BUCC, 13th Workshop on Building and Using Comparable Corpora

Co-located with LREC 2020
Live at https://bbb.limsi.fr/b/pie-636-a4v
Monday, May 11, 2020
Shared task: Bilingual dictionary induction from comparable corpora
Website: https://comparable.limsi.fr/bucc2020/
Invited speakers: Holger Schwenk, Facebook AI Research, and Jörg Tiedemann, University of Helsinki

INVITED SPEAKER

Holger Schwenk

Facebook AI Research

BUCC 2020 is held on-line, live

The workshop **takes place as an online event on May 11** from 9:15 to 16:35 **UTC+2** (Central European Summer Time).

See the Program page for details. Attendance is free but please send e-mail to reinhardrapp (at) gmx (dot) de to obtain further information.

Details for attendance

The workshop **takes place as an online event on May 11** from 9:15 to 16:35 **UTC+2** (Central European Summer Time).

Link to attend: https://bbb.limsi.fr/b/pie-636-a4v

This uses BigBlueButton (BBB), a free software (http://docs.bigbluebutton.org/) which works through WebRTC, meaning that you just need to use your Web browser, no need to install specific software. This works with Chrome, Firefox, Safari, and also with the most recent version of Edge (Jan. 2020). On your smartphone or tablet use your Web browser too: Chrome, Firefox, Safari. There is no way to connect using a simple phone on this installation of BBB.

Attendance is free but please send e-mail to reinhardrapp (at) gmx (dot) de to obtain further information.

All attendees

For planning purposes, we encourage (but do not require) attendees to announce their intention to participate to pz (at) limsi (dot) fr.

Just connect to the above-specified BBB link, enter your given name and family name, and join the meeting.

- When connecting, please choose the "**Listen Only**" option: this will lower the bandwidth and server load. This will keep your microphone inactive during the meeting. Unmute it only to ask a question at question time (in Listen Only mode, click on the headphones button, bottom center of the screen, to leave the audio, then click it again to join audio and choose the Microphone option). More conveniently, you can ask questions in the Public Chat pane of the BBB window. The session chair will collect the questions and ask them to the presenter at question time.
- Please **do not share your webcam** except when presenting.
- There is a Public Chat pane on the left of the BBB screen: you can display it or not.
- You can switch the presentation or video to full screen mode for more comfortable viewing (click on the arrows button at the bottom right of the presentation).
- Please see http://docs.bigbluebutton.org/ for more information.

Authors

You can either

- Upload a PDF presentation to BBB and page through it while presenting live. See below for details on how to present on BBB.
- Record a video presentation, upload it to a video sharing platform, and share the link to its address. YouTube, Vimeo, Instructure Media, Twitch and Daily Motion URLs are supported by BBB. See below for details on how to present on BBB.

Please test your connection and presentation before the workshop. For this purpose, the presentation room is already opened for authors ahead of time. Just connect to the above-specified BBB link, enter your name and join or start the meeting, click on the "+" button at the bottom left of the window, click on "become a presenter" if needed.

- PDF: once you have Presenter status, the same "+" button will give you access to an "Upload a presentation" entry. Drag your PDF presentation to the designated location, then click the Confirm button. Use the arrows at the bottom of the screen to move forward and backward.
- Recorded video: once you have Presenter status, the same "+" button will give you access to a "Share an external video" entry. Paste the URL of your video, then click the Share a new video button. Click the start button to play it.

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 and word embeddings

IMPORTANT DATES

4 March 2020	Paper submission deadline
12 March 2020	Notification to authors
19 March 2020	Early bird registration (reduced rates)
2 April 2020	Camera-ready final papers
11 May 2020	Workshop date

SUBMISSION INFORMATION

Please follow the style sheet and templates provided for the main conference at https://lrec2020.lrec-conf.org/en/submis
Papers should be submitted as a PDF file at https://www.softconf.com/lrec2020/BUCC2020/. 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.

