BUCC, 14th Workshop on Building and Using Comparable Corpora

Special Topic: Neural Networks in Comparable Corpora Research
Co-located with RANLP 2021, online
6 September 2021
Website: https://comparable.limsi.fr/bucc2021/
Invited speakers:
Pushpak Bhattacharyya, Indian Institute of Technology Bombay
Tomáš Mikolov, Czech Institute of Informatics, Robotics and Cybernetics
Sujith Ravi, SliceX AI
Program, connection, proceedings
New: Workshop proceedings (full PDF; full list of BibTeX entries)

INVITED SPEAKERS

Pushpak Bhattacharyya

Indian Institute of Technology
Mumbai

Machine Translation in Low Resource Setting

Abstract

AI now and in future will have to grapple continuously with the problem of low resource. AI will increasingly be ML intensive. But ML needs data often with annotation. However, annotation is costly.

Over the years, through work on multiple problems, we have developed insight into how to do language processing in low resource setting. Following 6 methods—individually and in combination—seem to be the way forward:

1. Artificially augment resource (e.g. subwords)
2. Cooperative NLP (e.g., pivot in MT)
3. Linguistic embellishment (e.g. factor based MT, source reordering)
4. Joint Modeling (e.g., Coref and NER, Sentiment and Emotion: each task helping the other to either boost accuracy or reduce resource requirement)
5. Multimodality (e.g., eye tracking based NLP, also picture+text+speech based Sentiment Analysis)
6. Cross Lingual Embedding (e.g., embedding from multiple languages helping MT, close to 2 above)

The present talk will focus on low resource machine translation. We describe the use of techniques from the above list and bring home the seriousness and methodology of doing Machine Translation in low resource settings.

Bio

Dr. Pushpak Bhattacharyya is Professor of Computer Science and Engineering Department IIT Bombay. His research areas are Natural Language Processing, Machine Learning and AI (NLP-ML-AI). Prof. Bhattacharyya has published more than 350 research papers in various areas of NLP. His textbook ‘Machine Translation’ sheds light on all paradigms of machine translation with abundant examples from Indian Languages. Two recent monographs co-authored by him called ‘Investigations in Computational Sarcasm’ and ‘Cognitively Inspired Natural Language Processing—An Investigation Based on Eye Tracking’ describe cutting edge research in NLP and ML. Prof. Bhattacharyya is Fellow of Indian National Academy of Engineering (FNAE) and Abdul Kalam National Fellow. For sustained contribution to technology he received the Manthan Award of the Ministry of IT, P.K. Patwardhan Award of IIT Bombay and VNMM Award of IIT Roorkey. He is also a Distinguished Alumnus of IIT Kharagpur.

Tomáš Mikolov

Czech Institute of Informatics, Robotics and Cybernetics

Language modeling and AI

Bio

Tomas Mikolov is a researcher at CIIRC, Prague. Currently he leads a research team focusing on development of novel techniques within the area of complex systems, artificial life and evolution. Previously, he did work at Facebook AI and Google Brain, where he led development of popular machine learning tools such as word2vec and fastText. He obtained PhD at the Brno University of Technology in 2012 for his work on neural language models (the RNNLM project). His main research interest is to understand intelligence, and to create artificial intelligence that can help people to solve complex problems.

Sujith Ravi

SliceX AI

Large-scale Deep Learning for Low-Resource AI

Bio

Dr. Sujith Ravi is Founder and CEO at SliceX AI. Previously, he was the Director of Amazon Alexa AI where he led efforts to build the future of multimodal conversational AI experiences at scale. Prior to that, he was leading and managing multiple ML and NLP teams and efforts in Google AI. He founded and headed Google’s large-scale graph-based semi-supervised learning platform, deep learning platform for structured and unstructured data as well as on-device machine learning efforts for products used by billions of people in Search, Ads, Assistant, Gmail, Photos, Android, Cloud and YouTube. These technologies power conversational AI (e.g., Smart Reply), Web and Image Search; On-Device predictions in Android and Assistant; and ML platforms like Neural Structured Learning in TensorFlow, Learn2Compress as Google Cloud service, TensorFlow Lite for edge devices.

