Special Issue on "Machine Translation Using Comparable Corpora"

Journal of Natural Language Engineering

Deadline for paper submission: closed
http://comparable.limsi.fr/jnle-bucc2015/
Submission site:
http://mc.manuscriptcentral.com/nle

MOTIVATION

Statistical machine translation based on parallel corpora has been very successful. The major search engines' translation systems, which are used by millions of people, are primarily using this approach, and it has been possible to come up with new language pairs in a fraction of the time that would be required when using more traditional rule-based methods.

In contrast, research on comparable corpora is still at an earlier stage. Comparable corpora can be defined as monolingual corpora covering roughly the same subject area in different languages but without being exact translations of each other.

However, despite its tremendous success, the use of parallel corpora in MT has a number of drawbacks:

1. It has been shown that translated language is somewhat different from original language, for example the "associative texture" is lost in translation.

2. As they require translation, parallel corpora will always be a far scarcer resource than comparable corpora. This is a severe drawback for a number of reasons:

   (a) Among the about 7000 world languages, of which 600 have a written form, the vast majority are of the "low resource" type.

   (b) The number of possible language pairs increases with the square of the number of languages. When using parallel corpora, one bitext is needed for each language pair. When using comparable corpora, one monolingual corpus per language suffices.

   (c) For improved translation quality, translation systems specialized on particular genres and domains are desirable. But it is far more difficult to acquire appropriate parallel rather than comparable training corpora.

   (d) As language evolves over time, the training corpora should be updated on a regular basis. Again, this is more difficult in the parallel case.

For such reasons it would be a big step forward if it were possible to base statistical machine translation on comparable rather than on parallel corpora: The acquisition of training data would be far easier, and the unnatural "translation bias" (source language shining through) within the training data would be avoided.

But is there any evidence that this is possible? Motivation for using comparable corpora in MT research comes from a cognitive perspective: Experience tells that persons who have learned a second language completely independently from their mother tongue can nevertheless translate between the languages. That is, human performance shows that there must be a way to bridge the gap between languages which does not rely on parallel data. Using parallel data for MT is of course a nice shortcut. But avoiding this shortcut by doing MT based on comparable corpora may well be a key to a better understanding of language, and to better MT quality.

Work on comparable corpora in the context of MT has been ongoing for almost 20 years. It has turned out that this is a very hard problem to solve, but as it is among the grand challenges in multilingual NLP, interest has steadily increased. Apart from the increase in publications this can be seen from the considerable number of research projects (such as ACCURAT and TTC) which are fully or partially devoted to MT using comparable corpora. Given also the success of the workshop series on Building and Using Comparable Corpora (BUCC), which is now in its seventh year, and following the publication of a related book (http://www.springer.com/computer/ai/book/978-3-642-20127-1), we think that it is now time to devote a journal special issue to the field. It is meant to bundle the latest top class research, make it available to everybody working in the field, and at the same time give an overview on the state of the art to all interested researchers.

TOPICS

We solicit contributions including but not limited to the following topics:

- Comparable-corpora-based MT systems (CCMTs)
- Architectures for CCMTs
- CCMTs for less-resourced languages
- CCMTs for less-resourced domains
- CCMTs dealing with morphologically rich languages
- CCMTs for spoken translation
- Applications of CCMTs
- CCMT evaluation
- Open source CCMT systems
- Hybrid systems combining SMT and CCMT
- Hybrid systems combining rule-based MT and CCMT
- Enhancing phrase-based SMT using comparable corpora
- Expanding phrase tables using comparable corpora
- Comparable-corpora-based processing tools/kits for MT
- Methods for mining comparable corpora from the Web
- Applying Harris' distributional hypothesis to comparable corpora
- Induction of morphological, grammatical, and translation rules from comparable corpora
- Machine learning techniques using comparable corpora
- Parallel corpora vs. pairs of non-parallel monolingual 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