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 : bucc2020-cfp.txt
PDF CFP : bucc2020-cfp.pdf
Last modified: 8 May 2020

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.

ORGANISERS

Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany), Chair and contact person: <reinhardrapp (at] gmx (dot) de>

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

Serge Sharoff (University of Leeds, United Kingdom)

SCIENTIFIC COMMITTEE

Ahmet Aker (University of Sheffield, UK)
Ebrahim Ansari (Institute 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)
Shervin 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, US)
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 (Université Paris-Saclay, CNRS, LIMSI, Orsay, France)

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 acquisition 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 extraction from comparable corpora, which is in the focus of this shared task. Typically, its aim is to extract word translations such as the following from comparable corpora, where a given source word may receive multiple translations:

Source (English)	**Target (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 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 competition between systems. This is accomplished by providing corpora and bilingual datasets for a number of language pairs involving Chinese, English, French, German, Russian and Spanish, and by comparing the results using a common evaluation framework. For the shared task we provide corpora as well as training data. However, as these corpora and data may not suit all needs, we divide the shared task into two tracks.

- In the "closed track," participants are required to only use the data provided by us. In this way equal conditions are ensured and, as the outcome of this track, the systems can be ranked according to the quality of their results.

- In the "open track," participants are free to use their own corpora and training data. If possible, they should still use our evaluation data, but this is also not mandatory. The participants can even work on languages for which the shared task provides no data. If relevant, the participants should describe why their systems are not suitable for the closed track, and discuss the pros and cons of their choices. If possible, they should also provide access to their data for the purpose of facilitating replication by others.

How to participate

Research groups or individual researchers can participate in one or both tracks (further details see below), choose the language pairs they wish to work on and can suggest new language pairs for which we will try to provide support. Participation in the shared task is expected to be accompanied by a system description paper (4 to 6 pages plus references). Ideally, this paper gives a description of the participating system in a way that allows replication of the work. As the shared task is supposed to compare as many different systems as possible (i.e. including systems based on well-established techniques) the scientific content of the paper needs not necessarily be novel. Nevertheless, the papers will be peer reviewed and (apart from novelty) the usual quality criteria for research papers will be applied for the papers to be published in the workshop proceedings.

Note that participation in the workshop, although we strongly encourage it, is not mandatory for participating in the shared task and for publication of the system description papers.

Checklist for participants

- Decide on the track you wish to participate in and on your language pairs.
- Express your interest to reinhardrapp (at) gmx (dot) de so that we can inform you about any possible updates, changes, issues etc. Please mention your track and the language pair(s) you are interested in. You may also suggest new language pairs, and we might be able to help you with data.
- Download the corpora from this webpage (WaCKy or Wikipedia, see below)
- Download the training data (bilingual word pairs) for your language pairs from this webpage, see below.
- Run your system on the words on the source side of the training data and compute the translations. Compare your results with target side of the training data and improve your system if necessary.
- Download the test data on the date specified in the time schedule below.
- Run your system on the test data. Format your output in the same way as you see in the training data.
- Before the deadline specified in the schedule (check for any extensions!), submit your results by e-mail to reinhardrapp (at) gmx (dot) de. Evaluation results will be sent to you after that deadline.
- Write and submit a system description paper.
- Present your paper at the workshop. (If you cannot participate, please let us know in time.) Please see the LREC website for registration information.

Time schedule for shared task

Any time	Expression of interest to reinhardrapp (at) gmx (dot) de (including suggestions for additional lang
12 January 2020	Release of shared task training sets (done)
16 February 2020	Release of shared task test sets (done)

5 March 2020	Submission of shared task results by e-mail to reinhardrapp (at) gmx (de)
15 March 2020	Submission of shared task system description papers
11 May 2020	Workshop date

Track 1: Closed Track

The supported language pairs (for which we provide data) are the following:

table-corpora

The cells in the table show which type of corpus should be used for both languages of a pair when conducting the dictionary induction task. The rationale behind these choices is that the WaCKy (Web-as-a-corpus initiative) corpora seem somewhat better suited for the dictionary induction task than Wikipedia, but they are not available for Chinese and Spanish. Language pairs involving Chinese and Spanish therefore use Wikipedia corpora, whereas other language pairs use WaCKy corpora.