Dr. Ravi has authored over 100 scientific publications and patents in top-tier machine learning and natural language processing conferences. His work has been featured in press: Wired, Forbes, Forrester, New York Times, TechCrunch, VentureBeat, Engadget, New Scientist, among others, and also won the SIGDIAL Best Paper Award in 2019 and ACM SIGKDD Best Research Paper Award in 2014. For multiple years, he was a mentor for Google Launchpad startups. Dr. Ravi was the Co-Chair (AI and deep learning) for the 2019 National Academy of Engineering (NAE) Frontiers of Engineering symposium. He was also the Co-Chair for ACL 2021, EMNLP 2020, ICML 2019, NAACL 2019, and NeurIPS 2018 ML workshops and regularly serves as Senior/Area Chair and PC of top-tier machine learning and natural language processing conferences like NeurIPS, ICML, ACL, NAACL, AAAI, EMNLP, COLING, KDD, and WSDM.

MOTIVATION

In the language engineering and the 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. Parallel corpora are on the one end of this spectrum, unrelated corpora on the other.

Comparable corpora have been used in a range of 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 applications of statistical NLP, for example to extract parallel corpora from comparable corpora for neural MT. As such, it is of great interest to bring together builders and users of such corpora.

TOPICS

This year our special topic is "Neural Networks in Comparable Corpora Research". But we solicit contributions on all topics related to comparable (and parallel) corpora, including but not limited to the following:

Building Comparable Corpora:

- Automatic and semi-automatic methods
- Methods to mine parallel and non-parallel corpora from the web
- Tools and criteria to evaluate the comparability of corpora
- Parallel vs non-parallel corpora, monolingual corpora
- Rare and minority languages, across language families
- Multi-media/multi-modal comparable corpora

Applications of comparable corpora:

- Human translation
- Language learning
- Cross-language information retrieval & document categorization
- Bilingual and multilingual projections
- Machine translation
- Writing assistance
- Machine learning techniques using comparable corpora

Mining from Comparable Corpora:

- Cross-language distributional semantics, word embeddings and pre-trained multilingual transformer models
- Extraction of parallel segments or paraphrases from comparable corpora
- Methods to derive parallel from non-parallel corpora (e.g. to provide for low-resource languages in neural machine translation)

- Extraction of bilingual and multilingual translations of single words and multi-word expressions, proper names, and named entities from comparable corpora
- Induction of morphological, grammatical, and translation rules from comparable corpora
- Induction of multilingual word classes from comparable corpora

PRACTICAL INFORMATION

The workshop proceedings (full PDF; full list of BibTeX entries) are published in the ACL Anthology. See the Program page for information about connection.

Workshop fees are 45 Euros for presenters and 15 Euros for non-presenters. For further details see https://ranlp.org/ranlp2021/fees.php

IMPORTANT DATES

extended to July 17, 2021	Paper submission deadline
July 31, 2021	Notification to authors
July 31, 2021	Early bird registration (reduced rates)
Aug 31, 2021	Camera-ready final papers
Sep 6, 2021	Workshop date

SUBMISSION INFORMATION

Please follow the style sheet and templates provided for the main conference at https://www.softconf.com/ranlp2021/BU
Papers should be submitted as a PDF file at https://www.softconf.com/ranlp2021/BUCC2021/. Submissions must describe original and unpublished work and range from four (4) to eight (8) pages plus unlimited references.

Reviewing will be double blind, so the papers should not reveal the authors' identity. Accepted papers will be published in the workshop proceedings (full PDF; full list of BibTeX entries).

Double submission policy: Parallel submission to other meetings or publications is possible but must be immediately notified to the workshop organizers.

In case of questions, please contact Reinhard Rapp: <reinhardrapp (at) gmx (dot) de>

Plain-text CFP : bucc2021-cfp.txt
PDF CFP : bucc2021-cfp.pdf
Last modified: 21 Jan 2022

DETAILS FOR ATTENDANCE

The workshop will take place online through Zoom. Details for participation are provided on the main conference Web site. This is the Zoom link to connect to the workshop. In the unlikely case of unforeseen problems with the Zoom session, a new link will be provided here.

The workshop proceedings (full PDF; full list of BibTeX entries) are available on the ACL Anthology. See also below direct links for each individual paper to ACL anthology page, PDF and BibTeX entry.

Last modified: 21 Jan 2022

WORKSHOP ORGANIZERS

Reinhard Rapp (Athena R.C.; Magdeburg-Stendal University of Applied Sciences; University of Mainz, Germany), Chair and contact person: reinhardrapp (at) gmx (dot) de

Serge Sharoff (University of Leeds, United Kingdom)

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

PROGRAMME COMMITTEE

- Ahmet Aker (University of Duisburg-Essen, Germany)
- Ebrahim Ansari (Institue 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 (Univerité de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Emmanuel Morin (Université de Nantes, France)
- Dragos Stefan Munteanu (RWS, USA)
- Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)
- 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)

Last modified: 3 Sep 2021