Coling 2008

22nd International Conference on Computational Linguistics

Proceedings of the

2nd workshop on Multi-source, Multilingual Information Extraction and Summarization

23 August 2008
Manchester, UK

Order copies of this and other Coling proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-905593-51-4

Design by Chimney Design, Brighton, UK
Production and manufacture by One Digital, Brighton, UK

Editors' Foreword

Information extraction (IE) and text summarization (TS) are key technologies aiming at extracting relevant information from texts and presenting the information to the user in a condensed form. The ongoing information explosion makes IE and TS particularly critical for successful functioning within the information society. These technologies, however, face new challenges with the adoption of the Web 2.0 paradigm (e.g., blogs, wikis) due to their inherent multi-source nature. These technologies must no longer deal only with isolated texts or narratives, but with large-scale repositories or sources—possibly in several languages—containing a multiplicity of views, opinions, and commentaries on particular topics, entities and events. There is thus a need to adapt and/or develop new techniques to deal with these new phenomena.

Recognising similar information across different sources and/or in different languages is of paramount importance in this multi-source, multi-lingual context. In information extraction, merging information from multiple sources can lead to increased accuracy, as compared to extraction from a single source. In text summarization, similar facts found across sources can inform sentence scoring algorithms. In question answering, the distribution of answers in similar contexts can inform answer-ranking components. Often, it is not the similarity of information that matters, but its complementary nature. In a multi-lingual context, information extraction and text summarization can provide solutions for cross-lingual access: key pieces of information can be extracted from different texts in one or many languages, merged, and then conveyed in natural language in concise form. Applications need to be able to cope with the idiosyncratic nature of the new Web 2.0 media: mixed input, new jargon, ungrammatical and mixed-language input, emotional discourse, etc. In this context, synthesizing or inferring opinions from multiple sources is a new and exciting challenge for NLP. On another level, profiling of individuals who engage in the new social Web, and identifying whether a particular opinion is appropriate/relevant in a given context are important topics to be addressed.

The objective of this second *Multi-source Multilingual Information Extraction and Summarization* (MMIES) workshop is to bring together researchers and practitioners in information-access technologies, to discuss recent approaches for dealing with multi-source and multi-lingual challenges. Each paper submitted to the workshop was reviewed by three members of an international Programme Committee. The selection process resulted in this volume of eight papers, covering the following key topics:

- Multilingual Named Entity Recognition,
- Automatic Construction of Multilingual Dictionaries for Information Retrieval,
- Multi-document Summaries for Geo-referenced Images,
- Keyword Extraction for Single-Document Summarization,
- Recognizing Similar News over Time and across Languages,
- Speech-to-Text Summarization,
- Automatic Annotation of Bibliographical References.

We are grateful to the members of the programme committee for their invaluable work, as well as to Roger Evans, Mark Stevenson and Harold Somers for their support.

We thank Robert Gaizauskas for giving the invited talk at the workshop.

July 2008.

Sivaji Bandyopadhyay, Jadavpur University (India)
Thierry Poibeau, CNRS / Université Paris 13 (France)
Horacio Saggion, University of Sheffield (UK)
Roman Yangarber, University of Helsinki (Finland)

Organizers

- Sivaji Bandyopadhyay, Jadavpur University (India)
- Thierry Poibeau, CNRS and University of Paris 13 (France)
- Horacio Saggion, University of Sheffield (United Kingdom)
- Roman Yangarber, University of Helsinki (Finland)

Programme Committee

- Javier Artiles, UNED (Spain)
- Kalina Bontcheva, University of Sheffield (UK)
- Nathalie Colineau, CSIRO (Australia)
- Nigel Collier, NII (Japan)
- Hercules Dalianis, KTH/Stockholm University (Sweden)
- Thierry Declerk, DFKI (Germany)
- Michel Généreux, LIPN-CNRS (France)
- Julio Gonzalo, UNED (Spain)
- Brigitte Grau, LIMSI-CNRS (France)
- Ralph Grishman, New York University (USA)
- Kentaro Inui, NAIST (Japan)
- Min-Yen Kan, National University of Singapore (Singapore)
- Guy Lapalme, University of Montreal (Canada)
- Diana Maynard, University of Sheffield (UK)
- Jean-Luc Minel, Modyco-CNRS (France)
- Constantin Orasan, University of Wolverhampton (UK)
- Cecile Paris, CSIRO (Australia)
- Maria Teresa Pazienza, University of Roma 'Tor Vergata' (Italy)
- Bruno Pouliquen, European Commission – Joint Research Centre (Italy)
- Patrick Saint-Dizier, IRIT-CNRS (France)

- Agnes Sandor, Xerox XRCE (France)
- Satoshi Sekine, NYU (USA)
- Ralf Steinberger, European Commission – Joint Research Centre (Italy)
- Stan Szpakowicz, University of Ottawa (Canada)
- Lucy Vanderwende, Microsoft Research (USA)
- José Luis Vicedo, Universidad de Alicante (Spain)

Table of Contents

Conference Programme

Wednesday, August 23, 2008

Invited Talk

9:30–10:30 *Generating Image Captions using Topic Focused Multi-document Summarization*
Robert Gaizauskas

10:30–11:00 Coffee break

Session 1: Named Entity and Lexical Resources for IE and Summarization

11:00–11:30 *Learning to Match Names Across Languages*
Inderjeet Mani, Alex Yeh and Sherri Condon

11:30–12:00 *Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic languages*
Sumithra Velupillai and Hercules Dalianis

12:00–12:30 *Graph-Based Keyword Extraction for Single-Document Summarization*
Marina Litvak and Mark Last

12:30–14:00 Lunch

Session 2: Multi-document Summarization

14:00–14:30 *MultiSum: Query-Based Multi-Document Summarization*
Mike Rosner and Carl Camilleri

14:30–15:00 *Mixed-Source Multi-Document Speech-to-Text Summarization*
Ricardo Ribeiro and David Martins de Matos

15:00–15:30 *Evaluating automatically generated user-focused multi-document summaries for geo-referenced images*
Ahmet Aker and Robert Gaizauskas

15:30–16:00 Coffee break

Wednesday, August 23, 2008 (continued)

Session 3: Applications

16:00–16:30 *Story tracking: linking similar news over time and across languages*
Bruno Pouliquen, Ralf Steinberger and Olivier Deguernel

16:30–17:00 *Automatic Annotation of Bibliographical References with target Language*
Harald Hammarström

17:00–17:30 Open Discussion