EACL 2012

# Hybrid 2012
# Innovative Hybrid Approaches to the Processing of Textual Data

## Proceedings of the Workshop

April 23 2012
Avignon France

ISBN 978-1-937284-19-0

Order copies of this and other ACL 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

# Introduction

The *hybrid approach* term covers a large set of situations in which different approaches are combined in order to better process textual data and to attempt a better achievement of the dedicated task. Hybrid approaches are commonly used in various NLP applications (*i.e.*, automatic creation of linguistic resources, POS tagging, building and structuring of terminologies, information retrieval and filtering, linguistic annotation, semantic labelling).

Among the hybridizations the possible combinations are unlimited. The most frequent combination, as stressed during The Balancing Act in 1994, addressed machine learning and rule-based systems. Beyond this, the hybridization can be augmented with distributional approaches, syntactic and morphological analyses, semantic distances and similarities, graph theory models, co-occurrences of linguistic units (e.g., word and their dependencies, word senses and postag, NEs and semantic roles,...), knowledge-based approaches (terminologies and ontologies), etc.

As a matter of fact, the hybridization implies to define a strategy to efficiently combine several approaches: cooperation between approaches, filtering, voting or ranking of the multiple system outputs, etc. Indeed, the combination of these different methods and approaches appears to provide more complete and efficient results. The reason is that each method is sensitive and efficient with given data and within given contexts. Hence, their combination may improve both precision and recall. The coverage is indeed improved, while the exploitation of different methods may also lead to the improvement of the precision since their use within filtering, voting etc. modes becomes possible.

This workshop has several objectives:

- To bring together researchers working on hybrid approaches independently from the topics and the applications. Indeed, the presented papers and posters address a great variety of applications: machine translation, lexicon and semantic relations acquisition, spell checking, indexing and annotation, syntactic analysis, summarization, named entity recognition, question-answering. We hope the exchange experienced during this workshop will be fruitful for the future research and collaborations.
- To outline future directions for the conception of novel hybrid approaches. For instance, the invited speaker Rada Mihalcea, University of North Texas, USA will give a presentation on the multilingual hybridization methods.

The Hybrid 2012 workshop received 27 submissions. Seven of these have been accepted as full papers and eight as poster presentations.

# Acknowledgments

First of all, we are grateful to the authors who chose the Hybrid 2012 workshop to submit and present their innovative work, especially as these authors come from different countries. Without them, this workshop could not have been organized.

We are grateful to the organizers of the EACL conference to have accepted this workshop as one of the joint events of the main conference. This is a really nice place and time to present the research work.

We are very grateful to the members of the Scientific committee who reviewed three submissions in a very short period of time and, for some of them, during the holiday period.

We are particularly grateful to Rada Mihalcea, University of North Texas, USA to have accepted to participate in this workshop and to give the invited speaker talk.

Finally, we are grateful to our labs and institutions for the sponsoring of this workshop.

**Organizers:**

Natalia Grabar, CNRS UMR 8163 STL, Université Lille 1&3, France
Marie Dupuch, CNRS UMR 8163 STL, Université Lille 1&3, France
Amandine Périnet, LIM&BIO, Université Paris 13, France
Thierry Hamon, LIM&BIO, Université Paris 13, France

**Program Committee:**

Anders Ardö, EIT, LTH, Lund University, Sweden
Delphine Bernhard, LiLPa, Université de Strasbourg, France
Wray Duntine, NICTA, Australia
Philipp Cimiano, CITEC, University of Bielefeld, Germany
Vincent Claveau, IRISA-CNRS, Rennes, France
Kevin Cohen, University of Colorado Health Sciences Center, USA
Marie-Claude l'Homme, OLST, Université de Montreal, Canada
Béatrice Daille, Université de Nantes, LINA, France
Stefan Th. Gries, University of California, Santa Barbara, USA
Anna Kazantseva, University of Ottawa, Canada
Mikaela Keller, CNRS LIFL UMR8022, Mostrare INRIA, Université Lille 1&3, France
Alistair Kennedy, University of Ottawa, Canada
Ben Leong, University of North Texas, USA
Pierre Nugues, CS, LTH, Lund University, Sweden
Bruno Pouliquen, WIPO, Geneva, Switzerland
Sampo Pyysalo, National Centre for Text Mining, University of Manchester, United Kingdom
Mathieu Roche, LIRMM, Université de Montpellier 2, France
Patrick Ruch, Haute École de Gestion de Genève, Switzerland
Paul Thompson, National Centre for Text Mining, University of Manchester, United Kingdom
Juan-Manuel Torres Moreno, LIA, Université d'Avignon et des Pays de Vaucluse, France
Özlem Uzuner, University at Albany, State University of New York, USA

**Invited Speaker:**

Rada Mihalcea, University of North Texas, USA

# Table of Contents

# Conference Program

## Monday April 23, 2012

09:00 Introduction

**(09:10) Session 1**

09:10 *Experiments on Hybrid Corpus-Based Sentiment Lexicon Acquisition*
Goran Glavaš, Jan Šnajder and Bojana Dalbelo Bašić

09:40 *A Study of Hybrid Similarity Measures for Semantic Relation Extraction*
Alexander Panchenko and Olga Morozova

10:10 Coffee break

**(10:30) Session 2**

10:30 *Hybrid Combination of Constituency and Dependency Trees into an Ensemble Dependency Parser*
Nathan Green and Zdeněk Žabokrtský

11:00 *Describing Video Contents in Natural Language*
Muhammad Usman Ghani Khan and Yoshihiko Gotoh

11:30 *An Unsupervised and Data-Driven Approach for Spell Checking in Vietnamese OCR-scanned Texts*
Cong Duy Vu Hoang and Ai Ti Aw

**(12:00) Short Presentation: Posters**

12:30 Lunch break

**Monday April 23, 2012 (continued)**

**(14:00) Invited speaker**

14:00 *Multilingual Natural Language Processing*
Rada Mihalcea

**(15:30) Coffee break and Poster Session**

15:30 *Contrasting Objective and Subjective Portuguese Texts from Heterogeneous Sources*
Michel Généreux and William Martinez

*A Joint Named Entity Recognition and Entity Linking System*
Rosa Stern, Benoît Sagot and Frédéric Béchet

*Collaborative Annotation of Dialogue Acts: Application of a New ISO Standard to the Switchboard Corpus*
Alex C. Fang, Harry Bunt, Jing Cao and Xiaoyue Liu

*Coupling Knowledge-Based and Data-Driven Systems for Named Entity Recognition*
Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger and Arnaud Soulet

*A Random Forest System Combination Approach for Error Detection in Digital Dictionaries*
Michael Bloodgood, Peng Ye, Paul Rodrigues, David Zajic and David Doermann

*Methods Combination and ML-based Re-ranking of Multiple Hypothesis for Question-Answering Systems*
Arnaud Grappy, Brigitte Grau and Sophie Rosset

*A Generalised Hybrid Architecture for NLP*
Alistair Willis, Hui Yang and Anne De Roeck

**Monday April 23, 2012 (continued)**

**(16:30) Session 3**

16:30 *Incorporating Linguistic Knowledge in Statistical Machine Translation: Translating Prepositions*
Reshef Shilon, Hanna Fadida and Shuly Wintner

17:00 *Combining Different Summarization Techniques for Legal Text*
Filippo Galgani, Paul Compton and Achim Hoffmann

17:30 Closing