

4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web 2011

(BUCC 2011)

Held at ACL HLT 2011

**Portland, Oregon, USA
24 June 2011**

ISBN: 978-1-61839-249-7

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2011) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2011)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2634
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Invited Presentation

<i>Putting a Value on Comparable Data</i>	
Kevin Knight	1
<i>The Copiale Cipher</i>	
Kevin Knight, Beáta Megyesi and Christiane Schaefer	2

Oral Presentations

<i>Learning the Optimal Use of Dependency-parsing Information for Finding Translations with Comparable Corpora</i>	
Daniel Andrade, Takuya Matsuzaki and Junichi Tsujii	10
<i>Building and Using Comparable Corpora for Domain-Specific Bilingual Lexicon Extraction</i>	
Darja Fišer, Nikola Ljubešić, Špela Vintar and Senja Pollak	19
<i>Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora</i>	
Emmanuel Morin and Emmanuel Prochasson	27
<i>Bilingual Lexicon Extraction from Comparable Corpora as Metasearch</i>	
Amir Hazem, Emmanuel Morin and Sebastian Peña Saldarriaga	35
<i>Two Ways to Use a Noisy Parallel News Corpus for Improving Statistical Machine Translation</i>	
Souhir Gahbiche-Braham, Hélène Bonneau-Maynard and François Yvon	44
<i>Paraphrase Fragment Extraction from Monolingual Comparable Corpora</i>	
Rui Wang and Chris Callison-Burch	52
<i>Extracting Parallel Phrases from Comparable Data</i>	
Sanjika Hewavitharana and Stephan Vogel	61
<i>Active Learning with Multiple Annotations for Comparable Data Classification Task</i>	
Vamshi Ambati, Sanjika Hewavitharana, Stephan Vogel and Jaime Carbonell	69
<i>How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives</i>	
Bruno Cartoni, Sandrine Zufferey, Thomas Meyer and Andrei Popescu-Belis	78
<i>Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia.</i>	
Alexandre Patry and Philippe Langlais	87
<i>Comparable Fora</i>	
Johanka Spoustová and Miroslav Spousta	96

Poster Presentations

<i>Unsupervised Alignment of Comparable Data and Text Resources</i>	
Anja Belz and Eric Kow	102
<i>Cross-lingual Slot Filling from Comparable Corpora</i>	
Matthew Snover, Xiang Li, Wen-Pin Lin, Zheng Chen, Suzanne Tamang, Mingmin Ge, Adam Lee, Qi Li, Hao Li, Sam Anzaroot and Heng Ji	110
<i>Towards a Data Model for the Universal Corpus</i>	
Steven Abney and Steven Bird	120
<i>An Expectation Maximization Algorithm for Textual Unit Alignment</i>	
Radu Ion, Alexandru Ceașu and Elena Irimia	128
<i>Building a Web-Based Parallel Corpus and Filtering Out Machine-Translated Text</i>	
Alexandra Antonova and Alexey Misyurev	136
<i>Language-Independent Context Aware Query Translation using Wikipedia</i>	
Rohit Bharadwaj G and Vasudeva Varma	145