BUCC, 11th Workshop on Building and Using Comparable Corpora

Shared task: Identifying parallel sentences in comparable corpora
Special theme: Comparable Corpora for Asian Languages
Co-located with LREC 2018,
Phoenix Seagaia Resort, Miyzaki, Japan
Tuesday, May 8, 2018
Web site: https://comparable.limsi.fr/bucc2018/
LREC and workshop registration

Invited Speakers:
Kyo Kageura, University of Tokyo
Yves Lepage, Waseda University

INVITED SPEAKERS

Kyo Kageura

The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
kyo@p.u-tokyo.ac.jp

Cross-lingual Correspondences of Terms in Texts and Terminologies: Theoretical Issues and Practical Implications

Terms are items in language that represent concepts. This relation of representation does not change through use. As such, terms have a unique status in language, second only to proper names. Due to this, clarifying the identity of concepts represented by terms becomes an important issue at the level of what is represented, and control of terms representing the same concept also becomes an important issue at the level of representation. These problems with which terminologists are concerned, though not clear at first glance, are in fact relevant to general words and vocabulary to a lesser extent. In this paper I first clarify theoretical issues of terms and terminologies and what they imply for terminology processing in particular and lexical and lexicological processing in general. I then pick up some terminological applications, examine their status and suggest a few issues that can be addressed in terminology processing.

Yves Lepage

Waseda University
808-0135 Fukuoka-ken, Kitakyûsyû-si, Wakamatu-ku, Hibikino 2-7, Japan
yves.lepage@waseda.jp

Quasi-Parallel Corpora: Hallucinating Translations for the Chinese–Japanese Language Pair

We show how to address the problem of bilingual data scarcity in machine translation. We propose a method that generates aligned sentences which may be not perfect translations. It consists in 'hallucinating' new sentences which contain small but well-attested variations extracted from unaligned unrelated monolingual data. We conducted various experiments in statistical machine translation between Chinese and Japanese to determine when adding such quasi-parallel data to a basic training corpus leads to increases in translation accuracy as measured by BLEU.

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 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

Given that LREC takes place for the first time in Asia, this year's special theme is *Comparable Corpora for Asian Languages*. But we solicit contributions also on all other 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:

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

IMPORTANT DATES

30 January 2018	Paper submission deadline (extended)
14 February 2018	Notification to authors
15 February 2018	Early bird registration (reduced rates)
25 February 2018	Camera-ready papers due
8 May 2018	Workshop date

SUBMISSION INFORMATION

Please follow the style sheet and templates provided for the main conference at http://lrec2018.lrec-conf.org/en/submission/authors-kit/. Papers should be submitted as a PDF file at http://softconf.com/lrec2018/BUCC2018/. Submissions must describe original and unpublished work and range from four (4) to eight (8) pages including 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, please contact Serge Sharoff <S (dot) Sharoff (at) leeds (dot) ac (dot) uk>

Plain-text CFP : bucc2018-cfp.txt
PDF CFP : bucc2018-cfp.pdf
Last modified: 22 Apr 2018

ORGANISERS

Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany), Chair

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

Serge Sharoff (University of Leeds, United Kingdom)

SCIENTIFIC COMMITTEE

Ahmet Aker (University of Sheffield, UK)
Caroline Barrière (CRIM, Montréal, Canada)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (INRIA, Saclay, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Kyo Kageura (University of Tokyo, Japan)
Natalie Kübler (Université Paris Diderot, France)
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 (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, US)
Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France)

SHARED TASK

Shared task: **identifying parallel sentences in comparable corpora**

As a continuation of the previous year's shared task, we announce a modified shared task for 2018. As is well known, a bottleneck in statistical machine translation is the scarceness of parallel resources for many language pairs and domains. Previous research has shown that this bottleneck can be reduced by utilizing parallel portions found within comparable corpora. These are useful for many purposes, including automatic terminology extraction and the training of statistical MT systems. The aim of the shared task is to quantitatively evaluate competing methods for extracting parallel sentences from comparable monolingual corpora, so as to give an overview on the state of the art and to identify the best performing approaches. This repetition of the same task with updated data aims to give new participants another opportunity to address this task and former participants an opportunity to test improved versions of their methods and systems.

Any submission to the shared task is expected to be accompanied by a short paper (4 pages plus references). This will be accepted for publication in the workshop proceedings after a basic quality check: hence the submission will go via Softconf with the standard peer-review process.

The training data for this task is the same as in the 2017 shared task.

