BUCC, 15th Workshop on Building and Using Comparable Corpora

Co-located with LREC 2022, Marseille, France, 25 June 2022
Invited speaker: **Alexander Fraser**
Shared Task: bilingual term alignment in comparable corpora
Website: https://comparable.limsi.fr/bucc2022/
Programme as of 17 June 2022
Proceedings

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 information 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 and pre-trained multilingual transformer models
- Creation of bilingual and multilingual embeddings 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

PRACTICAL INFORMATION

Registration will be via the main conference website LREC 2022

IMPORTANT DATES

April ~~10~~ →**20**, 2022	Paper submission deadline: extended to April 20
May 3, 2022	Notification to authors
May 23, 2022	Camera-ready final papers
June 25, 2022	Workshop date

SUBMISSION GUIDELINES

Please follow the style sheet and templates provided for the main conference at https://lrec2022.lrec-conf.org/en/submis
Papers should be submitted as a PDF file using the START conference manager at https://www.softconf.com/lrec2022/B
Submissions must describe original and unpublished work and range from 4 to 8 pages plus unlimited references.

It is the authors' choice whether or not to reveal their identities in their manuscripts submitted for review. 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 by e-mail.

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

Information from the LREC organizers

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 2022 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.

Plain-text CFP : bucc2022-cfp.txt
PDF CFP : bucc2022-cfp.pdf
Last modified: 25 June 2022

BUCC 2022 SHARED TASK: bilingual term alignment in comparable specialized corpora.

The BUCC 2022 shared task is on multilingual terminology alignment in comparable corpora. Many research groups are 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 comparison of systems. This is accomplished by providing corpora and evaluation datasets for a number of language pairs and domains.

Moreover, the importance of dealing with multi-word expressions in Natural Language Processing applications has been recognized for a long time. In particular, multi-word expressions pose serious challenges for machine translation systems because of their syntactic and semantic properties. Furthermore, multi-word expressions tend to be more frequent in domain-specific text, hence the need to handle them in tasks with specialized-domain corpora.

Through the 2022 BUCC shared task, we seek to evaluate methods that detect pairs of terms that are translations of each other in two comparable corpora, with an emphasis on multi-word terms in specialized domains.

Provided resources

The BUCC shared task provides several datasets of the following form:

- A pair of comparable corpora C_1 and C_2 in languages L_1 and L_2.
- A list of terms D_1 that occur in C_1 and a list of terms D_2 that occur in C_2. Term lists may include both single-word and multi-word terms.
- For training only, a gold standard dictionary $D_{1,2}$ in the form of a list of pairs of terms (t_1, t_2) that are translations of each other, with t_1 in D_1 and t_2 in D_2.

The task participants may additionally use any external resources, except the CCAligned corpora, from which the task datasets have been extracted. When reporting their results, participants are required to specify which resources they used. They are also encouraged to test conditions in which they only use the provided resources.

Task

Given a test dataset with comparable corpora C_1 and C_2, and lists of terms D_1 and D_2, participant systems are expected to produce an ordered list of term pairs in (D_1, D_2) that are translations of each other, in descending order of confidence.

Note that D_1 and D_2 may have different sizes, that not every term in D_1 may have a translation in D_2, that some terms in D_1 might have multiple translations, and conversely. For practical reasons, we limit the length of a submitted term pair list to a ceiling of 10 times the average length of D_1 and D_2. (This can be seen as meaning that, on average, a system may submit up to 10 alignment hypotheses for each term in D_1 or in D_2.)

The test datasets will include both the same language pairs as those provided for training and also other language pairs.

Participants can submit up to 5 system runs for each test dataset.

Evaluation

The evaluation metric will be the Average Precision of the predicted bilingual term pair list, where the relevance of a term pair is determined by its presence in the (hidden) gold standard dictionary $D_{1,2}$. This models the task as an information retrieval task: retrieve all relevant term pairs (t_1, t_2) (documents) from the cross-product $D_1 \times D_2$ (virtual pool of documents), presenting them in descending order of confidence. Average Precision is the area under the recall $\times$ precision curve. It is computed as the average over all m relevant term pairs (t_i, t_j) (i.e., all term pairs in the gold standard) of the precision value obtained for the set of top n_k term pairs existing after each relevant term pair (t_{i_k}, t_{j_k}) is retrieved, from the first to the last relevant term pair. Relevant term pairs that are not retrieved receive a precision of zero, hence decrease Average Precision. Average Precision (AP) is defined as:

$$AP = \frac{1}{m} \sum_{k=1}^{m} P(R_k)$$

where R_k is the set of ranked predicted term pairs from the top to the position at which k relevant term pairs have been retrieved. Given the gold standard dictionary $D_{1,2}$, the precision of a set of predicted term pairs R is defined as $P(R) = \frac{|R \cap D_{1,2}|}{|R|}$.

