora have been used in a range of applications, including, Information Retrieval, Machine Translation, Cross-lingual text classification, etc. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP, for example to extract parallel corpora from comparable corpora for neural MT. As such, it is of great interest to bring together builders and users of such corpora.

TOPICS

The special topic for this year is *Neural Networks for Building and Using Comparable Corpora*. More broadly, we solicit contributions to the following topics:

Building Comparable Corpora:

- Automatic and semi-automatic methods
- Methods to mine parallel and non-parallel corpora from the Web
- Tools and criteria to evaluate the comparability of corpora
- Parallel vs non-parallel corpora, monolingual corpora
- Rare and minority languages, across language families
- Multi-media/multi-modal comparable corpora

Applications of comparable corpora:

- Human translations
- Language learning
- Cross-language information retrieval & document categorization
- Bilingual projections
- Machine translation
- Writing assistance

Mining from Comparable Corpora:

- Cross-language distributional semantics and word embeddings
- Extraction of parallel segments or paraphrases from comparable corpora
- Methods to extract parallel from non-parallel corpora (e.g. to provide for low-resource languages in neural machine translation)
- Extraction of bilingual and multilingual translations of single words and multi-word expressions; proper names, named entities, etc., from comparable corpora

IMPORTANT DATES

11 July 2019	Paper submission deadline (extended)
28 July 2019	Notification to authors
20 August 2019	Camera-ready papers due
5 September 2019	Workshop date

SUBMISSION INFORMATION

Please follow the style sheet and templates provided for the main conference at http://lml.bas.bg/ranlp2019/submissions
Papers should be submitted as a PDF file at https://www.softconf.com/ranlp2019/BUCC/. Submissions must describe original and unpublished work and range from four (4) to eight (8) pages plus unlimited references.

Reviewing will be double blind, so the papers should not reveal the authors' identity. Accepted papers will be published in the workshop proceedings.

Double submission policy: Parallel submission to other meetings or publications is possible but must be immediately notified to the workshop organizers.

For further information, please contact Serge Sharoff <S (dot) Sharoff (at) leeds (dot) ac (dot) uk>

Plain-text CFP : bucc2019-cfp.txt
PDF CFP : bucc2019-cfp.pdf
Last modified: 9 November 2019

ORGANISERS

Serge Sharoff (University of Leeds, United Kingdom), Chair

Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany), Shared task organiser

Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France), Information chair

SCIENTIFIC COMMITTEE

Ahmet Aker (University of Sheffield, UK)
Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
Thierry Etchegoyhen (VicomTech, Spain)
Gregory Grefenstette (Florida Institute for Human and Machine Cognition, USA)
Hitoshi Isahara (Toyohashi University of Technology)
Kyo Kageura (University of Tokyo, Japan)
Philippe Langlais (Université de Montréal, Canada)
Yves Lepage (Waseda University, Japan)
Shervin Malmasi (Harvard Medical School, Boston, MA, US)
Pabitra Mitra Indian Institute of Technology Kharagpur, India
Michael Mohler (Language Computer Corp., US)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Lene Offersgaard (University of Copenhagen, Denmark)
Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)
Nasredine Semmar (CEA LIST, France)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France)

SHARED TASK

Previous BUCC shared tasks and datasets:

- Identifying parallel sentences in comparable corpora (BUCC 2018, 2017)
- Identifying comparable text (BUCC 2015)