BUCC, 8th Workshop on Building and Using Comparable Corpora

Co-located with ACL 2015
Beijing (China)
30 July 2015

http://comparable.limsi.fr/bucc2015/
Shared Task:
Description
Invited speaker: **Benjamin K. Tsou**
City University of Hong Kong
Augmented Comparable Corpora and Monitoring Corpus in Chinese: LIVAC and SKETCH search engine compared
Proceedings:
http://www.aclweb.org/anthology/W/W15#3400
Full volume:
http://www.aclweb.org/anthology/W/W15/W15-34.pdf

INVITED SPEAKER

Benjamin K. Tsou

City University of Hong Kong
The Chinese University of Hong Kong
Hong Kong University of Science and Technology

Augmented Comparable Corpora and Monitoring Corpus in Chinese: LIVAC and SKETCH search engine compared

The increasing availability of numerous corpora has significantly contributed to the understanding of words in terms of their underlying semantic structures and lexical networks (e.g. COBUILD, WordNet etc.). Through data mining and information retrieval, research in this area has vastly expanded our appreciation that what constitutes lexical knowledge goes beyond synonymy, hyponymy, metonymy, meronymy, grammatical and other collocations. Furthermore, they are fundamental to a universalistic conceptual base of ontologies and knowledge representation which are often enriched by deeper and newer analysis. In this context, each language foregrounds specific features or nodes within this knowledge base by usually non-uniform means.

At the same time, the arrival of the age of Big Data has attracted extensive studies on the actual and dynamic use of language as contextualized (ala. Jakobson 1960) within a given society, especially through the mass media. What are foregrounded in this medium tend to have graded cognitive saliency characterizing members of the common speech community, and such shared knowledge is usually at great variance with the thesaurus approach and show noticeable localized features. It is proposed here

that the two kinds of knowledge (thesauric vs cognitive-cultural) complement each other in human cognition, and are integral to it.

We draw on two large Chinese media databases Sketch (2.1 billion character tokens[1]) and LIVAC (550 million character tokens[2]) for illustration and discussion. The Sketch Engine in Chinese shows how *apple* is, as expected, primarily related to *orange*, *peach*, *fruit*, *vegetable*, *food* etc. At the same time three sub-corpora of LIVAC we draw on show that *apple* has a different set of saliency linkage with *computer*, *iPhone*, *Jobs*, *roll out*, *share price*, *company* etc. This linkage is related less to the universalistic semantic network for *apple*, than to the foregrounded awareness of *apple* as a cultural artifact in actual human social interaction and encoded as social knowledge (Park 1955, Longino 1990). We also show and examine how the salient information associated with *apple* varies across the three major Chinese speech communities: Beijing, Hong Kong and Taipei, reflecting social and societal differences, and regional developments, as well as variations over time. Similarly *free-freedom* in Chinese varies in associated saliency linkage in the three speech communities in interesting ways but also contrasts with the Sketch Engine results.

The above comparison in LIVAC is made possible by rigorous improvement to the common and simplistic approach to the cultivation and use of databases. The augmentation efforts included the rigorous cultivation of 3 comparable (sub-) corpora for Beijing, Hong Kong and Taipei through geographical (*horizontal*), chronological (*vertical*) and domain (*topical*) partitioning of what is often assumed to be a common linguistic database. This partitioning required well-reasoned pre-conceived criteria to ensure adequate equivalency in comparability in terms of size, period and depth of analysis.

To facilitate comparison we propose a Cognitive-cultural Salience Index (CSI) which draws on comparable corpus data (e.g. LIVAC) to provide comparison of the relative saliency of target words in the relevant corpus and presented as word clouds. The results are viewed in the light of the Sketch Engine output to explore how our appreciation of knowledge representation may be enhanced. It will also serve to echo the call to optimize our data collection efforts and to broaden our queries with data judiciously curated and cultivated.

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.

Parallel corpora are a key resource as training data for statistical machine translation, and for building or extending bilingual lexicons and terminologies. However, beyond a few language pairs such as English-French or English-Chinese and a few contexts such as parliamentary debates or legal texts, they remain a scarce resource, despite the creation of automated methods to collect parallel corpora from the Web. Interest in non-parallel forms of comparable corpora in language engineering primarily ensued from the scarcity of parallel corpora. This has motivated research concerning the use of comparable corpora: pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. Non-parallel yet comparable corpora overcome the two limitations of parallel corpora, since sources for original, monolingual texts are much more abundant than translated texts. However, because of their nature, mining translations in comparable corpora is much more challenging than in parallel corpora. What constitutes a good comparable corpus, for a given task or per se, also requires specific attention: while the definition of a parallel corpus is fairly straightforward, building a non-parallel corpus requires control over the selection of source texts in

[1] As per Sketch Engine website.

[2] As per LIVAC website.

both languages.

Research on comparable corpora spans a number of topics from machine translation to contrastive linguistics. Distributional analysis, a topic which has seen renewed interest in recent years, has formed the core of a large part of the methods used to identify translations in comparable corpora. As a matter of fact, the standard techniques of word alignment in comparable corpora can be seen as methods for cross-language distributional semantics.

