ACL 2013

Sixth Workshop on Building and Using Comparable Corpora

Proceedings of the Workshop

August 8, 2013
Sofia, Bulgaria

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

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: 978-1-937284-60-2

Introduction to BUCC 2013

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 five previous editions of the workshop which took place in Africa (LREC'08 in Marrakech), America (ACL'11 in Portland), Asia (ACL-IJCNLP'09 in Singapore), Europe (LREC'10 in Malta) and also on the border between Asia and Europe (LREC'12 in Istanbul), the workshop this year is co-located with ACL'13 in Sofia, Bulgaria. The main theme for the current edition is "Terminology mining". We have received 27 submissions, accepted 10 oral presenations and 7 posters, including four oral presentations on the special topic.

We would like to thank all people who in one way or another helped in making this workshop once again a success. Our special thanks go to Hinrich Schütze 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'13 workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum

Organizers:

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

Invited Speaker:

Hinrich Schütze, Ludwig-Maximilians-Universität München, Germany

Scientific Committee:

Caroline Barriere (Computer Research Institute of Montreal, 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)
Dragos Munteanu (Language Weaver, US)
Emmanuel Morin (Université de Nantes, France)
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)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Justin Washtell (365 Media Inc, US)
Michael Zock (Laboratoire d'Informatique Fondamentale, CNRS, Marseille)
Pierre Zweigenbaum (LIMSI-CNRS, France)

Table of Contents

Cross-lingual WSD for Translation Extraction from Comparable Corpora
Marianna Apidianaki, Nikola Ljubešić and Darja Fišer 1

Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tool
Hong-seok Kwon, Hyeong-won Seo and Jae-hoon Kim 11

Using WordNet and Semantic Similarity for Bilingual Terminology Mining from Comparable Corpora
Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum 16

A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora
Amir HAZEM and Emmanuel MORIN 24

Chinese–Japanese Parallel Sentence Extraction from Quasi–Comparable Corpora
Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi 34

A modular open-source focused crawler for mining monolingual and bilingual corpora from the web
Vassilis Papavassiliou, Prokopis Prokopidis and Gregor Thurmair 43

Building basic vocabulary across 40 languages
Judit Acs, Katalin Pajkossy and Andras Kornai 52

Scientific registers and disciplinary diversification: a comparable corpus approach
Elke Teich, Stefania Degaetano-Ortlieb, Hannah Kermes and Ekaterina Lapshinova-Koltunski . . 59

Improving MT System Using Extracted Parallel Fragments of Text from Comparable Corpora
Rajdeep Gupta, Santanu Pal and Sivaji Bandyopadhyay 69

VARTRA: A Comparable Corpus for Analysis of Translation Variation
Ekaterina Lapshinova-Koltunski 77

Building Ontologies from Collaborative Knowledge Bases to Search and Interpret Multilingual Corpora
Yegin Genc, Elizabeth Lennon, Winter Mason and Jeffrey Nickerson 87

Using a Random Forest Classifier to recognise translations of biomedical terms across languages
Georgios Kontonatsios, Ioannis Korkontzelos, Sophia Ananiadou and Jun'ichi Tsujii 95

Comparing Multilingual Comparable Articles Based On Opinions
Motaz Saad, David Langlois and Kamel Smaili 105

Mining for Domain-specific Parallel Text from Wikipedia
Magdalena Plamada and Martin Volk 112

Gathering and Generating Paraphrases from Twitter with Application to Normalization
Wei Xu, Alan Ritter and Ralph Grishman 121

Learning Comparable Corpora from Latent Semantic Analysis Simplified Document Space
Ekaterina Stambolieva 129

Finding More Bilingual Webpages with High Credibility via Link Analysis
Chengzhi Zhang, Xuchen Yao and Chunyu Kit 138

Conference Program

Session: Invited talk

9:00–10:00 *Three dimensions of comparable corpora: same or different language, given or inferred comparability, means to an end or end in itself*
Hinrich Schütze

Session: (10:00-12:30) Terminology

10:00–10:30 *Cross-lingual WSD for Translation Extraction from Comparable Corpora*
Marianna Apidianaki, Nikola Ljubešić and Darja Fišer

Coffee break: (10:30-11:00)

11:00–11:30 *Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tool*
Hong-seok Kwon, Hyeong-won Seo and Jae-hoon Kim

11:30–12:00 *Using WordNet and Semantic Similarity for Bilingual Terminology Mining from Comparable Corpora*
Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum

12:00–12:30 *A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora*
Amir Hazem and Emmanuel Morin

Session: (14:00-15:00) Comparable corpora

14:00–14:30 *Finding More Bilingual Webpages with High Credibility via Link Analysis*
Chengzhi Zhang, Xuchen Yao and Chunyu Kit

14:30–15:00 *A modular open-source focused crawler for mining monolingual and bilingual corpora from the web*
Vassilis Papavassiliou, Prokopis Prokopidis and Gregor Thurmair

Session: (15:00-15:30) Posters with Booster Session

15:00–15:03 *Building basic vocabulary across 40 languages*
Judit Acs, Katalin Pajkossy and Andras Kornai

15:04–15:07 *Scientific registers and disciplinary diversification: a comparable corpus approach*
Elke Teich, Stefania Degaetano-Ortlieb, Hannah Kermes and Ekaterina Lapshinova-Koltunski

15:08–15:11 *Improving MT System Using Extracted Parallel Fragments of Text from Comparable Corpora*
Rajdeep Gupta, Santanu Pal and Sivaji Bandyopadhyay

15:12–15:15 *VARTRA: A Comparable Corpus for Analysis of Translation Variation*
Ekaterina Lapshinova-Koltunski

8 August 2013 (continued)

15:16–15:19 *Building Ontologies from Collaborative Knowledge Bases to Search and Interpret Multilingual Corpora*
Yegin Genc, Elizabeth Lennon, Winter Mason and Jeffrey Nickerson

15:20–15:23 *Using a Random Forest Classifier to recognise translations of biomedical terms across languages*
Georgios Kontonatsios, Ioannis Korkontzelos, Sophia Ananiadou and Jun'ichi Tsujii

15:24–15:27 *Comparing Multilingual Comparable Articles Based On Opinions*
Motaz Saad, David Langlois and Kamel Smaili

Coffee break: (15:30-16:00)

Session: (16:00-18:00) Comparable corpora

16:00–16:30 *Mining for Domain-specific Parallel Text from Wikipedia*
Magdalena Plamada and Martin Volk

16:30–17:00 *Gathering and Generating Paraphrases from Twitter with Application to Normalization*
Wei Xu, Alan Ritter and Ralph Grishman

17:00–17:30 *Learning Comparable Corpora from Latent Semantic Analysis Simplified Document Space*
Ekaterina Stambolieva

17:30–18:00 *Chinese–Japanese Parallel Sentence Extraction from Quasi–Comparable Corpora*
Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi