ACL-IJCNLP 2015

Eighth Workshop on Building and Using Comparable Corpora

Proceedings of the Workshop

July 30, 2015
Beijing, China

Production and Manufacturing by
Taberg Media Group AB
Box 94, 562 02 Taberg
Sweden

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 978-1-941643-60-0

Introduction to BUCC 2015

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.

Research on comparable corpora spans a number of topics from machine translation to contrastive linguistics. Distributional analysis, a topic which has seen renewed interest in recent years, has formed the core of a large part of the methods used to identify translations in comparable corpora. As a matter of fact, the standard techniques of word alignment in comparable corpora can be seen as methods for cross-language distributional semantics.

Following the seven previous editions of the workshop which took place at LREC 2008 (Marrakech), ACL-IJCNLP 2009 (Singapore), LREC 2010 (Malta), ACL-HLT 2011 (Portland), LREC 2012 (Istanbul), ACL 2013 (Sofia), LREC 2014 (Reykjavik), the workshop this year is co-located with ACL-IJCNLP 2015 in Beijing, China.

This year's workshop also hosts a companion shared task which is the first evaluation exercise on the identification of comparable texts: given a large multilingual collection of texts derived from Wikipedia, detecting the most similar texts across languages. Evaluation is performed using a gold standard based on actual inter-language links. Three teams submitted eleven runs to link text in three languages to comparable English texts. A special section in this proceedings volume reports on this shared task.

Finally, 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 Benjamin K. Tsou for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers, and to the ACL-IJCNLP 2015 workshop chairs and organizers. We also thank LIMSI-CNRS for financial support to our invited speaker. Last but not least we would like to thank our authors and the participants of the workshop.

Pierre Zweigenbaum, Serge Sharoff, Reinhard Rapp

Organizers

Pierre Zweigenbaum LIMSI, CNRS, Orsay (France), Chair
Serge Sharoff University of Leeds (UK), Shared Task Chair
Reinhard Rapp University of Mainz (Germany)

Programme Committee

Ahmet Aker, University of Sheffield (UK)
Srinivas Bangalore (AT&T Labs, US)
Caroline Barrière (CRIM, Montréal, Canada)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Kurt Eberle (Lingenio, Heidelberg, Germany)
Andreas Eisele (European Commission, Luxembourg)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (INRIA, Saclay, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology)
Kyo Kageura (University of Tokyo, Japan)
Adam Kilgarriff (Lexical Computing Ltd, UK)
Natalie Kübler (Université Paris Diderot, France)
Philippe Langlais (Université de Montréal, Canada)
Michael Mohler (Language Computer Corp., US)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Lene Offersgaard (University of Copenhagen, Denmark)
Ted Pedersen (University of Minnesota, Duluth, US)
Reinhard Rapp (Université Aix-Marseille, France)
Sujith Ravi (Google, US)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Tim Van de Cruys (IRIT-CNRS, Toulouse, France)
Stephan Vogel, QCRI (Qatar)
Guillaume Wisniewski (Université Paris Sud & LIMSI-CNRS, Orsay, France)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

Invited Speaker

Benjamin K. Tsou (City University of Hong Kong)

Table of Contents

Workshop Program

Thursday, July 30, 2015

Session 1: 09:00–10:30 Opening Session

09:00–09:05 *Introduction to the BUCC Workshop*
Pierre Zweigenbaum, Serge Sharoff, Reinhard Rapp

09:05–10:05 **Invited presentation:** *Augmented Comparative Corpora and Monitoring Corpus in Chinese: LIVAC and Sketch Search Engine Compared*
Benjamin K. Tsou

10:05–10:30 *A Factory of Comparable Corpora from Wikipedia*
Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba and Lluís Màrquez

Session 2: 11:00–12:30

11:00–11:25 *Knowledge-lean projection of coreference chains across languages*
Yulia Grishina and Manfred Stede

11:25–11:50 *Projective methods for mining missing translations in DBpedia*
Laurent Jakubina and Philippe Langlais

11:50–12:05 *Attempting to Bypass Alignment from Comparable Corpora via Pivot Language*
Alexis Linard, Béatrice Daille and Emmanuel Morin

12:05–12:20 *Application of a Corpus to Identify Gaps between English Learners and Native Speakers*
Katsunori Kotani and Takehiko Yoshimi

Thursday, July 30, 2015 (continued)

Session 3: 14:00–15:30 Alignment

14:00–14:25 *A Generative Model for Extracting Parallel Fragments from Comparable Documents*
Somayeh Bakhshaei, Shahram Khadivi and Reza Safabakhsh

14:25–14:50 *Evaluating Features for Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families*
Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto

14:50–15:05 *Extracting Bilingual Lexica from Comparable Corpora Using Self-Organizing Maps*
Hyeong-Won Seo, Minah Cheon and Jae-Hoon Kim

15:05–15:20 *Obtaining SMT dictionaries for related languages*
Miguel Rios and Serge Sharoff

Session 4: 16:00–17:00 Shared Task

16:00–16:15 *BUCC Shared Task: Cross-Language Document Similarity*
Serge Sharoff, Pierre Zweigenbaum and Reinhard Rapp

16:15–16:30 *AUT Document Alignment Framework for BUCC Workshop Shared Task*
Atefeh Zafarian, Amir Pouya Agha Sadeghi, Fatemeh Azadi, Sonia Ghiasifard, Zeinab Ali Panahloo, Somayeh Bakhshaei and Seyyed Mohammad Mohammadzadeh Ziabary

16:30–16:45 *LINA: Identifying Comparable Documents from Wikipedia*
Emmanuel Morin, Amir Hazem, Florian Boudin and Elizaveta Loginova-Clouet

16:45–17:00 *Shared Task: General Discussion*

Closing: 17:00