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

Note that an edited book “Building and Using Comparable Corpora” has recently been published by Springer. Chapter 1, an introduction and state of the art on the topic, is now freely available on Springer’s Web site: Overviewing Important Aspects of the Last 20 Years of Research in Comparable Corpora (click on "Download Sample pages 1 (pdf, 346 kB)").

IMPORTANT DATES

15 May 2015	Deadline for submission of full papers
4 June 2015	Notification of acceptance
21 June 2015	Camera-ready papers due
30 July 2015	Workshop date

SUBMISSION INFORMATION

Submissions should follow the ACL 2015 length and formatting requirements found at http://acl2015.org/call_for_papers.html: long papers can have a maximum of eight (8) pages of content plus two (2) extra pages for references, while short papers can have a maximum of four (4) pages of content plus two (2) extra pages for references. They should be submitted as PDF documents to the following address:

https://www.softconf.com/acl2015/BUCC/

Papers will be blind reviewed by at least two members of the Program Committee. Therefore, authors' names and affiliations should not appear in the paper. Accepted papers will be published in the workshop proceedings.

Authors may submit the same paper at several meetings, but a paper published at this workshop cannot also be published elsewhere. In case of double submission, the authors must notify the workshop organizers in a separate e-mail, so we know that the paper might be withdrawn depending on the results at some other meeting. However, after notification authors will be asked to make a final decision.

For further information, please contact Pierre Zweigenbaum pz(erase_at)limsi(erase_dot)fr

Plain-text CFP : bucc2015-cfp.txt
PDF CFP : bucc2015-cfp.pdf
Last modified: 2 August 2015

ORGANISERS

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

Serge Sharoff University of Leeds (UK), Shared Task Chair

Reinhard Rapp University of Mainz (Germany)

SCIENTIFIC COMMITTEE

Ahmet Aker, University of Sheffield (UK)
Srinivas Bangalore (AT&T Labs, US)
Caroline Barrière (CRIM, Montréal, Canada)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Kurt Eberle (Lingenio, Heidelberg, Germany)
Andreas Eisele (European Commission, Luxembourg)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (INRIA, Saclay, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology)
Kyo Kageura (University of Tokyo, Japan)
Adam Kilgarriff (Lexical Computing Ltd, UK)
Natalie Kübler (Université Paris Diderot, France)
Philippe Langlais (Université de Montréal, Canada)
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)shared.bucc2015@gmail.com
Sujith Ravi (Google, US)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Tim Van de Cruys (IRIT-CNRS, Toulouse, France)
Stephan Vogel, QCRI (Qatar)
Guillaume Wisniewski (Université Paris Sud & LIMSI-CNRS, Orsay, France)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

SHARED TASK

A shared task is organized together with the workshop. This will be the first evaluation exercise on the identification of comparable texts: given a large multilingual collection of texts (we will be using Wikipedia documents in several languages), the task is to identify the most similar texts across languages. Evaluation will be done by measuring precision, recall and F-measure on links between pages, with a gold standard based on actual inter-language links.

Task description

Parallel corpora of original texts with their translations provide the basis for multilingual NLP applications since the beginning of the 1990s. Relative scarcity of such resources led to greater attention to comparable (=less parallel) resources to mine information about possible translations. Many studies have been produced within the paradigm of comparable corpora, including publications in the BUCC workshop series since 2008, see bucc-introduction.html.

However, the community so far has not conducted an evaluation which compared different approaches for identifying more or less parallel resources in a large amount of multilingual data. Also, it is not clear how language-specific such approaches are. In this shared task we propose the first evaluation exercise, which is aimed at detecting the most similar texts in a large collection.

Data set

The data for each language pair has been split into two sets:

Training set pages with information about the correct links for the respective language pairs;

Test set pages without the links.

The task is for each page in the test set to submit up to five ranked suggestions to its linked page, assuming that the gold standard contains its counterpart in another language. The submissions will have to be in the tab-separated format as used in the submissions to TREC with six fields:
`id1 X id2 Y score run.name`

The X and Y fields are not used, but they are reserved by the TREC evaluation script (and it does not use them either). Please keep them with constant values X and Y. `id1` and `id2` are the articles ids in a language of evaluation and in English. The score should reflect the similarity between `id1` and `id2`, the higher the closer. The participants are invited to submit up to five runs of their system with different parameters, as identified by a keyword in the last field. This field should include the name of the team and an identifier for the run, e.g., Leeds.run1, or LIMSI.BM25. For the evaluation script and for more information about the format, please visit: http://trec.nist.gov/trec_eval/

The languages in the shared task will be Chinese, French, German, Russian and Turkish. Pages in these languages need to be linked to a page in English.

Submission procedure

Please register by sending a message to shared.bucc2015@gmail.com and giving the name of the contact person, and the language pairs you'd like to work on.

In response you will receive links to the training sets and the scoring script.

Deadlines

1 February 2015	Training set available
20 April 2015	Test set available
24 April 2015	Test submission deadline
1 May 2015	System results to participants
15 May 2015	Paper submission deadline
4 June 2015	Notification of acceptance
21 June 2015	Camera-ready papers due
30 July 2015	Workshop date