4th Workshop on Building and Using Comparable Corpora

Co-located with ACL-HLT 2011
Portland, Oregon
24 June 2011

Deadline for papers: Extended to 11 April 2011
http://comparable.limsi.fr/bucc2011-comparable-corpora/
Submission: https://www.softconf.com/acl2011/comparable/

Endorsed by ACL SIGWAC (Special Interest Group on Web as Corpus)
and FLaReNet (Fostering Language Resources Network)

INVITED SPEAKER

Kevin Knight Information Sciences Institute, USC

"Putting a Value on Comparable Data"

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 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. 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.

With the advent of online data, the potential for building and exploring comparable corpora is growing exponentially. Comparable documents in languages that are very different from each other pose special challenges as very often, the non-parallelness in sentences can result from cultural and political differences.

TOPICS

The theme of the workshop will be “Comparable Corpora and the Web”. Nevertheless we solicit contributions to other topics as well, including the following:

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

11 April 2011	Deadline for submission
27 April 2011	Notification
6 May 2011	Final version
24 June 2011	Workshop

SUBMISSION INFORMATION

Submissions should follow the ACL HLT 2011 length and formatting requirements for long papers of six to eight (6–8) pages of content with two (2) additional pages of references, found at http://www.acl2011.org/call.shtml. They should be submitted as PDF documents to the following address:

https://www.softconf.com/acl2011/comparable/

For further information, please contact

Pierre Zweigenbaum mailto:pz(erase_at)limsi(erase_dot)fr

Plain-text CFP : bucc2011-cfp.txt
PDF CFP : bucc2011-comparable-corpora.pdf
Last modified: Date: 2011/04/03 16:07:12 (Revision: 1.11)

ORGANISERS

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

Reinhard Rapp Universities of Mainz (Germany) and Tarragona (Spain)

Serge Sharoff University of Leeds (UK)

SCIENTIFIC COMMITTEE

Srinivas Bangalore (AT&T Labs, US)
Caroline Barrière (National Research Council Canada)
Chris Biemann (Microsoft / Powerset, San Francisco, US)
Lynne Bowker (University of Ottawa, Canada)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Kurt Eberle (Lingenio, Heidelberg, Germany)
Andreas Eisele (European Commission, Luxembourg)
Pascale Fung (Hong Kong University of Science & Technology, Hong Kong)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (Exalead, Paris, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (National Institute of Information and Communications Technology, Kyoto, Japan)
Kyo Kageura (University of Tokyo, Japan)
Adam Kilgarriff (Lexical Computing Ltd, UK)
Natalie Kübler (Université Paris Diderot, France)
Philippe Langlais (Université de Montréal, Canada)
Tony McEnery (Lancaster University, UK)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Reinhard Rapp (University of Tarragona, Spain)
Sujith Ravi (Information Sciences Institute, University of Southern California, US)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council, Canada)
Monique Slodzian (INALCO, Paris, France)
Richard Sproat (OGI School of Science & Technology, US)
Benjamin T'sou (The Hong Kong Institute of Education, Hong Kong)
Yujie Zhang (National Institute of Information and Communications Technology, Japan)
Michael Zock (Laboratoire d'Informatique Fondamentale, CNRS, Marseille, France)
Pierre Zweigenbaum (LIMSI-CNRS, France)