Schedule

Shared task **sample and training sets released**	22 December 2017
Shared task **test set released**	24 January 2018
Shared task test submission deadline	31 January 2018
Shared task paper submission deadline	2 February 2018
Shared task camera ready papers	25 February 2018

Task definition

The task definition is the same as in the 2017 shared task: Given two sentence-split monolingual corpora, participant systems are expected to identify pairs of sentences that are translations of each other.

Evaluation will be performed using balanced F-score. In the results of a system, a true positive TP is a pair of sentences that is present in the gold standard and a false positive FP is a pair of sentences that is not present in the gold standard. A false negative FN is a pair of sentences present in the gold standard but absent from system results. Precision P, recall R and F1-score $F1$ are then computed as:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad F1 = \frac{2 \times P \times R}{P+R}$$

Shared task data contents

Sample, training and test data provide monolingual corpora split into sentences, with the following format (utf-8 text, with Unix end-of-lines; identifiers are made of a two-letter language code + 9 digits, separated by a dash '-'):

- Monolingual EN corpus (where EN stands for English), one tab-separated sentence_id + sentence per line
- Monolingual FR corpus (where FR stands for Foreign, e.g. French), one tab-separated sentence_id + sentence per line
- Gold standard list of tab-separated EN-FR sentence_id pairs (held out for the test data)

Sample and training datasets are the same as in the 2017 shared task. They are provided for French-English, German-English, Russian-English, and Chinese-English (see links below).

Important information and requirements:

- The papers that describe the data preparation process (BUCC 2016, BUCC 2017) transparently explain that the data come from two sources: Wikipedia (now 20161201 dumps from December 2016) and News Commentary (now version 11).
- As a consequence, the use of Wikipedia and of News Commentary (other than what is distributed in the present shared task corpora) is not allowed in this task, because they trivially contain the solutions (the latter in a positive way, and the former in a negative way).

Sample data

Sample data is provided for the following language pairs (note that the monolingual English data vary in each language pair):

- de-en (German-English)
- fr-en (French-English)
- ru-en (Russian-English)
- zh-en (Chinese-English)

Each sample dataset contains two monolingual corpora of about 10–70k sentences including 200–2,300 parallel sentences and is provided as a .tar.bz2 archive (1–4MB).

Training and **test data**

Training and **test data** are provided for the following language pairs (note that the monolingual English data vary in each language pair):

- de-en (German-English)
- fr-en (French-English)
- ru-en (Russian-English)
- zh-en (Chinese-English)
- download training data
- download test data

Each training or **test dataset** contains two monolingual corpora of about 100–550k sentences including 2,000–14,000 parallel sentences and is provided as a .tar.bz2 archive (6–36MB). Training data includes gold standard links, test data will not.

Submission details

Each team is allowed to submit up to three (3) runs for each language. In other words, a team can test several methods or parameter settings and submit the three they prefer.

Please structure your test results as follows:

- one file per language, named <team><N>.<fr>-en.test, where
  - <team> stands for you team name (please use only ASCII letters, digits and “-” or “_”)
  - <N> (*1, 2* or *3*) is the run number
  - <fr> stands for the language (among *de, fr, ru, zh*)
- the file contents and format should be the same as the gold standard files provided with the sample and training data, and contain only those sentence pairs that the system believes are translation pairs:

  - One sentence_id pair per line, tab-separated, of the form <fr>-<id1><tab><en>-<id2> where <fr> is one of *de, fr, ru, zh* and <fr>-<id1> and en-<id2> are 9-digit identifiers found in the <fr> and en parts of the test corpus. For instance, for de-en (German-English):

    de-000000003<tab>en-000007818
    de-000000004<tab>en-000013032
    ...

- put all files in one directory called <team>

- create an archive with the contents of this directory (either <team>.tar.bz2, <team>.tar.gz, or <team>.zip)

Send the archive as an attachment in a message together with factual summary information on your team and method:

To: bucc2018st-submission@limsi.fr
Subject: <team> submission

Team name: <team>
Number of runs submitted: <1,2,3>
Participants:
<person1> <email> <affiliation> <country>
<person2> <email> <affiliation> <country>
...
Resources used: <dictionary X>, <corpus Y>, ...
Tools used: <POS tagger X>, <IR system Y>, <word alignment system Z>, <machine learning library T>, ...

You will receive a human acknowledgment in a maximum of 8 hours (depending on the difference between your time zone and CEST).