BUCC, 6th Workshop on Building and Using Comparable Corpora

Co-located with ACL 2013
Sofia, Bulgaria
8 August 2013

Extended deadline for papers: 3 May 2013
http://comparable.limsi.fr/bucc2013/
Submission: https://www.softconf.com/acl2013/BUCC2013/
Invited Speaker: **Hinrich Schütze**, University of Munich

INVITED SPEAKER

Hinrich Schütze University of Munich

Three dimensions of comparable corpora: same or different language, given or inferred comparability, means to an end or end in itself

MOTIVATION

In the language engineering and the linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, 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 intra-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. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

Parallel corpora are a key resource as training data for statistical machine translation, and for building or extending bilingual lexicons and terminologies. However, beyond a few language pairs such as English-French or English-Chinese and a few contexts such as parliamentary debates or legal texts, they remain a scarce resource, despite the creation of automated methods to collect parallel corpora from the Web. Interest in non-parallel forms of comparable corpora in language engineering primarily ensued from the scarcity of parallel corpora. This has motivated research concerning the use of comparable corpora: pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. Non-parallel yet comparable corpora overcome the two limitations of parallel corpora, since sources for original, monolingual texts are much more abundant than translated texts. However, because of their nature, mining translations in comparable corpora is much more challenging than in parallel corpora. What constitutes a good comparable corpus, for a given task or per se, also requires specific attention: while the definition of a parallel corpus is fairly straightforward, building a non-parallel corpus requires control over the selection of source texts in both languages.

TOPICS

The special theme for this edition is terminology mining, which featured in a number of submissions in the past years, and this time it will serve as the highlighted theme for the workshop.

In addition to this special theme, we solicit contributions including but not limited to other relevant topics:

Building Comparable Corpora:

- Human translations
- 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:

- Extraction of parallel segments or paraphrases from comparable corpora
- Extraction of bilingual and multilingual translations of single words and multi-word expressions; proper names, named entities, etc.

IMPORTANT DATES

3 May 2013	Extended deadline for submission
24 May 013	Notification of acceptance
7 June 2013	Camera-ready deadline
8 August 2013	Workshop

SUBMISSION INFORMATION

Submissions should follow the ACL 2013 length and formatting requirements for long papers of maximum eight (8) pages of content plus two (2) extra pages for references, found at http://www.acl2013.org/call.html. They should be submitted as PDF documents to the following address:

https://www.softconf.com/acl2013/BUCC2013/

Papers will be blind reviewed by at least two members of the Program Committee. Therefore, authors' names and affiliations should not appear in the paper. Accepted papers will be published in the workshop proceedings.

Authors may submit the same paper at several meetings, but a paper published at this workshop cannot also be published elsewhere. In case of double submission, the authors must notify the workshop organizers in a separate e-mail, so we know that the paper might be withdrawn depending on the results at some other meeting. However, after notification authors will be asked to make a final decision.

For further information, please contact

Serge Sharoff mailto:s(erase_dot)sharoff(erase_at)leeds(erase_dot)ac(erase_dot)uk

Plain-text CFP : bucc2013-cfp.txt
PDF CFP : bucc2013-cfp.pdf
Last modified: 6 July 2013 Date: 2013/07/06 15:06:29 (Revision: 1.12)

ORGANISERS

Serge Sharoff University of Leeds, UK (Chair)

Pierre Zweigenbaum LIMSI, CNRS, Orsay, and ERTIM, INALCO, Paris, France

Reinhard Rapp Universities of Mainz, Germany, and Aix-Marseille, France

SCIENTIFIC COMMITTEE

Caroline Barrière (CRIM, Montréal, Canada)

Chris Biemann (TU Darmstadt, Germany)

Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)

Kurt Eberle (Lingenio, Heidelberg, Germany)

Andreas Eisele (European Commission, Luxembourg)

Gregory Grefenstette (Exalead, Paris, France)

Silvia Hansen-Schirra (University of Mainz, Germany)

Hitoshi Isahara (Toyohashi University of Technology)

Kyo Kageura (University of Tokyo, Japan)

Natalie Kübler (Université Paris Diderot, France)

Philippe Langlais (Université de Montréal, Canada)

Emmanuel Morin (Université de Nantes, France)

Dragos Stefan Munteanu (Language Weaver, Inc., US)

Lene Offersgaard (University of Copenhagen, Denmark)

Reinhard Rapp (Université Aix-Marseille, France)

Serge Sharoff (University of Leeds, UK)

Mandel Shi (Xiamen University, China)

Michel Simard (National Research Council Canada)

Richard Sproat (OGI School of Science & Technology, US)

Justin Washtell (365 Media Inc, US)

Michael Zock (Laboratoire d'Informatique Fondamentale, CNRS, Marseille)

Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)