IMPORTANT DATES

15 December 2014	Paper submission deadline (extended, closed)
1 February 2015	Notification (done)
1 May 2015	Deadline for revised papers
1 July 2015	Final notification
1 September 2015	Final paper due

GUEST EDITORS

Reinhard Rapp University of Mainz (Germany)

Serge Sharoff University of Leeds (UK)

Pierre Zweigenbaum LIMSI, CNRS, Orsay (France)

SUBMISSION INFORMATION

Manuscripts, submitted as PDF files, will be processed via the JNLE site at http://mc.manuscriptcentral.com/nle, according to the instructions in http://assets.cambridge.org/NLE/NLE_ifc.pdf.

Register as an author and select the manuscript type "Special Issue: Machine Translation Using Comparable Corpora". Further details will appear in this space closer to the submission deadline.

The recommended format is LaTeX, using the JNLE style files, a copy of which is here.

Please use the following e-mail address if you need to contact the guest editors: mailto:jnle.bucc(erase_at)limsi(erase_dot)fr

Plain-text CFP : jnle-bucc2015-cfp.txt
PDF CFP : jnle-bucc2015-cfp.pdf
Last modified: 8 Feb 2015

GUEST EDITORIAL BOARD

Akiko Aizawa (University of Tokyo, NII, Japan)

Ahmet Aker (University of Sheffield, UK)

Marianna Apidianaki (LIMSI, CNRS, Orsay, France)

Núria Bel (Universitat Pompeu Fabra, Barcelona, Spain)

Helena Blancafort (Syllabs, Paris, France)

Dhouha Bouamor (LIMSI, CNRS, Orsay, France)

Chenhui Chu (Kyoto University, Japan)

Kenneth W. Church (IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA)

Béatrice Daille (Université de Nantes, France)

Estelle Delpech (Nomao Labs, Paris, France)

Silvia Hansen-Schirra (Universität Mainz, Germany)

Amir Hazem (Université de Nantes, France)

Diana Inkpen (University of Ottawa, Canada)

Kyo Kageura (University of Tokyo, Japan)

Kevin Knight (University of Southern California, ISI, Marina del Rey, CA, USA)

Philipp Koehn (Johns Hopkins University, Baltimore, MD, USA)

Bo Li (Central China Normal University, Wuhan, China)

Belinda Maia (University of Porto, Portugal)

Maria Teresa Martin-Valdivia (Universidad de Jaén, Spain)

Tomas Mikolov (Facebook, Menlo Park, CA, USA)

Emmanuel Morin (Université de Nantes, France)

Santanu Pal (Saarland University, Saarbrücken, Germany)

Uwe Quasthoff (Universität Leipzig, Germany)

Reinhard Rapp (Universität Mainz, Germany)

Iñaki San Vicente Roncal (Elhuyar Fundazioa R&D, Usurbil, Spain)

Xabier Saralegi (Elhuyar Fundazioa R&D, Usurbil, Spain)

Carolina Scarton (University of Sheffield, UK)

Serge Sharoff (University of Leeds, UK)

Inguna Skadiņa (Tilde and Liepāja University, Latvia)

Kamel Smaili (Université de Lorraine, Vandoeuvre-lès-Nancy, France)

Marko Tadić (University of Zagreb, Croatia)

George Tambouratzis (Institute for Language and Speech Processing, Athens, Greece)

Benjamin Tsou (The Hong Kong Institute of Education, China)

Stephan Vogel (Qatar Computing Research Institute, Doha, Qatar)

Yorick Wilks (Florida Institute of Human and Machine Cognition, Ocala, FL, USA)

Geoffrey Williams (Université de Bretagne-Sud, Lorient, France)

Krzysztof Wołk (Polish-Japanese Institute of Information Technology, Warsaw, Poland)

Chengzhi Zhang (Nanjing University of Science and Technology, China)

Pierre Zweigenbaum (LIMSI, CNRS, Orsay, France)