INTERNATIONAL WORKSHOP

NATURAL LANGUAGE PROCESSING METHODS AND CORPORA IN TRANSLATION, LEXICOGRAPHY, AND LANGUAGE LEARNING

held in conjunction with the International Conference

RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria

PROCEEDINGS

Edited by
Iustina Ilisei, Viktor Pekar and Silvia Bernardini

Borovets, Bulgaria

17 September 2009

International Workshop

NATURAL LANGUAGE PROCESSING METHODS AND CORPORA IN TRANSLATION, LEXICOGRAPHY, AND LANGUAGE LEARNING

PROCEEDINGS

Borovets, Bulgaria
17 September 2009

ISBN 978-954-452-010-6

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Foreword

In recent years corpora have become an indispensable tool in research and everyday practice for translators, lexicographers, second language learners. Specialists in these areas share a general goal in using corpora in their work: corpora provide the possibility of finding and analysing linguistic patterns characteristic of various kinds of language users, monitoring language change, and revealing important similarities and divergences across different languages.
By this time, Natural Language Processing (NLP) technologies have matured to the point where much more complex analysis of corpora becomes possible: more complex grammatical and lexical patterns can be discovered, and new, more complex aspects of text (pragmatic, stylistic, etc.) can be analysed computationally.
For professional translators, corpora represent an invaluable linguistic and cultural awareness tool. For language learners, they serve as a means to gain insights into specifics of competent language use as well as to analyse typical errors of fellow learners. For lexicographers, corpora are key for monitoring the development of language vocabularies, making informed decisions as to lexicographic relevance of the lexical material, and for general verification of all varieties of lexicographic data.

While simple corpus analysis tools such as concordancers have long been in use in these specialist areas, in the past decade there have been important developments in Natural Language Processing technologies: it has become much easier to construct corpora, and powerful NLP methods have become available that can be used to analyse corpora not only at the surface level, but also at the syntactic, and even semantic, pragmatic, and stylistic levels.

We believe that 2009 was an appropriate moment for the RANLP workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning. It presented recent studies covering the following topics: term and collocation extraction, corpora in translator training, construction of lexical resource, lexical substitution techniques, word alignment, and automatic tree alignment. The event was complemented by two invited speakers who presented several studies where NLP methods and corpora have proved to be helpful.
The workshop brought together the developers and the users of NLP technologies for the purposes of translation, translation studies, lexicography, terminology, and language learning in order to present their research and discuss new possibilities and challenges in these research areas.
We are grateful to the organisers of the Seventh International Conference on Recent Advances in Natural Language Processing, RANLP 2009, for holding this workshop in conjunction to the main conference. We are also thankful to the Programme Committee for their commitment and support in the reviewing process and to the researchers who submitted papers to this workshop.

Iustina Ilisei, Viktor Pekar, and Silvia Bernardini

17th September, 2009

Programme Committee

Marco Baroni, University of Trento
Jill Burstein, Educational Testing Service
Michael Carl, Copenhagen Business School
Gloria Corpas Pastor, University of Málaga
Le An Ha, University of Wolverhampton
Patrick Hanks, Masaryk University
Federico Gaspari, University of Bologna
Adam Kilgarriff, Lexical Computing
Marie-Claude L'Homme, Université de Montréal
Ruslan Mitkov, University of Wolverhampton
Roberto Navigli, University of Rome "La Sapienza"
Miriam Seghiri, University of Málaga
Pete Whitelock, Oxford University Press
Richard Xiao, Edge Hill University
Federico Zanettin, University of Perugia

Organising Committee

Iustina Ilisei, University of Wolverhampton, United Kingdom
Viktor Pekar, Oxford University Press, United Kingdom
Silvia Bernardini, University of Bologna, Italy

Table of Contents

Conference Program

Thursday, September 17, 2009

10:30 - 11:00 Finding Domain Specific Collocations and Concordances on the Web
Caroline Barrière

11:00 - 11:30 HMMs, GRs, and N-Grams as Lexical Substitution Techniques – Are They Portable to Other Languages?
Judita Preiss, Andrew Coonce and Brittany Baker

11:30 - 13:00 The Web a Corpus: Going Beyond Page Hit Frequencies
Preslav Nakov (invited talk)

13:00 - 14:00 Lunch Break

14:00 - 14:30 Unsupervised Construction of a Multilingual WordNet from Parallel Corpora
Dimitar Kazakov and Ahmad R. Shahid

14:30 - 15:00 Search Techniques in Corpora for the Training of Translators
Verónica Pastor and Amparo Alcina

15:00 - 16:00 Translation Universals: Experiments on Simplification, Convergence and Transfer
Gloria Corpas and Ruslan Mitkov (invited talk)

16:00 - 16:30 Break

16:30 - 17:00 Evidence-Based Word Alignment
Jörg Tiedemann

17:00 - 17:30 A Discriminative Approach to Tree Alignment
Jörg Tiedemann and Gideon Kotzé