

IEKA 2011

**Proceedings of the Workshop on
Information Extraction and Knowledge
Acquisition**

held in conjunction with
**The 8th International Conference on
Recent Advances in Natural Language Processing
(RANLP 2011)**

16 September, 2011
Hissar, Bulgaria

INTERNATIONAL WORKSHOP
INFORMATION EXTRACTION AND KNOWLEDGE ACQUISITION

PROCEEDINGS

Hissar, Bulgaria
16 September 2011

ISBN 978-954-452-018-2

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

The RANLP 2011 Workshop on *Information Extraction and Knowledge Acquisition* (IEKA 2011) took place on September 16, 2011 in Hissar, Bulgaria, following the conference on Recent Advances in Natural Language Processing (RANLP 2011).

The workshop was envisioned as a meeting place for those concerned with the fields of information extractions (IE) and knowledge acquisition (KA). Until a decade ago, these fields were mostly limited to identifying and extracting named entities, semantic and ontological relations, events, templates, and facts in relatively small text corpora using a small variety of external resources such as gazetteers, thesauri, and lexical hierarchies.

Today everything has changed. The size of corpora has grown dramatically: using Gigaword-scale data is common, and it is almost standard to use the Web, which contains quadrillions of words, or at least the Google Web 1T 5-grams. More importantly, new types of communication have emerged, such as chats, blogs and, in the last 2-3 years, Twitter, whose informal language poses many challenges to automatic IE and KA, yet they are becoming increasingly important, e.g., for learning customer opinions on various products and services. Social network analysis is another emerging topic, where data is naturally much more interconnected than in the rest of the Web.

All these recent developments have posed not only new challenges, but have also created a number of opportunities, opening new research directions, and offering new useful resources. For example, the growth of Wikipedia has given rise to DBpedia and other collaboratively-created resources such as Freebase. Today, IE and KA researchers can even create annotations and resources on demand as they need them for a very low price using crowd-sourcing tools such as Amazon's Mechanical Turk.

We received 12 submissions, and, given our limited capacity as a one-day workshop, we were only able to accept seven full papers for oral presentation: an acceptance rate of 58%. The workshop also featured two invited talks (by Ralph Grishman and by Ralf Steinberger), and a panel discussion (involving Kevin Cohen, Georgi Georgiev, Petya Osenova, Elena Paskaleva, and Kiril Simov).

We would like to thank the members of the Program Committee for their timely reviews. We would also like to thank the authors for their valuable contributions.

Preslav Nakov, Zornitsa Kozareva, Kuzman Ganchev, Jerry Hobbs
Co-Organizers

Organizers:

Preslav Nakov, National University of Singapore
Zornitsa Kozareva, University of Southern California
Kuzman Ganchev, Google Inc.
Jerry Hobbs, University of Southern California

Program Committee:

Javier Artiles, Queens College
Kalina Bontcheva, University of Sheffield
Razvan Bunescu, Ohio State University
Hamish Cunningham, University of Sheffield
Mona Diab, Columbia University
Greg Druck, University of Massachusetts Amherst
Georgi Georgiev, Ontotext AD
Filip Ginter, University of Turku
Jennifer Gillenwater, University of Pennsylvania
Andrew Gordon, USC/ICT
João Graça, University of Pennsylvania
Chikara Hashimoto, NICT
Nitin Indurkha, eBay
Hagen Fürstena, Saarland University
Alex Kulesza, University of Pennsylvania
Maussam, University of Washington
Su Nam Kim, The University of Melbourne
Alexandre Klementiev, Johns Hopkins University
Ana-Maria Popescu, Yahoo!
Ryan McDonald, Google
Andrés Montoyo, University of Alicante
Rafael Muñoz, University of Alicante
Roberto Navigli, University of Rome
Manuel Palomar, University of Alicante
Siddharth Patwardhan, IBM
Marco Pennacchiotti, Yahoo!
Simone Paolo Ponzetto, University of Heidelberg
Sampo Pyysalo, University of Tokyo
Sujith Ravi, ISI, University of Southern California
Ellen Riloff, University of Utah
Alan Ritter, University of Washington
Paolo Rosso, Technical University of Valencia
Kenji Sagae, USC/ICT
Tamar Solorio, University of Alabama at Birmingham
Ralf Steinberger, JRC
Mihai Surdeanu, Stanford University
Idan Szpektor, Yahoo! Research
Valentin Tablan, University of Sheffield
Jörg Tiedemann, University of Uppsala
Ivan Titov, Saarland University
Aline Villavicencio, Federal University of Rio Grande do Sul
Alexander Yates, Temple University