The WaCKy corpora can be downloaded from the links below (download can take from less than one minute to hours depending on your connection speed). For convenience, we also provide pre-trained fastText embeddings for these corpora (see below () for details):

Corpus	Language	Corpus	fastText embeddings
UKWaC	English	http://corpus.leeds.ac.uk/serge/bucc/ukwac.ol.xz (3.0Gb)	bin (3.2Gb), vec.xz (0.3Gb)
FRWAC	French	http://corpus.leeds.ac.uk/serge/bucc/frwac.ol.xz (1.8Gb)	bin (3.0Gb), vec.xz (0.3Gb)
DEWAC	German	http://corpus.leeds.ac.uk/serge/bucc/dewac.ol.xz (2.4Gb)	bin (3.0Gb), vec.xz (0.5Gb)
RUWAC	Russian	http://corpus.leeds.ac.uk/serge/bucc/ruwac.ol.xz (3.1Gb)	bin (4.1Gb), vec.xz (0.7Gb)

The WaCKy corpora are cleaned-up web crawls of approximately 2 billion words per language. They are kindly provided by the Web-as-a-corpus initiative (WaCKy). For further information see http://wacky.sslmit.unibo.it/doku.php?id=corpora.

If you use the WaCKy-corpora, please cite the following paper:

- M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Lar *Language Resources and Evaluation* 43(3): 209–226.

The Wikipedia corpora can be downloaded from the links below (download can take from less than one minute to hours depending on your connection speed). For convenience, we also point at pre-trained fastText embeddings for these corpora prepared at Facebook (see below () for details):

Corpus	Language	Corpus	fastText embeddings
enWiki	English	http://corpus.leeds.ac.uk/serge/bucc/en.ol.xz (3.6Gb)	bin+vec, zipped (9.6Gb) vec (6.1Gb)
esWiki	Spanish	http://corpus.leeds.ac.uk/serge/bucc/es.ol.xz (0.9Gb)	bin+vec, zipped (5.1Gb) vec (2.4Gb)
zhWiki	Chinese	http://corpus.leeds.ac.uk/serge/bucc/zh.ol.xz (0.4Gb)	bin+vec, zipped (3.1Gb) vec (0.8Gb)

These corpora are in a one-line per document format. The first tab-separated field in each line contains metadata, the second field contains the text. Paragraph boundaries are marked with HTML tags. As cleaning up the original Wikipedia dump files is not trivial, occasionally there can be some noise in the form of not fully cleaned HTML and Javascript fragments.

Bilingual word pairs. For checking and improving the performance of your systems, please use the following training data which consists of tab-separated bilingual word pairs:

table-training-pairs

Rather than providing one large set of word pairs for each language pair, by splitting into frequency ranges we provide three smaller sets. Looking at different frequency ranges is of scientific interest as algorithms typically work best for high frequency words, whereas the performance at low frequencies is of higher practical relevance.

We split the data into three sets corresponding to frequency ranges of the source language words: The high frequency set provides bilingual word pairs where the frequency is among the 5000 most frequent words. The mid frequency sets consist of words ranking between 5001 and 20000, and the

low frequency set belongs to ranks 20001 to 50000. (For languages where not enough data is available, we had to reduce the size of the bins.)

Each set is a random sample extracted from the MUSE data kindly provided by facebook AI Research and comprises 2000 different source language words together with their translations. Like in the original MUSE data, the source language words are ordered according to frequency (most frequent first). All three sets (per language pair) taken together, this gives 6000 source language words together with their translations, whereby each translation is listed in a separate line.

