ACL HLT 2011

4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web BUCC

Proceedings of the Workshop

24 June, 2011
Portland, Oregon, USA

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53704 USA

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

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-13 9781937284015

Introduction

Comparable corpora are collections of documents that are comparable in content and form in various degrees and dimensions. This definition includes many types of parallel and non-parallel multilingual corpora, but also sets of monolingual corpora that are used for comparative purposes. Research on comparable corpora is active but used to be scattered among many workshops and conferences. The workshop series on “Building and Using Comparable Corpora” (BUCC) aims at promoting progress in this exciting emerging field by bundling its research, thereby making it more visible and giving it a better platform.

Following the three previous editions of the workshop which took place at LREC 2008 in Marrakech, at ACL-IJCNLP 2009 in Singapore, and at LREC 2010 in Malta, this year the workshop was co-located with ACL-HLT in Portland and its theme was “Comparable Corpora and the Web”. Among the topics solicited in the call for papers, three are particularly well represented in this year’s workshop:

- Mining word translations from comparable corpora, an early favorite, continues to be explored;
- Identifying parallel sub-sentential segments from comparable corpora is gaining impetus;
- Building comparable corpora and assessing their comparability is a basic need for the field.

Additionally, statistical machine translation and cross-language information access are recurring motivating applications.

We would like to thank all people who in one way or another helped in making this workshop a particularly successful. This year the workshop has been formally endorsed by ACL SIGWAC (Special Interest Group on Web as Corpus) and FLaReNet (Fostering Language Resources Network). Our special thanks go to Kevin Knight for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the ACL-HLT workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Pierre Zweigenbaum, Reinhard Rapp, Serge Sharoff

Organizers:

Pierre Zweigenbaum (LIMSI, CNRS, Orsay, and ERTIM, INALCO, Paris, France)
Reinhard Rapp (Universities of Leeds, UK, and Mainz, Germany)
Serge Sharoff (University of Leeds, UK)

Invited Speaker:

Kevin Knight (Information Sciences Institute, USC)

Program Committee:

Srinivas Bangalore (AT&T Labs, USA)
Caroline Barrière (National Research Council Canada)
Chris Biemann (Microsoft / Powerset, San Francisco, USA)
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., USA)
Reinhard Rapp (Universities of Leeds, UK, and Mainz, Germany)
Sujith Ravi (Information Sciences Institute, University of Southern California, USA)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council, Canada)
Monique Slodzian (INALCO, Paris, France)
Richard Sproat (OGI School of Science & Technology, USA)
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, and ERTIM, INALCO, Paris, France)

Table of Contents

Invited Presentation

Oral Presentations

Poster Presentations

Conference Program

Friday June 24, 2011

Session 1: (09:00) Bilingual Lexicon Extraction From Comparable Corpora

9:00 *Learning the Optimal Use of Dependency-parsing Information for Finding Translations with Comparable Corpora*
Daniel Andrade, Takuya Matsuzaki and Junichi Tsujii

9:20 *Building and Using Comparable Corpora for Domain-Specific Bilingual Lexicon Extraction*
Darja Fišer, Nikola Ljubešić, Špela Vintar and Senja Pollak

9:40 *Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora*
Emmanuel Morin and Emmanuel Prochasson

10:00 *Bilingual Lexicon Extraction from Comparable Corpora as Metasearch*
Amir Hazem, Emmanuel Morin and Sebastian Peña Saldarriaga

Session 2: (11:00) Extracting Parallel Segments From Comparable Corpora

11:00 *Two Ways to Use a Noisy Parallel News Corpus for Improving Statistical Machine Translation*
Souhir Gahbiche-Braham, Hélène Bonneau-Maynard and François Yvon

11:20 *Paraphrase Fragment Extraction from Monolingual Comparable Corpora*
Rui Wang and Chris Callison-Burch

11:40 *Extracting Parallel Phrases from Comparable Data*
Sanjika Hewavitharana and Stephan Vogel

12:00 *Active Learning with Multiple Annotations for Comparable Data Classification Task*
Vamshi Ambati, Sanjika Hewavitharana, Stephan Vogel and Jaime Carbonell

Friday June 24, 2011 (continued)

Session 3: (14:00) Invited Presentation

14:00 *Putting a Value on Comparable Data* / *The Copiale Cipher*
Kevin Knight (in collaboration with Beáta Megyesi and Christiane Schaefer)

Session 4: (14:50) Poster Presentations (including Booster Session)

14:50 *Unsupervised Alignment of Comparable Data and Text Resources*
Anja Belz and Eric Kow

14:55 *Cross-lingual Slot Filling from Comparable Corpora*
Matthew Snover, Xiang Li, Wen-Pin Lin, Zheng Chen, Suzanne Tamang, Mingmin Ge, Adam Lee, Qi Li, Hao Li, Sam Anzaroot and Heng Ji

15:00 *Towards a Data Model for the Universal Corpus*
Steven Abney and Steven Bird

15:05 *An Expectation Maximization Algorithm for Textual Unit Alignment*
Radu Ion, Alexandru Ceauşu and Elena Irimia

15:10 *Building a Web-Based Parallel Corpus and Filtering Out Machine-Translated Text*
Alexandra Antonova and Alexey Misyurev

15:15 *Language-Independent Context Aware Query Translation using Wikipedia*
Rohit Bharadwaj G and Vasudeva Varma

Session 5: (16:00) Building and Assessing Comparable Corpora

16:00 *How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives*
Bruno Cartoni, Sandrine Zufferey, Thomas Meyer and Andrei Popescu-Belis

16:20 *Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia.*
Alexandre Patry and Philippe Langlais

16:40 *Comparable Fora*
Johanka Spoustová and Miroslav Spousta