Additional Reviewers:

Gabor Angeli

Invited Speakers:

Ralph Grishman, New York University

Ralf Steinberger, European Commission Joint Research Centre

Panelists:

Kevin Cohen, University of Colorado, School of Medicine

Georgi Georgiev, Ontotext AD

Petya Osenova, Bulgarian Academy of Sciences

Elena Paskaleva, Bulgarian Academy of Sciences

Kiril Simov, Bulgarian Academy of Sciences

Table of Contents

| | |
|---|----|
| <i>INVITED TALK 1: The Knowledge Base Population Task: Challenges for Information Extraction</i> | |
| Ralph Grishman | 1 |
| <i>Fine-grained Entity Set Refinement with User Feedback</i> | |
| Bonan Min and Ralph Grishman | 2 |
| <i>Extraction of Domain-specific Opinion Words for Similar Domains</i> | |
| Iliia Chetviorkin and Natalia Loukachevitch | 7 |
| <i>The Role of Predicates in Opinion Holder Extraction</i> | |
| Michael Wiegand and Dietrich Klakow | 13 |
| <i>Dependency-Based Text Compression for Semantic Relation Extraction</i> | |
| Marcos Garcia and Pablo Gamallo | 21 |
| <i>How to Distinguish a Kidney Theft from a Death Car? Experiments in Clustering Urban-Legend Texts</i> | |
| Roman Grundkiewicz and Filip Graliński | 29 |
| <i>Machine Reading Between the Lines: A Simple Evaluation Framework for Extracted Knowledge Bases</i> | |
| Avirup Sil and Alexander Yates | 37 |
| <i>Temporal Expressions Extraction in SMS messages</i> | |
| Stéphanie Weiser, Louis-Amélie Cougnon and Patrick Watrin | 41 |
| <i>INVITED TALK 2: Bringing Multilingual Information Extraction to the User</i> | |
| Ralf Steinberger | 45 |

Workshop Program

Friday: September 16, 2011

09:00–09:05 **Welcome**

Session I

09:05–10:05 **Invited Talk 1**

The Knowledge Base Population Task: Challenges for Information Extraction
Ralph Grishman, New York University

10:05–10:30 *Fine-grained Entity Set Refinement with User Feedback*
Bonan Min and Ralph Grishman

10:30–11:00 **BREAK**

Session II

11:00–11:25 *Extraction of Domain-specific Opinion Words for Similar Domains*
Ilia Chetviorkin and Natalia Loukachevitch

11:25–11:50 *The Role of Predicates in Opinion Holder Extraction*
Michael Wiegand and Dietrich Klakow

11:50–12:15 *Dependency-Based Text Compression for Semantic Relation Extraction*
Marcos Garcia and Pablo Gamallo

12:15–14:15 **LUNCH**

Session III

14:15–14:40 *How to Distinguish a Kidney Theft from a Death Car?*
Experiments in Clustering Urban-Legend Texts
Roman Grundkiewicz and Filip Graliński

14:40–15:05 *Machine Reading Between the Lines:*
A Simple Evaluation Framework for Extracted Knowledge Bases
Avirup Sil and Alexander Yates

15:05–15:30 *Temporal Expressions Extraction in SMS messages*
Stéphanie Weiser, Louis-Amélie Cougnon and Patrick Watrin

15:30–16:00 **BREAK**

Friday: September 16, 2011 (continued)

Session IV

16:00–17:00 **Invited Talk 2**

Bringing Multilingual Information Extraction to the User

Ralf Steinberger, European Commission Joint Research Centre

17:00–18:00 **Panel**

Moderator: Preslav Nakov

Panelists:

- Kevin Cohen, University of Colorado, School of Medicine
- Georgi Georgiev, Ontotext AD
- Petya Osenova, Bulgarian Academy of Sciences
- Elena Pakaleva, Bulgarian Academy of Sciences
- Kiril Simov, Bulgarian Academy of Sciences

18:00–18:05 **Closing Remarks**