CFP: 13th BUCC Workshop at LREC with Shared Task on Bilingual Dictionary Induction from Comparable Corpora

.
****

13th WORKSHOP ON BUILDING AND USING COMPARABLE CORPORA

Co-located with LREC 2020, Pharo Palace, Marseille, France

Monday, May 11, 2020

Submission deadline: February 20, 2018

SHARED TASK: Bilingual dictionary induction from comparable corpora

Website: https://comparable.limsi.fr/bucc2020/

****

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 in themselves by making possible cross-language 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.

TOPICS

We solicit contributions on all topics related to comparable corpora, including but not limited to the following:

Building Comparable Corpora:

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

Mining from Comparable Corpora:

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

- Induction of multilingual word classes from comparable corpora
- Cross-language distributional semantics

SUBMISSION INFORMATION

Please follow the style sheet and templates provided for the main conference at
http://lrec2020.lrec-conf.org/en/submission/authors-kit/
Further details on the submission procedure will be provided on the workshop website later on.
Papers should be submitted as a PDF file. Submissions must describe original and unpublished work and range from 4 to 8 pages excluding references.
Reviewing will be double blind, so the papers should not reveal the authors' identity. Accepted papers will be published in the workshop proceedings.
Double submission policy: Parallel submission to other meetings or publications is possible but must be immediately notified to the workshop organizers.

For further information see the BUCC 2018 website:
http://comparable.limsi.fr/bucc2020/

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

IMPORTANT DATES

25 February 2020: Paper submission deadline
12 March 2020: Notification of acceptance
mid March 2020 (tentative): Early bird registration (reduced rates)
2 April, 2020: Camera ready final papers
May 11, 2020: Workshop date

SHARED TASK: Bilingual dictionary induction from comparable corpora

In the framework of machine translation, the extraction of bilingual dictionaries from parallel corpora has been conducted very successfully. On the other hand, human second language acquisition appears not to be based on parallel data. This means that there must be a way of acquiring and relating lexical knowledge in two or more languages without the use of parallel data.

It has been suggested that it might also be possible to extract multilingual lexical knowledge from comparable rather than from parallel corpora. From a theoretical perspective, this suggestion might lead to advances in understanding human second language acquisition. From a practical perspective, as comparable corpora are available in much larger quantities than parallel corpora, this approach might help in relieving the data acqisition bottleneck which tends to be especially severe when dealing with language pairs involving low resource languages.

A well established practical task to approach this topic is bilingual lexicon induction from comparable corpora, which is in the focus of the current shared task. Typically, its aim is to extract word equations such as the following from comparable corpora:

English / French

baby <-> bébé
baby <-> poupon
bath <-> bain
bed <-> lit
bed <-> plumard
convenience <-> commodité
doctor <-> médecin
doctor <-> docteur
eagle <-> aigle
mountain <-> montagne
nervous <-> nerveux
work <-> travail

Quite a few research groups have been working on this problem using a wide variety of approaches. However, as there is no standard way to measure the performance of the systems, the published results are not comparable and the pros and cons of the various approaches are not clear. The shared task aims at solving these problems by organizing a fair competition between systems. This is accomplished by providing corpora and evaluation datasets for a number of language pairs involving English, French, and German, and by comparing the results using a common evaluation framework. Other language pairs might be added on request.

Any submission to the shared task is expected to be accompanied by a short paper (4 to 6 pages plus references). This will be accepted for publication in the workshop proceedings after a basic quality check.

Note that participation in the workshop, although we strongly encourage it, is not mandatory for participating in the shared task.

Further information on the shared task as well as the data sets will be provided on the workshop website at https://comparable.limsi.fr/bucc2020/

SHARED TASK SCHEDULE

Any time: Expression of interest (not compulsory)
December 31, 2019: Release of shared task training sets
1 February 2020: Release of shared task test sets
25 February 2020: Submission deadline for shared task results
29 February 2020: Shared task paper submission deadline
12 March 2020: Reviewers' feedback
2 April 2020: Shared task camera ready papers
May 11, 2020: Workshop taking place at LREC 2020

For further information concerning the shared task see https://comparable.limsi.fr/bucc2018/bucc2018-task.html or contact reinhardrapp (at] gmx (dot) de

WORKSHOP AND SHARED TASK ORGANIZERS

Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany), Chair and contact person: reinhardrapp (at] gmx (dot) de
Pierre Zweigenbaum (LIMS, CNRS, Université Paris-Saclay, Orsay, France)
Serge Sharoff (University of Leeds, United Kingdom)

PROGRAMME COMMITTEE

Ahmet Aker (University of Sheffield, UK)
Ebrahim Ansari (Institue for Advanced Studies in Basic Sciences, Iran)
Hervé Déjean (Naver Labs Europe, Grenoble, France)
Thierry Etchegoyhen (Vicomtech, Spain)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology, Japan)
Kyo Kageura (The University of Tokyo, Japan)
Yves Lepage (Waseda University, Japan)
Sheervin Malmasi (Harvard Medical School, Boston, MA, USA)
Michael Mohler (Language Computer Corp., USA)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., USA)
Ted Pedersen (University of Minnesota, Duluth, USA)
Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Richard Sproat (OGI School of Science & Technology, USA)
Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France)

INFORMATION FROM THE LREC ORGANIZERS

Please make sure that your papers take into account the following information about the LRE Map, the "Share your LRs!" initiative and the ISLRN number:

Describing your LRs in the LRE Map is now a normal practice in the submission procedure of LREC (introduced in 2010 and adopted by other conferences). To continue the efforts initiated at LREC 2014 about “Sharing LRs” (data, tools, web-services, etc.), authors will have the possibility, when submitting a paper, to upload LRs in a special LREC repository. This effort of sharing LRs, linked to the LRE Map for their description, may become a new “regular” feature for conferences in our field, thus contributing to creating a common repository where everyone can deposit and share data.

As scientific work requires accurate citations of referenced work so as to allow the community to understand the whole context and also replicate the experiments conducted by other researchers, LREC 2020 endorses the need to uniquely Identify LRs through the use of the International Standard Language Resource Number (ISLRN, www.islrn.org), a Persistent Unique Identifier to be assigned to each Language Resource. The assignment of ISLRNs to LRs cited in LREC papers will be offered at submission time.