BUCC 2021

Proceedings
of the
14th Workshop on Building and Using Comparable Corpora (BUCC 2021)

in conjunction with

International Conference on Recent Advances in Natural Language Processing (RANLP 2021)

Edited by

Reinhard Rapp, Serge Sharoff, Pierre Zweigenbaum

September 6, 2021

14th Workshop on Building and Using Comparable Corpora (BUCC 2021)
in conjuction with
the International Conference Recent Advances in Natural Language Processing (RANLP 2021)

PROCEEDINGS

6 September 2021

ISBN 978-954-452-076-2

Designed by INCOMA Ltd.
Shoumen, BULGARIA

14th BUCC Workshop at RANLP 2021 – Preface

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 field by bundling some of its research, thereby making it more visible and giving it a better platform.

The first 12 editions of the workshop 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). Due to the corona crisis, the 13th edition took place online as an LREC'20 workshop. This year's 14th edition was held again online and took place as an RANLP'21 workshop.

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 Ruslan Mitkov, Galia Angelova, Ivelina Nikolova, Kiril Simov and the whole RANLP team for their excellent support.

Our special thanks go to Pushpak Bhattacharyya, Tomas Mikolov and Sujith Ravi 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, presenters and all participants of the workshop.

Reinhard Rapp, Serge Sharoff, Pierre Zweigenbaum September 2021

Workshop Organizers:

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

Programme Committee:

Ahmet Aker (University of Sheffield, UK)
Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
Thierry Etchegoyhen (VicomTech, Spain)
Hitoshi Isahara (Otemon Gakuin University, Japan)
Kyo Kageura (The University of Tokyo, Japan)
Natalie Kübler (CLILLAC-ARP, Université de Paris, France
Philippe Langlais (Université de Montréal, Canada
Yves Lepage (Waseda University, Japan)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., USA)
Reinhard Rapp (Athena R.C., Magdeburg-Stendal University of Applied Sciences, University of Mainz)
Nasredine Semmar (CEA LIST, Paris, France)
Serge Sharoff (University of Leeds, UK)
Richard Sproat (OGI School of Science & Technology, USA)
Tim Van de Cruys (KU Leuven, Belgium)
Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Invited Speakers:

Pushpak Bhattacharyya, Indian Institute of Technology Bombai
Tomas Mikolov, Czech Institute of Informatics, Robotics and Cybernetics
Sujith Ravi, SliceX AI

Table of Contents

BUCC 2021 Workshop Programme

Monday, September 6, 2021

Times refer to UTC + 0

08:00–8:05 ***Opening***

Session 1: Invited Presentation

08:05–9:00 *Machine Translation in Low Resource Setting*
Pushpak Bhattacharyya, IIT Bombay

Session 2: Corpus Construction

9:00–9:25 *EM Corpus: a comparable corpus for a less-resourced language pair Manipuri-English*
Rudali Huidrom, Yves Lepage and Khogendra Khomdram

9:25–9:40 ***Coffee Break***

Session 3: Data Extraction and Corpus Annotation

9:40–10:05 *Mining Bilingual Word Pairs from Comparable Corpus using Apache Spark Framework*
Sanjanasri JP, Vijay Krishna Menon, Soman KP and Krzysztof Wolk

10:05–10:30 *Employing Wikipedia as a resource for Named Entity Recognition in Morphologically complex under-resourced languages*
Aravind Krishnan, Stefan Ziehe, Franziska Pannach and Caroline Sporleder

10:30–10:55 *Semi-Automated Labeling of Requirement Datasets for Relation Extraction*
Jeremias Bohn, Jannik Fischbach, Martin Schmitt, Hinrich Schütze and Andreas Vogelsang

10:55–11:20 *A Dutch Dataset for Cross-lingual Multilabel Toxicity Detection*
Ben Burtenshaw and Mike Kestemont

11:20–12:10 ***Lunch Break***

Session 4: Invited Presentation

12:10–13:05 *Language Modeling and AI*
Tomas Mikolov, Czech Institute of Informatics, Robotics and Cybernetics

Session 5: Neural MT and Bitext Extraction

13:05–13:30 *Syntax-aware Transformers for Neural Machine Translation: The Case of Text to Sign Gloss Translation*
Santiago Egea Gómez, Euan McGill and Horacio Saggion

13:30–13:55 *Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches*
Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson and Andy Way

13:55–14:10 ***Coffee Break***

Session 6: Bitext Retrieval and Dictionary Extraction

14:10–14:35 *Majority Voting with Bidirectional Pre-translation For Bitext Retrieval*
Alexander Jones and Derry Tanti Wijaya

14:35–15:00 *On Pronunciations in Wiktionary: Extraction and Experiments on Multilingual Syllabification and Stress Prediction*
Winston Wu and David Yarowsky

Session 7: Invited Presentation

15:00–15:55 *Large-scale Deep Learning for Low-Resource AI*
Sujith Ravi, SliceX AI

15:55–16:00 ***Closing***