LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

13th Workshop on Building and Using Comparable Corpora

PROCEEDINGS

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff (eds.)

Proceedings of the LREC 2020
13th Workshop on Building and Using Comparable Corpora

Edited by: Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

ISBN: 979-10-95546-42-9
EAN: 9791095546429

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Preface – 13th BUCC at LREC 2020

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 primarily 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 twelve previous editions of the workshop which took place in Africa (LREC’08 in Marrakech), America (ACL’11 in Portland and ACL’17 in Vancouver), Asia (ACL-IJCNLP’09 in Singapore, ACL-IJCNLP’15 in Beijing, LREC’18 in Miyazaki, Japan), Europe (LREC’10 in Malta, ACL’13 in Sofia, LREC’14 in Reykjavik, LREC’16 in Portoroz, RANLP’19 in Varna) and also on the border between Asia and Europe (LREC’12 in Istanbul), this year the 13th edition of the BUCC workshop was supposed to be held in Marseille.

However, due to the corona crisis, unfortunately LREC 2020 could not be held in Marseille this year. Therefore, with full support of the LREC organizers, we decided to hold the BUCC workshop as a free online event on the planned date. This not only causes problems, but also offers chances which we are eager to explore. Fortunately, the fourth BUCC shared task on "Bilingual Dictionary Induction from Comparable Corpora" was not strongly affected by this change and could be successfully conducted with surprisingly good results. Several papers by the shared task participants in this volume as well as an overview paper provide more information on this.

We would like to thank all people who in one way or another helped in making this workshop once again a success. We are especially grateful to Khalid Choukri for his excellent and almost magical guidance concerning the proceedings, to Nicoletta Calzolari for her continuous support of our workshop, and to Hélène Mazo, Sara Goggi and the whole team of LREC organisers for finding solutions to all matters of concern.

Our special thanks go to Holger Schwenk and Jörg Tiedemann for accepting to give invited presentations and to the members of the programme committee who did an excellent job in reviewing the submitted papers under strict time constraints. Last but not least we would like to thank our authors, shared task teams and all participants of the workshop.

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff May 2020

Workshop Organizers:

Reinhard Rapp, Athena R.C., Magdeburg-Stendal University of Applied Sciences, University of Mainz (Chair)
Pierre Zweigenbaum, LIMSI, CNRS, Université Paris-Saclay
Serge Sharoff, University of Leeds

Programme Committee:

Ahmet Aker (University of Sheffield, UK)
Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
Hervé Déjean (Naver Labs Europe, Grenoble, France)
Thierry Etchegoyhen (VicomTech, Spain)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology, Japan)
Kyo Kageura (The University of Tokyo, Japan)
Yves Lepage (Waseda University, Japan)
Shervin Malmasi (Harvard Medical School, Boston, MA, USA)
Michael Mohler (Language Computer Corp., USA)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., USA)
Ted Pedersen (University of Minnesota, Duluth, US)
Reinhard Rapp (Athena R.C., Magdeburg-Stendal University of Applied Sciences, University of Mainz)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Richard Sproat (OGI School of Science & Technology, USA)
Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LIMSI, Orsay, France)

Invited Speakers:

Holger Schwenk, Facebook Artificial Intelligence Research
Jörg Tiedemann, University of Helsinki

Table of Contents

BUCC 2020 Workshop Programme

Monday, May 11, 2020

Times refer to Central European Summer Time (UTC + 2)
https://www.timeanddate.com/worldclock/france/marseille

09:15–9:30 ***Opening***

Session 1: Invited Presentation

09:30–10:20 Holger Schwenk, Facebook AI Research

Session 2: Shared Task: Bilingual Dictionary Induction from Comparable Corpora

10:20–10:40 *Overview of the Fourth BUCC Shared Task: Bilingual Dictionary Induction from Comparable Corpora*
Reinhard Rapp, Pierre Zweigenbaum and Serge Sharoff

10:40–11:00 *TALN/LS2N Participation at the BUCC Shared Task: Bilingual Dictionary Induction from Comparable Corpora*
Martin Laville, Amir Hazem and Emmanuel Morin

11:00–11:20 ***Coffee Break***

11:20–11:40 *LMU Bilingual Dictionary Induction System with Word Surface Similarity Scores for BUCC 2020*
Silvia Severini, Viktor Hangya, Alexander Fraser and Hinrich Schütze

11:40–12:00 *BUCC2020: Bilingual Dictionary Induction using Cross-lingual Embedding*
Sanjanasri JP, Vijay Krishna Menon and Soman KP

12:00–13:00 ***Lunch Break***

Session 3: Invited Presentation

13:00–13:50 Jörg Tiedemann, University of Helsinki

Session 4: Corpus Construction

13:50–14:10 *Constructing a Bilingual Corpus of Parallel Tweets*
Hamdy Mubarak, Sabit Hassan and Ahmed Abdelali

14:10–14:30 *cEnTam: Creation and Validation of a New English-Tamil Bilingual Corpus*
Sanjanasri JP, Premjith B, Vijay Krishna Menon and Soman KP

14:30–14:50 ***Coffee Break***

Session 5: Semantics

14:50–15:10 *Automatic Creation of Correspondence Table of Meaning Tags from Two Dictionaries in One Language Using Bilingual Word Embedding*
Teruo Hirabayashi, Kanako Komiya, Masayuki Asahara and Hiroyuki Shinnou

15:10–15:30 *Mining Semantic Relations from Comparable Corpora through Intersections of Word Embeddings*
Špela Vintar, Larisa Grčić Simeunović, Matej Martinc, Senja Pollak and Uroš Stepišnik

Session 6: Machine Translation

15:30–15:50 *Benchmarking Multidomain English-Indonesian Machine Translation*
Tri Wahyu Guntara, Alham Fikri Aji and Radityo Eko Prasojo

15:50–16:10 *Reducing the Search Space for Parallel Sentences in Comparable Corpora*
Rémi Cardon and Natalia Grabar

16:10–16:30 *Line-a-line: A Tool for Annotating Word-Alignments*
Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson and Rickard Domeij

16:30–16:35 ***Closing***