Building and Using Comparable Corpora

LREC 2008 workshop
Marrakech, Morocco
31 May 2008

Call for papers: Deadline extended to 18 February 2008
http://www.limsi.fr/~pz/lrec2008-comparable-corpora/
Submission: https://www.softconf.com/LREC2008/CompCorp/submit.html

Context and focus

Research in comparable corpora is motivated by the scarcity of parallel corpora. Parallel corpora are a key resource to mine translations for statistical machine translation or 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. A more fundamental limitation is that translated texts, whatever the skills of translators, are generally influenced by the very translation process and by the language of source texts, so that they may not be fully adequate for the task at hand.

This has motivated research into the use of comparable corpora: pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. 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 comparable corpus requires control over the selection of source texts in both languages.

Topics

This workshop aims to bring together researchers interested in the constitution and use of comparable corpora. Contributions are solicited on the constitution and application of comparable corpora, including the following topics:

Applications of comparable corpora:

- tools for translators;
- tools for language learning;
- cross-language information retrieval;
- cross-language document categorization;
- machine translation;
- monolingual comparable corpora for writing assistance;
- extraction of parallel segments in comparable corpora.

Units aligned in comparable corpora:

- single words and multi-word expressions; proper names; alignment across different scripts.

Constitution of comparable corpora:

- criteria of comparability;
- degree of comparability;
- methods for mining comparable corpora.

Important dates

18 February 2008	Extended deadline for submission
14 March 2008	Notification
31 March 2008	Final version
31 May 2008	Workshop

Organisers

Pierre Zweigenbaum LIMSI, CNRS, Orsay, France

Eric Gaussier LIG, Université J. Fourier, Grenoble, France

Pascale Fung Department of Electronic & Computer Engineering, University of Science & Technology, Hong Kong

Submission information

We expect short papers of max 3500 words (about 4-6 pages) describing research addressing one of the above topics, to be submitted as PDF documents at the following address:

https://www.softconf.com/LREC2008/CompCorp/submit.html

In case of problem with the above address, please send your PDF document by email to:

Pierre Zweigenbaum mailto:pz@limsi.fr

The final papers should not have more than 6 pages, adhering to the stylesheet that will be adopted for the LREC Proceedings (to be announced later on the Conference web site).

Scientific Committee

Lynne Bowker (University of Ottawa, Canada)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (CEA/LIST, Fontenay-aux-Roses, France)
Pascale Fung (University of Science & Technology, Hong Kong)
Natalie Kübler (Université Paris Diderot, France)
Tony McEnery (Lancaster University, UK)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Information Sciences Institute, Marina Del Rey, USA)
Carol Peters (ISTI-CNR, Pisa, Italy)
Reinhard Rapp (Johannes Gutenberg-Universität Mainz, Germany)
Serge Sharoff (University of Leeds, UK)
Monique Slodzian (INALCO, Paris, France)
Richard Sproat (University of Illinois at Urbana-Champaign, USA)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

Plain-text CFP : cfp-lrec2008-comparable-corpora.txt
PDF CFP : cfp-lrec2008-comparable-corpora.pdf