RANLP 2019

12th Workshop on Building and Using Comparable Corpora

PROCEEDINGS

Edited by

Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp

5 September 2019

Proceedings of the 12th Workshop on
Building and Using Comparable Corpora, 8 September 2019 – RANLP 2019, Varna, Bulgaria

Edited by Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp

https://comparable.limsi.fr/bucc2019/

Organising Committee

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

Programme Committee

- Ahmet Aker (University of Sheffield, UK)
- Ebrahim Ansari (Department of Computer Science and Information Technology, IASBS, Iran)
- Thierry Etchegoyhen (Vicomtech, Spain)
- Gregory Grefenstette (INRIA, Saclay, France)
- Askar Hamdulla (Xinjiang University, China)
- Hitoshi Isahara (Toyohashi University of Technology)
- Kyo Kageura (University of Tokyo, Japan)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Shervin Malmasi (Harvard Medical School, US)
- Pabitra Mitra (Indian Institute of Technology, Kharagpur, India)
- Michael Mohler (Language Computer Corp, US)
- Emmanuel Morin (Université de Nantes, France)
- Dragos Stefan Munteanu (SDL Research, Los Angeles, US)
- Lene Offersgaard (University of Copenhagen, Denmark)
- Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)
- Serge Sharoff (University of Leeds, UK)
- Nasredine Semmar (CEA LIST, France)
- Michel Simard (National Research Council, Canada)
- Richard Sproat (OGI School of Science Technology, US)
- 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, Université Paris-Saclay, Orsay, France)

Preface – 12th BUCC at RANLP'19

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 eleven 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 and LREC'16 in Portoroz) and also on the border between Asia and Europe (LREC'12 in Istanbul), this year the 12th edition of the BUCC workshop is back to Bulgaria (the first time the BUCC workshop returns to a country).

A major paradigm change in the field concerns the prevalence of Artificial Neural Networks, also appearing under the more catchy title of *Deep Learning*. Within the last five years, the Deep Learning methods shifted the balance in multilingual NLP processing towards less parallel and more comparable resources, e.g., by providing multilingual embedding spaces from monolingual corpora and by enabling Neural MT with minimal or no reliance on parallel data. Neural Networks finally make it possible to take long distance dependencies (e.g. between the words within a sentence) into account, thus overcoming a fundamental limitation of traditional n-gram-based approaches. The proceedings of this workshop present the new horizons for multilingual research with limited resources.

We would like to thank all people who in one way or another helped in making this workshop once again a success. We're especially grateful to Ruslan Mitkov and the team of the RANLP organisers for helping us with the event.

Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp September 2019

Programme

9:20–9:30 ***Opening Remarks***

Session 1: Panel

09:30–10:30 Invited panel: Ekaterina Lapshinova-Koltunski, Sebastian Pado, Alexander Panchenko
Future directions of research in comparable corpora

Session 2: Applications of Neural Networks to Comparable Corpora

10:30–11:00 Yuri Bizzoni, Elke Teich
Analyzing variation in translation through neural semantic spaces

11:00–11:30 Maria Kunilovskaya, Shiva Taslimipoor, Tatyana Ilyushchenya
Functional Text Representations for Building Cross-Linguistic Comparable Corpora in English and Russian

11:30–12:00 ***Coffee Break***

Session 3: Inducing Lexicons from Comparable Corpora

12:00–12:30 Ebrahim Ansari, M.H. Sadreddini, Mahsa Radinmehr and Ziba Khosravan
Extracting Bilingual Persian Italian Lexicon from Comparable Corpora Using Different Seed Dictionaries

12:30–13:00 Kim Steyaert and Ayla Rigouts Terryn
Multilingual Term Extraction from Comparable Corpora: Informativeness of Monolingual Term Extraction Features

13:00–14:30 ***Lunch Break***

Session 4: Building Comparable corpora

14:30–15:00 Benjamin K. Tsou and Kapo Chow
From the cultivation of comparable corpora to harvesting from them: A quantitative and qualitative exploration

Session 5: Shared Tasks

15:00–16:00 Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff
Towards the Fourth BUCC Shared Task: inducing bilingual lexicons through comparable corpora

16:00–16:15 *Closing remarks*

Table of Contents