LREC-COLING 2024

The 17th Workshop on
Building and Using Comparable Corpora (BUCC)
@ LREC-COLING-2024

Workshop Proceedings

Editors
Pierre Zweigenbaum, Reinhard Rapp and Serge Sharoff

20 May, 2024
Torino, Italy

Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024

ISBN 978-2-493814-31-9
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface

The 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024

In the language engineering and linguistics communities, research in 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 NLP applications such as statistical and neural machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest because they enable cross-language discoveries and comparisons. It is generally accepted in both communities that comparable corpora consist of documents that are comparable in content and form in various degrees and dimensions across several languages or language varieties. Parallel corpora are on the one end of this spectrum, unrelated corpora on the other.

Comparable corpora have been used in various applications, including Information Retrieval, Machine Translation, Cross-lingual text classification, etc. The linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for statistical natural language processing applications, for example, to extract parallel corpora from comparable corpora for neural machine translation. As such, it is of great interest to bring together builders and users of such corpora. The aim of the workshop series on "Building and Using Comparable Corpora" (BUCC) is to promote progress in this field.

The previous editions of the workshop took place in Africa (LREC 2008 in Marrakech), America (ACL 2011 in Portland and ACL 2017 in Vancouver), Asia (ACL-IJCNLP 2009 in Singapore, ACL-IJCNLP 2015 in Beijing, LREC 2018 in Miyazaki, Japan), Europe (LREC 2010 in Malta, ACL 2013 in Sofia, LREC 2014 in Reykjavik, LREC 2016 in Portoroz, RANLP 2019 and RANLP 2023 in Varna, LREC 2022 in Marseille) and also on the border between Asia and Europe (LREC 2012 in Istanbul). Due to the Corona crisis, the workshop was also held online in conjunction with LREC 2020 and RANLP 2021. The materials of the past workshops and related studies have also been summarised in a recent textbook from Springer:
`https://link.springer.com/book/10.1007/978-3-031-31384-4`.

We want to thank all the people who, in one way or another, helped make this workshop once again a success, especially the LREC management chairs, workshop chairs, and publication chairs.

Our special thanks go to our invited speaker, François Yvon, and to the members of the program committee, who did a great job in reviewing the submitted papers under strict time constraints. Last but not least, we would like to thank the authors and all workshop participants.

Pierre Zweigenbaum, Reinhard Rapp, Serge Sharoff May 2024

Workshop Chairs

Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Reinhard Rapp (University of Mainz and Magdeburg-Stendal University of Applied Sciences, Germany)

Serge Sharoff (University of Leeds, United Kingdom)

Program Committee

- Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
- Thierry Etchegoyhen (Vicomtech, Spain)
- Kyo Kageura (University of Tokyo, Japan)
- Natalie Kübler (Université Paris Cité, France)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Shervin Malmasi (Amazon, USA)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Nantes Université, France)
- Dragos Stefan Munteanu (Language Weaver, Inc., USA)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Ayla Rigouts Terryn (KU Leuven, Belgium)
- Reinhard Rapp (University of Mainz and Magdeburg-Stendal University of Applied Sciences, Germany)
- Nasredine Semmar (CEA LIST, Paris, France)
- Silvia Severini (Leonardo Labs, Italy)
- 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)

Acknowledgements

The BUCC workshop was supported by ANR-20-IADJ-0005-01 under the trilateral ANR-DFG-JST AI Research project KEEPHA, *Knowledge-enhanced information extraction across languages for pharmacovigilance*, through support to the participation of Pierre Zweigenbaum and François Yvon.

Table of Contents

Posters

Workshop Program

Monday, 20 May, 2024

9:00–10:30 **Session 1**

9:00–9:30 *On a Novel Application of Wasserstein-Procrustes for Unsupervised Cross-Lingual Alignment of Embeddings*
Guillem Ramírez, Rumen Dangovski, Preslav Nakov and Marin Soljacic

9:30–10:00 *Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal*
Julius Steuer, Marie-Pauline Krielke, Stefan Fischer, Stefania Degaetano-Ortlieb, Marius Mosbach and Dietrich Klakow

10:00–10:30 *PORTULAN ExtraGLUE Datasets and Models: Kick-starting a Benchmark for the Neural Processing of Portuguese*
Tomás Freitas Osório, Bernardo Leite, Henrique Lopes Cardoso, Luís Gomes, João Rodrigues, Rodrigo Santos and António Branco

10:30–11:00 **Coffee break**

11:00–12:00 **Invited talk**

11:00–12:00 *The Way Towards Massively Multilingual Language Models*
François Yvon

12:00–13:00 **Session 2**

12:00–12:30 *Quality and Quantity of Machine Translation References for Automatic Metrics*
Vilém Zouhar and Ondřej Bojar

12:30–13:00 *Exploring the Necessity of Visual Modality in Multimodal Machine Translation using Authentic Datasets*
Zi Long, ZhenHao Tang, Xianghua Fu, Jian Chen, Shilong Hou and Jinze Lyu

13:00–14:00 **Lunch break**

14:00–16:00 **Session 3**

14:00–14:30 *Exploring the Potential of Large Language Models in Adaptive Machine Translation for Generic Text and Subtitles*
Abdelhadi Soudi, Mohamed Hannani, Kristof Van Laerhoven and Eleftherios Avramidis

14:30–15:00 *INCLURE: a Dataset and Toolkit for Inclusive French Translation*
Paul Lerner and Cyril Grouin

15:00–15:30 *BnPC: A Gold Standard Corpus for Paraphrase Detection in Bangla, and its Evaluation*
Sourav Saha, Zeshan Ahmed Nobin, Mufassir Ahmad Chowdhury, Md. Shakirul Hasan Khan Mobin, Mohammad Ruhul Amin and Sudipta Kar

Monday, 20 May, 2024 (continued)

15:30–16:00 ***Booster presentations***
poster authors

16:00–16:30 Coffee break

16:30–18:00 Poster session

Creating Clustered Comparable Corpora from Wikipedia with Different Fuzziness Levels and Language Representativity
Anna Laskina, Eric Gaussier and Gaelle Calvary

EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research
Marc Kupietz, Piotr Banski, Nils Diewald, Beata Trawinski and Andreas Witt

Building Annotated Parallel Corpora Using the ATIS Dataset: Two UD-style treebanks in English and Turkish
Neslihan Cesur, Aslı Kuzgun, Mehmet Kose and Olcay Taner Yıldız

Bootstrapping the Annotation of UD Learner Treebanks
Arianna Masciolini

SweDiagnostics: A Diagnostics Natural Language Inference Dataset for Swedish
Felix Morger

Multiple Discourse Relations in English TED Talks and Their Translation into Lithuanian, Portuguese and Turkish
Deniz Zeyrek, Giedrė Valūnaitė Oleškevičienė and Amalia Mendes

mini-CIEP+ : A Shareable Parallel Corpus of Prose
Annemarie Verkerk and Luigi Talamo