ACL-IJCNLP 2009

BUCC 2009

2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora

Proceedings of the Workshop

6 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-932432-53-4 / 1-932432-53-1

Introduction

Research in comparable corpora has been motivated by two main reasons in the language engineering and the linguistics communities. 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-language information retrieval. In linguistics, on the other hand, comparable corpora are of interest themselves in providing intra-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one to many languages, that are comparable in content and form in various degrees and dimensions. It was pointed out that parallel corpora are at one end of the spectrum of comparability whereas quasi-comparable corpora are at the other end. We believe that the linguistic definitions and observations in comparable corpora can improve methods to mine such corpora for applications to 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 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. 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-parallel-ness in sentences can result from cultural and political differences.

Following the success of the first workshop on Building and Using Comparable Corpora at LREC 2008 in Marrakech, this second workshop again brings together language engineers as well as linguists interested in the constitution and use of comparable corpora, ranging from parallel to non-parallel corpora. In the larger context of the joint ACL-IJCNLP conference, this time the workshop specifically aimed to solicit contributions from researchers in different geographical regions, in order to highlight in particular the issues with comparable corpora across languages that are very different from each other, such as across Asian and European languages. Research in minority languages is also of particular interest. We are very glad to include papers on languages as varied as Arabic, Chinese, English, French, Japanese, Uyghur and even sign language.

We would like to thank all people who in one way or another helped in making this workshop a success. Our particular thanks go to Ken Church for accepting to give the invited presentation, to the participants of the panel discussion, to the members of the program committee, to the ACL-IJCNLP workshop co-chairs Jimmy Lin and Yuji Matsumoto, and to the members of the local organizing committee. Last but not least we would like to thank our authors and the participants of the workshop.

Pascale Fung, Pierre Zweigenbaum, Reinhard Rapp

Organizers:

Pascale Fung (Hong Kong University of Science & Technology—HKUST)
Pierre Zweigenbaum (LIMSI-CNRS, France)
Reinhard Rapp (University of Mainz, Germany & University of Tarragona, Spain)

Program Committee:

Askar Hamdulla (Xinjiang University, China)
Srinivas Bangalore (AT&T Labs, US)
Lynne Bowker (University of Ottawa, Canada)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (Exalead, Paris, France)
Satoshi Isahara (National Institute of Information and Communications Technology, Japan)
Min-Yen Kan (National University of Singapore)
Adam Kilgarriff (Lexical Computing Ltd, UK)
Philippe Langlais (Université de Montral, Canada)
Rada Mihalcea (University of North Texas, US)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Grace Ngai (Hong Kong Polytechnic University, Hong Kong)
Carol Peters (ISTI-CNR, Pisa, Italy)
Serge Sharoff (University of Leeds, UK)
Richard Sproat (OGI School of Science & Technology, US)
Mandel Shi (Xiamen University, China)
Yujie Zhang (National Institute of Information and Communications Technology, Japan)

Invited Speaker:

Kenneth Ward Church (Chief Scientist, Human Language Technology Center of Excellence, Johns Hopkins University, US)

Workshop Technical Support:

Ricky Chan Ho Yin (Hong Kong University of Science & Technology)

Table of Contents

Conference Program

Thursday, August 6, 2009

8:45 Welcome and Introduction

Invited Presentation

9:00 *Repetition and Language Models and Comparable Corpora*
Ken Church

10:00 Coffee break

Information Extraction and Summarization

10:30 *Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora*
Louise Deléger and Pierre Zweigenbaum

10:55 *An Extensible Crosslinguistic Readability Framework*
Jesse Kirchner, Justin Nuger and Yi Zhang

11:20 *An Analysis of the Calque Phenomena Based on Comparable Corpora*
Marie Garnier and Patrick Saint-Dizier

11:45 *Active Learning of Extractive Reference Summaries for Lecture Speech Summarization*
Jian Zhang and Pascale Fung

12:10 Lunch break

Thursday, August 6, 2009 (continued)

Statistical Machine Translation

13:50 *Train the Machine with What It Can Learn—Corpus Selection for SMT*
Xiwu Han, Hanzhang Li and Tiejun Zhao

14:15 *Mining Name Translations from Comparable Corpora by Creating Bilingual Information Networks*
Heng Ji

14:40 *Chinese-Uyghur Sentence Alignment: An Approach Based on Anchor Sentences*
Samat Mamitimin and Min Hou

15:05 *Exploiting Comparable Corpora with TER and TERp*
Sadaf Abdul Rauf and Holger Schwenk

15:30 Coffee break

Building Comparable Corpora

16:00 *Compilation of Specialized Comparable Corpora in French and Japanese*
Lorraine Goeuriot, Emmanuel Morin and Béatrice Daille

16:25 *Toward Categorization of Sign Language Corpora*
Jérémie Segouat and Annelies Braffort

16:50 **Panel Session**
Multilingual Information Processing: from Parallel to Comparable Corpora

17:50 End of Workshop