To optimize Average Precision, a system must find all relevant term pairs and put them at the top of the list. Average Precision increases when true predictions (relevant term pairs) are added anywhere in the prediction list. Average Precision also increases when false predictions, if any, are pushed towards the bottom of the list. Note that Average Precision cannot decrease when more predictions, whether true or false, are added to the bottom of the list. Also note that Average Precision is equivalent to Mean Average Precision (MAP) with exactly one query (*find all term pairs in $D_1 \times D_2$ that are translations of each other*).

File format

All files use UTF-8 encoding, with LF end-of-line markers.

- Single-term lists D_1 and D_2 contain one term per line.
- Corpora C_1 and C_2 contain one sentence per line.
- The gold standard dictionary $D_{1,2}$ contains two terms per line, separated by a tabulation: t_1<TAB>t_2
- The system output submitted by a participant contains two terms per line, separated by a tabulation <TAB>. Its lines are ordered in decreasing order of confidence.

Sample data

A small sample dataset is provided in bucc2022_sample.zip for the English-French language pair. It contains:

- A pair of comparable corpora C_1 =*src_corpus_sample.txt* and C_2 =*tgt_corpus_sample.txt* in languages L_1 =*en* and L_2 =*fr*.
- A list of terms D_1 =*src_term_list_sample.txt* that occur in C_1 and a list of terms D_2 =*tgt_term_list_sample.txt* that occur in C_2. Term lists may include both single-word and multi-word terms.
- For training only, a gold standard dictionary $D_{1,2}$ =*gold_dictionary_sample.txt* in *en*×*fr*.

Training data

A training dataset is provided in bucc2022_training.zip for the English-French language pair. It contains:

- A pair of comparable corpora C_1 =*corpus-en.txt* and C_2 =*corpus-fr.txt* in languages L_1 =*en* and L_2 =*fr*.
- A list of terms D_1 =*terms-en.txt* that occur in C_1 and a list of terms D_2 =*terms-fr.txt* that occur in C_2. Term lists may include both single-word and multi-word terms.
- For training only, a gold standard dictionary $D_{1,2}$ =*terms-en-fr.txt* in *en*×*fr*.

Note that the sizes of D_1, D_2 and $D_{1,2}$as well as the proportions of terms in D_1or D_2 that have a translation in $D_{1,2}$ are likely to be different in the test datasets.

Test data

Test datasets (en-fr: bucc2022_test_enfr_nogold.zip; en-de; en-ru). A test dataset contains:

- A pair of comparable corpora C_1 =*corpus-en.txt* and C_2 =*corpus-fr.txt* in languages L_1 =*en* and L_2 =*fr*.
- A list of terms D_1 =*terms-en.txt* that occur in C_1 and a list of terms D_2 =*terms-fr.txt* that occur in C_2. Term lists may include both single-word and multi-word terms.

Time schedule

Any time	Expression of interest to all three contact points of the shared task. This will allow us to register y
19 January 2022	Sample dataset release
13 February 2022	Training data release (en-fr)
19 March 2022	Test data release (1: en-fr)
26 March 2022	Submission of system runs by participants (up to 5 per dataset) by e-mail to all three contact poin
30 March 2022	Evaluation sent to participants
10 April 2022	Submission of shared task papers to the BUCC workshop
25 June 2022	Workshop date

Shared task organizers and contact

Omar Adjali (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Emmanuel Morin (Nantes Université, LS2N, Nantes, France)

Serge Sharoff (University of Leeds, United Kingdom)

Reinhard Rapp (Athena R.C., Greece; Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)

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

Shared task contact points: please send expressions of interest to:

- omar (dot) adjali (at) universite-paris-saclay (dot) fr
- CC emmanuel (dot) morin (at) ls2n (dot) fr
- CC pz (at) lisn (dot) fr

Last modified: 19 March 2022

WORKSHOP ORGANIZERS AND CONTACT

Reinhard Rapp (Athena R.C., Greece; Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)

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

Serge Sharoff (University of Leeds, United Kingdom)

Contact for workshop: reinhardrapp (at) gmx (dot) de
Contact for shared task: pz (at) lisn (dot) fr

PROGRAMME COMMITTEE

- Ahmet Aker (University of Sheffield, UK)
- 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 (Université de Paris, France)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Université de Nantes, France)
- Dragos Stefan Munteanu (Language Weaver, Inc., USA)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Reinhard Rapp (Athena R.C., Greece; 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)
- Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Last modified: 26 Apr 2022