BUCC, 9th Workshop on Building and Using Comparable Corpora

Co-located with LREC 2016
Portorož (Slovenia)
23 May 2016, room Adria
Special Topic: Continuous Vector Space Models and Comparable Corpora
https://comparable.limsi.fr/bucc2016/

Invited Speakers: Ruslan Mitkov, Gregory Grefenstette
Shared Task preparation panel: Identifying Parallel Sentences in Comparable Corpora

INVITED SPEAKERS

Ruslan Mitkov

University of Wolverhampton

The Name of the Game is Comparable Corpora

Comparable corpora are the most versatile and valuable resource for multilingual Natural Language Processing. The speaker will argue that comparable corpora can support a wider range of applications than has been demonstrated so far in the state of the art. The talk will present completed and ongoing work conducted by the speaker and colleagues from his research group where comparable corpora are employed for different tasks including but not limited to the identification of cognates and false friends, validation of translation universals, language change and translation of multiword expressions.

Gregory Grefenstette

Inria Saclay/TAO, Université Paris-Saclay

Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research

Comparable Corpora have been used to improve statistical machine translation, for augmenting linked open data, for finding terminology equivalents, and to create other linguistic resources for natural language processing and language learning applications. Recently, continuous vector space models, creating and exploiting word embeddings, have been gaining in popularity in more powerful

solutions to creating, and sometimes replacing, these resources. Both classical comparable corpora solutions and vector space models require the presence of a large quantity of multilingual content. In this talk, we will discuss the breadth of this content on the internet to provide some type of intuition in how successful comparable corpus approaches will be in achieving its goals of providing multilingual and cross lingual resources. We examine current estimates of language presence and growth on the web, and of the availability of the type of resources needed to continue and extend comparable corpus research.

MOTIVATION

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 chiefly 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.

TOPICS

Beyond this year's special topic "Continuous Vector Space Models and Comparable Corpora" and the shared task on "Identifying Parallel Segments in Comparable Corpora", we solicit contributions including but not limited to the following topics:

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

Mining from Comparable Corpora:

- Cross-language distributional semantics

- Extraction of parallel segments or paraphrases from comparable corpora
- Extraction of translations of single words and multi-word expressions, proper names, named entities, etc., from comparable corpora

IMPORTANT DATES

24 February 2016	Deadline for submission of full papers (extended)
10 March 2016	Notification of acceptance
27 March 2016	Camera-ready papers due
23 May 2016	Workshop date

SUBMISSION INFORMATION

Papers should follow the LREC main conference formatting details (to be announced on the conference website http://lrec2016.lrec-conf.org/) and should be submitted as a PDF-file via the START workshop manager at https://www.softconf.com/lrec2016/BUCC2016/.

Contributions can be short or long papers. Short paper submission must describe original and unpublished work without exceeding six (6) pages. Characteristics of short papers include: a small, focused contribution; work in progress; a negative result; an opinion piece; an interesting application nugget. Long paper submissions must describe substantial, original, completed and unpublished work without exceeding ten (10) pages.

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.

Please also observe the following two paragraphs which are applicable to all LREC workshops as well as to the main conference:

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 2016 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.

For further information, please contact Reinhard Rapp <reinhardrapp (at) gmx (dot) de>

Plain-text CFP : bucc2016-cfp.txt
PDF CFP : bucc2016-cfp.pdf
Last modified: 23 April 2016

ORGANISERS

Reinhard Rapp University of Mainz (Germany), Chair

Pierre Zweigenbaum LIMSI, CNRS, Orsay (France), Shared Task Chair

Serge Sharoff University of Leeds (UK)

SCIENTIFIC COMMITTEE

Ahmet Aker, University of Sheffield (UK)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Vishal Goyal (Punjabi University, Patiala, India)
Gregory Grefenstette (INRIA, Saclay, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology)
Kyo Kageura (University of Tokyo, Japan)
Philippe Langlais (Université de Montréal, Canada)
Shervin Malmasi (Harvard Medical School, Boston, MA, USA)
Michael Mohler (Language Computer Corp., US)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Lene Offersgaard (University of Copenhagen, Denmark)
Ted Pedersen (University of Minnesota, Duluth, US)
Reinhard Rapp (Université Aix-Marseille, France)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

SHARED TASK

A shared task on "Identifying Parallel Sentences in Comparable Corpora" will be discussed in the panel. See the shared task data preparation paper.