NLP for Medicine and Biology 2013

Proceedings of the Workshop on NLP for Medicine and Biology

associated with

The 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013)

13 September, 2013
Hissar, Bulgaria

WORKSHOP ON NLP FOR MEDICINE AND BIOLOGY
ASSOCIATED WITH THE INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING'2013

PROCEEDINGS

Hissar, Bulgaria
13 September 2013

ISBN 978-954-452-024-3

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

Biomedical NLP deals with the processing of healthcare-related text—clinical documents created by physicians and other healthcare providers at the point of care, scientific publications in the areas of biology and medicine, and consumer healthcare text such as social media blogs. Recent years have seen dramatic changes in the types and amount of data available to researchers in this field. Where most research on publications in the past has dealt with the abstracts of journal articles, we now have access to the full texts of journal articles via PubMedCentral. Where research on clinical documents has been hampered by a lack of availability of data, we now have access to large bodies of data through the auspices of the Cincinnati Children's Hospital NLP Challenge, the i2b2 shared tasks (www.i2b2.org), the TREC Electronic Medical Records track, the US-funded Strategic Health Advanced Research Projects Area 4 (www.sharpn.org) and the Shared Annotated Resources (ShARe; https://sites.google.com/site/shareclefehealth/taskdescription; www.clinicalnlpannotations.org) project. Meanwhile, the number of abstracts in PubMed continues to grow exponentially. Text in the form of blogs created by patients discussing various healthcare topics has emerged as another data source, with a new perspective on healthrelated issues. Connecting the information from the three main sources in multiple languages to the scientific community, the healthcare provider, and the healthcare consumer presents new challenges.

The Natural language processing for medicine and biology workshop at RANLP 2013 provided a venue for presentations of current work in this field. The topics of papers and posters presented at the workshop included finding domainspecific symptoms in patient records, helping parents understand diseases, phenotyping, and deidentification of clinical text. We gratefully acknowledge the contributions of

- Sophia Ananiadou, University of Manchester, UK
- William A. Baumgartner Jr., University of Colorado School of Medicine, USA
- Svetla Boytcheva, American University in Bulgaria, BG
- Dina Demner-Fushman, US National Library of Medicine, USA
- Dmitriy Dligach, Childrens Hospital Boston and Harvard Medical School, USA
- Timothy Miller, Childrens Hospital Boston and Harvard Medical School, USA
- Sameer Pradhan, Childrens Hospital Boston and Harvard Medical School, USA
- Angus Roberts, University of Sheffield, UK

Organizers:

Guergana Savova (Children's Hospital Boston and Harvard Medical School)
Kevin Bretonnel Cohen (University of Colorado School of Medicine)
Galia Angelova (IICT Bulgarian Academy of Sciences)

Table of Contents

Workshop Programme

Friday September 13, 2013

9:00–9:15 Opening

Session 1

9:15–10:15 **Invited talk: Guergana Savova** (Harvard Medical School and Childrens Hospital Boston, USA) *Temporal Relations in the Clinical Domain and Apache cTAKES*

The presentation will consist of two parts. Part 1 will present an overview of methods and software development behind the Apache cTAKES platform (ctakes.apache.org). The second part of the presentation will shift to current research on temporal relations in the clinical domain. The research is done as a collaboration among Harvard, University of Colorado and Mayo Clinic.

10:15–10:45 *Finding Negative Symptoms of Schizophrenia in Patient Records*
Genevieve Gorrell, Angus Roberts, Richard Jackson and Robert Stewart

10:45–11:15 Coffee break

Session 2

11:15–11:45 *NLP can help parents to understand rare diseases*
Marina Sokolova, Ilya Ioshikhes, Hamid Poursepanj and Alex MacKenzie

11:45–12:15 *Active Learning for Phenotyping Tasks*
Dmitriy Dligach, Timothy Miller and Guergana Savova

12:15–12:35 *De-Identification of Clinical Free Text in Dutch with Limited Training Data: A Case Study*
Elyne Scheurwegs, Kim Luyckx, Filip Van der Schueren and Tim Van den Bulcke