The 17th Workshop on Building and Using Comparable Corpora (BUCC)

Co-located with LREC-COLING 2024,
Torino, Italia, 20 May 2024
Invited speaker: François Yvon, Sorbonne Université, CNRS, ISIR

INVITED SPEAKER

François Yvon

Sorbonne Université, CNRS, ISIR

The way towards massively multilingual language models

Abstract

In this talk, I will discuss the training and evaluation of massively multilingual language models, that can handle dozens or even hundreds of languages. After motivating the development of such models, I will draw some lessons learned in the course of developing Glot500, a language model covering 500 languages, and some associated resources such as language identification softwares. I will also focus on the challenges raised by "low resourced" languages, i.e. languages for which (a) the available training data is often incomplete, highly specialised and also possibly very noisy; (b) the evaluation data are non existent, requiring to use innovative evaluation strategy, eg. based on various cross-lingual alignment tasks.

Bio

François Yvon is a senior CNRS researcher at the ISIR laboratory of Sorbonne-Université in Paris, France, working on Machine Translation and Multilingual Language Models. Before this, F. Yvon has been leading activities in Machine Translation at LISN / LIMSI in Orsay for about 15 years, resulting in more than one hundred scientific publications on all aspects related to the development and evaluation of multilingual language processing technologies, from word and sentence alignment to translation modelling and evaluation, including recent work on multi-domain adaptation in Machine Translation and on cross-lingual transfert learning issues. He has acted as coordinator or Principal Investigator in multiple past national and international projects in Machine Translation and has supervised more than 20 PhDs on related topics. Between 2013 and 2020, Dr. Yvon has also been the general director of the LIMSI laboratory in Orsay. He is a board member of the European chapter of the Association for Computational Linguistics, of the MetaNet network, and has recently contributed as an expert on linguistic technologies for the French language to several European projects (European Language Resource Collection, ELE - European Language Equality, ELG - European Language Grid).

MOTIVATION

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. 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

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
- (Unsupervised) 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, multi-word expressions, proper names, named entities, sentences, and paraphrases from comparable corpora, etc.
- Induction of morphological, grammatical, and translation rules from comparable corpora
- Induction of multilingual word classes from comparable corpora

Comparable Corpora in the Humanities:

- Comparing linguistic phenomena across languages in contrastive linguistics
- Analyzing properties of translated language in translation studies
- Studying language change over time in diachronic linguistics
- Assigning texts to authors via authors' corpora in forensic linguistics
- Comparing rhetorical features in discourse analysis
- Studying cultural differences in sociolinguistics
- Analyzing language universals in typological research

PRACTICAL INFORMATION

Workshop registration is via the main conference registration site.

The workshop proceedings will be published in the ACL Anthology.

IMPORTANT DATES

Deadlines are "anywhere on Earth."

6 Mar 2024	Extended paper submission deadline
25 Mar 2024	Notification of acceptance
7 Apr 2024	Camera-ready final papers
20 May 2024	Workshop date

For updates, please follow the present Web page.

SUBMISSION GUIDELINES

Please follow the style sheet and templates (for LaTeX, Overleaf, Open Office, and MS-Word) provided for the main conference.

Papers should be submitted as a PDF file using the START conference manager.

Submissions must describe original and unpublished work and range from 4 to 8 pages plus unlimited references. **Camera-ready final versions may use one more page than the initial submission to take into account the reviewers' comments.**

Reviewing will be double blind, so the papers should not reveal the authors' identity. Accepted papers will be published in the workshop proceedings, which will be included in the ACL Anthology.

Double submission policy: Parallel submission to other meetings or publications is possible but must be immediately (i.e. as soon as known to the authors) notified to the workshop organizers by e-mail.

Presentation slides Due to the size of the conference rooms it is recommended to use 36 pt fonts for the presentation slides. posters

Posters The size of posters holders is 90cm x 150 cm and the format is vertical (Portrait). The Poster Boards cannot accommodate Landscape posters. You can print your poster in Portrait A0 (84,1 x 118,9cm).

For further information and updates see the present Web page.

Plain-text CFP : bucc2024-cfp.txt
PDF CFP : bucc2024-cfp.pdf
Last modified: 16 Apr 2024

ORGANIZERS AND CONTACT

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)

Contact for workshop: pz (at) lisn (dot) fr

PROGRAMME 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)

Last modified: 14 Jan 2024