ACL 2017

Tenth Workshop on Building and Using Comparable Corpora

Proceedings of the Workshop

August 3, 2017
Vancouver, Canada

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 978-1-945626-61-6

Introduction

In the language engineering and the linguistics communities, research on comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical Natural Language Processing applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

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 nine previous editions of the workshop which took place in Africa (LREC’08 in Marrakech), North America (ACL’11 in Portland), Asia (ACL-IJCNLP’09 in Singapore and ACL-IJCNLP’15 in Beijing), Europe (LREC’10 in Malta, ACL’13 in Sofia, and LREC’14 in Reykjavik) and also on the border between Asia and Europe (LREC’12 in Istanbul), the workshop this year has returned to North America, first time in Canada in Vancouver.

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 Philippe Langlais for accepting to give the keynote talk, 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’17 workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

This year the workshop included a shared task to quantitatively evaluate competing methods for extracting parallel sentences from comparable monolingual corpora, so as to give an overview on the state of the art and to identify the best performing approaches. 13 runs were submitted in time to the shared task by 4 teams, covering three of the four proposed language pairs: French-English (7 runs), German-English (3 runs), and Chinese-English (3 runs). We make the datasets are available on the workshop Web page at `https://comparable.limsi.fr/bucc2017/bucc2017-task.html`.

Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp — August 2017

Organizers:

Serge Sharoff	University of Leeds, UK
Pierre Zweigenbaum	LIMSI, CNRS, Université Paris-Saclay, Orsay, France
Reinhard Rapp	Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany

Program Committee:

Ahmet Aker	University of Sheffield, UK
Hervé Déjean	Xerox Research Centre Europe, Grenoble, France
Kurt Eberle	Lingenio, Germany
Andreas Eisele	European Commission, Luxembourg
Éric Gaussier	Université Joseph Fourier, Grenoble, France
Vishal Goyal	Punjabi University, Patiala, India
Silvia Hansen-Schirra	University of Mainz, Germany
Hitoshi Isahara	Toyohashi University of Technology
Kyo Kageura	University of Tokyo, Japan
Philippe Langlais	Université de Montrèal, Canada
Shervin Malmasi	Harvard Medical School, Boston, MA, USA
Michael Mohler	Language Computer Corp., US
Emmanuel Morin	Université de Nantes, France
Dragos Stefan Munteanu	Language Weaver, Inc., US
Ted Pedersen	University of Minnesota, Duluth, US
Reinhard Rapp	Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany
Serge Sharoff	University of Leeds, UK
Michel Simard	National Research Council Canada
Pierre Zweigenbaum	LIMSI-CNRS, Orsay, France

Invited Speaker:

Philippe Langlais, Université de Montrèal, Canada

Table of Contents

Workshop Program

9:00 - 9:05	**Opening**
9:05 - 10:00	**Invited presentation** *Users and Data: The Two Neglected Children of Bilingual Natural Language Processing Research* Phillippe Langlais
Session 1:	**Plagiarism detection**
10:00 - 10:30	*Deep Investigation of Cross-Language Plagiarism Detection Methods* Jérémy Ferrero, Laurent Besacier, Didier Schwab and Frédéric Agnès
10:30 - 11:00	**Coffee break**
Session 2:	**Sentence alignment and lexicon acquisition**
11:00 - 11:30	*Sentence Alignment using Unfolding Recursive Autoencoders* Jeenu Grover and Pabitra Mitra
11:30 - 12:00	*Acquisition of Translation Lexicons for Historically Unwritten Languages via Bridging Loanwords* Michael Bloodgood and Benjamin Strauss
12:00 - 14:00	**Lunch**
Session 3:	**Building comparable corpora**
14:00 - 14:30	*Toward a Comparable Corpus of Latvian, Russian and English Tweets* Dmitrijs Milajevs
14:30 - 15:00	*Automatic Extraction of Parallel Speech Corpora from Dubbed Movies* Alp Öktem, Mireia Farrús and Leo Wanner
15:00 - 15:30	*A parallel collection of clinical trials in Portuguese and English* Mariana Neves
15:30 - 16:00	**Coffee break**
Session 4:	**Shared task session**
16:00 - 16:20	*Overview on the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora* Pierre Zweigenbaum, Serge Sharoff, Reinhard Rapp
16:20 - 16:40	*Weighted Set-Theoretic Alignment of Comparable Sentences* Andoni Azpeitia, Thierry Etchegoyhen and Eva Martínez Garcia
16:40 - 17:00	*BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora* Francis Grégoire and Philippe Langlais
17:00 - 17:20	*zNLP: Identifying Parallel Sentences in Chinese-English Comparable Corpora* Zheng Zhang and Pierre Zweigenbaum
17:20 - 17:40	*BUCC2017: A Hybrid Approach for Identifying Parallel Sentences in Comparable Corpora* Sainik Mahata, Dipankar Das and Sivaji Bandyopadhyay
17:40 - 17:50	**Closing**