INTERNATIONAL WORKSHOP

BIOMEDICAL INFORMATION EXTRACTION

held in conjunction with the International Conference

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

PROCEEDINGS

Edited by
Guergana Savova, Vangelis Karkaletsis and Galia Angelova

Borovets, Bulgaria

18 September 2009

International Workshop

BIOMEDICAL INFORMATION EXTRACTION

PROCEEDINGS

Borovets, Bulgaria
18 September 2009

ISBN 978-954-452-013-7

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Welcome to the Biomedical Information Extraction Workshop at RANLP09!

Undoubtly, the availability of information for biomedicine in an electronic format has been rapidly increasing. The Medical Literature Analysis and Retrieval system (Medline) houses millions of biomedical scientific literature publications. PubMed offers the power of a search engine for accessing the Medline content. On the other hand, the electronization of clinical data within the Electronic Medical Record (EMR) provides another powerful source for information extraction. Access to integrated information is critical for health care improvement, research, and the overall science of healthcare delivery and personalized medicine. Information extraction from the scientific literature is distinct from information extraction from the clinical narrative as these two types of genre have their own stylistic characteristics and pose different methodological challenges. Moreover, biomedical information spans multiple languages thus necessitating methods for multi-lingual information extraction.

In addition to the biomedical scientific literature and clinical data, we need also to consider the large number of health related web resources that is increasing day by day. The content of these resources is rather variable and difficult to assess. Furthermore, the number of people searching for health-related information is also increasing. The development of tools to support the process of describing the content of medical web resources with meta-data that facilitate their retrieval, and with quality labels by certified authorities, is crucial for the delivery of content of better quality to health information consumers. Multi-lingual information extraction has a significant role to play there also. The focus of this workshop is natural language processing and information extraction for biomedicine, including scientific and clinical free-text as well as health-related web resources within one or many languages. The presentations accepted for inclusion in the workshop are divided into two main groups: full papers and pilot project notes. The full papers describe mature investigative efforts, while the project notes showcase preliminary results on work-in-progress. The topics covered by the full-papers and the pilot project notes span a range:

- Natural language processing techniques for basic tasks, e.g. sentence boundary detection, tokenization, part of speech tagging, shallow parsing and deep parsing. Evaluation and comparison with the general domain;
- Efforts to create sharable biomedical lexical resources, including the annotation of biomedical data for linguistic and domain events;
- Biomedical named entity recognition;
- Methods for higher level biomedical language processing, e.g. relation discovery, temporal relation discovery, anaphoric relation discovery;
- Terminology/ontology biomedical named entity mapping;
- Integrated knowledge management of scientific and clinical free-text data;

- Knowledge representation and management technologies (e.g. OWL, RDF, Annotation Schemas, etc.) that enable the creation of machine-processable descriptions of health-related web resources;
- Content collection and information extraction techniques that allow the quality labeling of web resources and the continuous monitoring of already labeled ones;
- Multi-lingual information extraction.

We hope that this 1st workshop on Biomedical Information Extraction becomes a tradition within the RANLP conference. We would like to thank all the authors for their efforts in making it a highly productive workshop and a lively venue for exchange of scientific ideas! We invite you to consider submitting to the 2nd edition of the workshop as part of RANLP-2011.

September 2009

Guergana Savova
Vangelis Karkaletsis
Galia Angelova

Organisers and Sponsors

The International Workshop on Biomedical Information Extraction is organised by:

Guergana Savova, PhD, Assistant Professor in Medical Informatics, Mayo Clinic School of Medicine, Rochester, Minnesota, USA (Chair)

Vangelis Karkaletsis, Research Director, Institute of Informatics and Telecommunications, National Centre for Scientific Research (NCSR) Demokritos, Athens, Greece

Galia Angelova, PhD, Associate Professor in Computer Science, Head of Linguistic Modeling Department, Institute of Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria

The International Workshop on Biomedical Information Extraction is partially supported by:

The National Science Fund, Bulgaria,
via contract EVTIMA DO 02-292/December 2008
with the Institute of Parallel Processing, Bulgarian Academy of Sciences

PROGRAMME COMMITTEE

Werner Ceusters, Psychiatry and Ontology, SUNY at Buffalo
Wendy Chapman, Biomedical informatics, University of Pittsburgh
Cheryl Clark, MITRE Corporation
Kevin Cohen, University of Colorado
Noemie Elhadad, Biomedical Informatics, Columbia University
Udo Hahn, Jena University
Dimitris Kokkinakis, Gothenburg University
Stasinos Konstantopoulos, Institute of Informatics and Telecommunications, Athens
Anastassia Krithara, Institute of Informatics and Telecommunications, Athens
John Pestian, Biomedical Informatics, Cincinnatis Childrens Hospital
Sunghwan Sohn, Biomedical statistics and informatics, Mayo Clinic
Vojtech Svatek, University of Economics, Prague

REVIEWERS

In addition to the members of the Programme Committee and the Organisers, the following colleagues were involved in the reviewing process:

Svetla Boytcheva, State University of Library Studies and Information Technologies, Bulgaria
Georgi Georgiev, Ontotext AD, Bulgaria
Pythagoras Karampiperis, Institute of Informatics and Telecommunications, Athens, Greece
Preslav Nakov, National University of Singapore, Singapore
Dimitar Tcharaktchiev, Medical University, Sofia, Bulgaria

Table of Contents

Workshop Program

18 September 2010

Full Paper Presentations

Cascading Classifiers for Named Entity Recognition in Clinical Notes
Yefeng Wang and Jon Patrick

Issues on Quality Assessment of SNOMED CT® Subsets Term Validation and Term Extraction
Dimitrios Kokkinakis and Ulla Gerdin

A Joint Model for Normalizing Gene and Organism Mentions in Text
Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Deyan Peychev and Vassil Momchev

Natural Language Processing to Detect Risk Patterns Related to Hospital Acquired Infections
Denys Proux, Pierre Marchal, Frédérique Segond, Ivan Kergourlay, Stéfan Darmoni, Suzanne Pereira, Quentin Gicquel and Marie Hélène Metzger

Extraction and Exploration of Correlations in Patient Status Data
Svetla Boytcheva, Ivelina Nikolova, Elena Paskaleva, Galia Angelova, Dimitar Tcharaktchiev and Nadya Dimitrova

Pilot Project Notes Presentations

Corpus Study of Kidney-related Experimental Data in Scientific Papers
Brigitte Grau, Anne-Laure Ligozat and Anne-Lyse Minard

Semantic Portals in Biomedicine: Case Study
Irina Efimenko, Sergey Minor, Anatoli Starostin and Vladimir Khoroshevsky

Discussion