BUCC, 10th Workshop on Building and Using Comparable Corpora

Co-located with ACL 2017
Vancouver (Canada)
August 3rd, 2017
https://comparable.limsi.fr/bucc2017/
Program: sessions with full papers

Invited Speaker: Philippe Langlais

Shared task: parallel sentence extraction from comparable corpora
Sample, training and test data are available for four language pairs

INVITED SPEAKER

Philippe Langlais

Université de Montréal

Users and Data: The Two Neglected Children of Bilingual Natural Language Processing Research

Despite numerous studies devoted to mining parallel material from bilingual data, we have yet to see the resulting technologies wholeheartedly adopted by professional translators and terminologists alike. I argue that this state of affairs is mainly due to two factors: the emphasis published authors put on models (even though data is as important), and the conspicuous lack of concern for actual end-users.

Philippe Langlais is full professor within the computer science department (DIRO) at University of Montreal (UdeM) in the area of computational linguistics.

MOTIVATION

In the language engineering and the linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, it is chiefly motivated by the need to use comparable corpora as training data for statistical NLP 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 intra-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

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
- 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

IMPORTANT DATES

27 April 2017	Deadline for submission of full papers
19 May 2017	Notification to authors
26 May 2017	Camera-ready papers due
3 August 2017	Workshop date

SUBMISSION INFORMATION

Papers should follow the ACL main conference formatting details (see the ACL conference website http://acl2017.org/calls/papers/) and should be submitted as a PDF-file via the START workshop manager at https://www.softconf.com/acl2017/bucc/.

Contributions can be short or long papers. Short paper submission must describe original and unpublished work without exceeding four (4) pages of content, plus unlimited references. 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 eight (8) pages of content, 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.

For further information, please contact Serge Sharoff <S (dot) Sharoff (at) leeds (dot) ac (dot) uk>

Plain-text CFP : bucc2017-cfp.txt
PDF CFP : bucc2017-cfp.pdf
Last modified: 17 May 2017

ORGANISERS

Serge Sharoff University of Leeds (UK), Chair

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

Reinhard Rapp University of Mainz (Germany)

SCIENTIFIC COMMITTEE

Ahmet Aker (University of Sheffield, UK)
Marianna Apidianaki (LIMSI, CNRS, Université Paris-Saclay, France)
Caroline Barrière (CRIM, Montréal, Canada)
Kurt Eberle (Lingenio, Heidelberg, Germany)
Andreas Eisele (European Commission, Luxembourg)
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)
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)
Ted Pedersen (University of Minnesota, Duluth, US)
Reinhard Rapp (University of Mainz, Germany)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Richard Sproat (Google, US)
Tim Van de Cruys (IRIT-CNRS, Toulouse, France)
Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France)

SHARED TASK

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

We announce a new shared task for 2017. 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.

Shared task **sample set released**	6 February, 2017
Shared task **training set released**	20 February, 2017
(Chinese **training set released**)	3 March 2017
Shared task **test set released**	21 April, 2017
Shared task **test submission deadline**	28 April, 2017
Shared task **paper submission deadline**	2 May, 2017
Shared task camera ready papers	26 May, 2017

Any submission to the shared task is expected to be followed by a short paper (4 pages plus references) describing the methods and resources used to perform the task. This will be accepted for publication in the workshop proceedings automatically, although the submission will go via Softconf with the standard peer-review process.

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)

Datasets are provided for French-English, German-English, Russian-English, and Chinese-English (see links below).

Important information and requirements:

- The paper that describes the data preparation process (BUCC 2016) transparently explains that the data come from two sources: Wikipedia (now 20161201 dumps from December 2016) and News Commentary (now version 11). The details of corpus preparation have changed since the paper, but its overall principles remain the same.
- 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 does not.

Task definition

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:

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

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: bucc2017st-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).