If you use any of these datasets, please cite the following paper:

- Conneau, Alexis; Lample, Guillaume; Ranzato, Marc Aurelio ; Denoyer, Ludovic ; Jégou, Hervé (2017). Word translation without parallel data. arXiv preprint 1710.04087.

As described in this paper, the MUSE dictionaries, which take the polysemy of words into account, were created using a facebook internal translation tool. Given that they were generated automatically, they are of high quality, but still contain a few errors. Participants of the shared task are encouraged to report to us such errors, so that, as a positive side effect of the shared task, the datasets can be improved.

For testing the systems, lists of source language test words were provided on the day listed in the above time schedule, which are likewise split into three sets of 2000 words each:

table-test-words

If your algorithm for inducing dictionaries from comparable corpora requires a **seed lexicon**, then please use an arbitrary part of the training data for this purpose. We hope that with its 6000 source language words and (depending on the language pair) roughly twice as many translation pairs, the training set is large enough to provide for your needs. If not, please consider using your own data and participating in Track 2 of the shared task.

Pre-trained embedding models such as fastText or BERT can be used only if (re)trained on the provided corpora. The following fastText embeddings have been trained on Wikipedia or WaCKy corpora and can be readily used in this track (specific links are provided in the table above ()):

- Wikipedia: fastText embeddings for the Wikipedia corpora are available from Facebook: https://fasttext.cc/docs/e
- WaCKy: pre-trained fastText embeddings are available as follows:
  - The .vec.xz files are text representations, widely used in various tools.
  - The .bin files are the binary versions for use in Fasttext.
  - The following parameters were used: method: skipgram; minCount: 30; dim: 300; ws (context window): 7; epochs: 10; neg (number of negatives sampled): 10. The other parameters are as defaults for fastText.

Track 2: Open Track

In this track, participants are free to work on other language pairs, use their own data and–if desired–conduct their own evaluation procedure. However, it would be very helpful if in their papers they described their reasons and motivation for deviating from the procedures of Track 1 and–if possible–provided access to their data.

Please also let us know about your plans in time as we may be able to support you with corpora and datasets.

As this appears to be the first shared task on the topic of dictionary induction from comparable corpora, we cannot draw on previous experiences. Due to this pilot character, in Track 1 we are trying to keep things as clear and unsophisticated as possible. But in Track 2 we encourage you to challenge this simplicity, to freely experiment and to come up with new ideas in the hope that the resulting insights will promote future progress in the field.

Evaluation

For evaluation, Track 1 participants (for Track 2 participants this is optional) are asked to provide their results on the test data sets for the test words in each of the three frequency ranges. Hereby

it is expected that for each source language word all its major translations are provided (where the definition of "major" is supposed to be inferred from the training data). The shared task organizers compare these translations to the translations as found in their (internal) gold standard data which is structurally similar to the training data. Only identical strings are considered correct, and the performance of the respective system is determined by computing precision, recall, and F1-score, the latter being the official score for system ranking. All data sets are in utf-8 encoding.

More precisely: the input to the system is a list of source language words, one per line. A system should return, for each input word w^s, one or more candidate translations w_i^t, in the form of tab-separated word pairs $w^s \backslash t\, w_i^t$, each on its own line. For instance, in the English-French case, given the following gold standard, test word list, and system output (tab-separated word pairs), the system would get credited for two true positives, one false positive, and two false negatives, hence $P = 2/3 = 0.67, R = 2/4 = 0.50, F = 0.57$.

gold standard

bed	lit
bed	plumard
doctor	médecin
doctor	docteur

test set

bed
doctor

system output

bed	lit
bed	futon
doctor	docteur

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 (Université Paris-Saclay, CNRS, LIMSI, Orsay, France)
- Serge Sharoff (University of Leeds, United Kingdom)

Previous BUCC shared tasks and datasets:

- Identifying parallel sentences in comparable corpora (BUCC 2018, 2017)
- Identifying comparable text (BUCC 2015)

Last modified: 8 May 2020