EACL 2014

14th Conference of the European Chapter of the Association for Computational Linguistics

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)

April 27, 2014
Gothenburg, Sweden

Stockholm
University

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

ISBN 978-1-937284-90-9

Introduction

Welcome to Louhi 2014: The Fifth International Workshop on Health Text Mining and Information Analysis, in Gothenburg, Sweden. The Louhi workshop series is an international, scientific, forum for researchers and practitioners in the multidisciplinary area of health text mining and information analysis. This research area has progressed and grown since the First Louhi workshop in Turku, 2008, addressing challenging research issues in the health and biomedical domain, leading to more openly available annotated corpora, tools, terminologies, etc. Moreover, work on other languages than the previously dominating English is both increasing and maturing.

The importance of accurate and specific information extraction and classification from diverse health and biomedical documents such as Electronic Health Records (EHRs), scientific literature and online health forums is evident for several differing purposes, e.g. detecting drugs and medications, adverse events, building timelines, understanding information needs. Ontologies and terminologies for such tasks are also crucial. The papers presented in this workshop all address these issues from different perspectives, and on different languages such as Basque, Spanish, French, Danish, Swedish, German and English. The Fifth Louhi workshop will provide a platform for important and useful discussions in this vivid research area, and hopefully lead to many more fruitful endeavours.

We received in total 21 submissions from 11 countries and three continents, and after a rigorous double-blind peer-review process we could accept 17 of these submissions (nine long papers and eight short papers) to be published in the 2014 Louhi proceedings. The acceptance rate for long papers was 60%, while the overall acceptance rate for the workshop was 81%.

A short description of each paper follows in order of appearance.

Long papers:

Medical queries in a Swedish health portal are studied from the perspective of supporting information needs by using semantic- and graph-based methods (*Moradi et al.*).

Collier et al. experiment with five strategies for mitigating the impact of near domain transference for biomedical named entity recognition. Distributional dissimilarities of domains need to be adequately compensated during learning, or else lower performance and higher annotation costs are expected.

Zhao and Tou Ng also address domain adaptation, but in the area of coreference resolution, using active learning methods and target domain instance weighting.

Discourse parsing is addressed in the paper by *Stepanov and Riccardi*, where cross-domain evaluation of a discourse relation parser trained on one domain generalises well across domains with feature-level domain adaptation.

Perez-de-Viñaspre and Oronoz present initial work on semi-automatically translating SNOMED CT into Basque, using the English version of SNOMED CT as source and then adapting the terms to Basque utilizing various rules.

A method for building FrameNet-like corpora using ontologies is described in the paper by *Tan*. The system includes algorithms for selecting and describing appropriate concepts to be translated into semantic frames.

Quan and Ren describe work on gene-disease association extraction by combining information filtering, grammar parsing and network analysis. With breast cancer as testing disease, they achieve 83.9% accuracy for the testing genes and diseases and 74.2% accuracy for the testing genes.

Segura-Bedmar et al. describe work on detecting drugs and adverse events from Spanish health social media posts. A gold standard is created for evaluation, and a multilingual text analysis engine, Textalytics, was applied for automatic detection, achieving 80%/56% precision, and 87%/85% recall for drugs/adverse events.

The paper by *Moen et al.* presents several methods for information retrieval, focusing on care episode retrieval based on textual similarity using distributional semantics and ICD-10 codes of diagnoses, to retrieve the most similar care episodes among the records.

Short papers:

Engel Thomas et al. present work on text mining of Danish health records, with a focus on handling spelling and ending variations, gaps and shuffling of terms, as well as negation identification and scope. Spelling variation was found to be the most important functionality.

A new annotated corpus for identifying phenotype information for congestive heart failure is presented by *Alnazzawi et al.* This corpus is unique in that it integrates information both from electronic health records and literature articles.

Medication extraction, as defined in the 2009 i2b2 challenge, is addressed by using agile text mining methods in the paper by *Shivade et al.* They report results of 92% precision and 71.5% recall.

Roller and Stevenson describe work using the Unified Medical Language System (UMLS) for distantly supervised relation extraction.

Casillas et al. describe work on extracting cause-effect relations between drugs and diseases, applied on Spanish health records.

Semantic relations are integrated in a vector space model to tackle the problem of context-unawareness and applied on an electronic health record corpus (*Périnet and Hamon*).

Kreuzthaler and Schulz present work on disambiguating period characters in German clinical discharge summaries. An accuracy of 93% is reported for abbreviation detection and sentence delimitation.

The Heideltime system is used for identifying time expressions in English and French in the paper by *Hamon and Grabar*. Results of 0.94 (French) and 0.85 (English) F-measure are reported on their adapted version of the system.

Dear reader, most welcome to study these proceedings, which we hope will raise interest, open new perspectives and generate new exciting research questions in health text mining and information analysis.

Stockholm and San Diego, March 2014

Sumithra Velupillai, Martin Duneld, Maria Kvist and Hercules Dalianis

Chair:

Sumithra Velupillai, DSV/Stockholm University

Program Chairs:

Hercules Dalianis, DSV/Stockholm University
Maria Kvist, DSV/Stockholm University
Martin Duneld, DSV/Stockholm University

Local Organizing Committee:

Maria Skeppstedt, DSV/Stockholm University
Aron Henriksson, DSV/Stockholm University

Publication Chair:

Martin Duneld, DSV/Stockholm University

Program Committee:

Anette Hulth, Swedish Institute for Infectious Disease Control, Karolinska Institutet, Sweden
Antti Airola, University of Turku, Finland
Beáta Megyesi, Uppsala University, Sweden
David Martinez, NICTA, Australia
Dimitris Kokkinakis, University of Gothenburg, Sweden
Filip Ginter, University of Turku, Finland
Gintaré Grigonyté, Stockholm University, Sweden
Hanna Suominen, NICTA, Australia
Henning Müller University of Applied Sciences Western Switzerland, Switzerland
Jon D. Patrick, Health Language Laboratories, Australia
Jong C. Park, KAIST Computer Science, Korea
Jussi Karlgren, KTH, Royal Institute of Technology, Sweden
Mats Wirén, Stockholm University, Stockholm
Özlem Uzuner, MIT, U.S.
Pierre Zweigenbaum, LIMSI, France
Richárd Farkas, Institute of Informatics, Hungary
Sabine Bergler, Concordia University, Canada
Sampo Pyysalo, University of Tokyo, Japan
Sanna Salanterä, University of Turku, Finland
Sophia Ananiadou, University of Manchester, U.K.
Stefan Schulz, Graz General Hospital and University Clinics, Austria
Tapio Salakoski, University of Turku, Finland
Thomas Brox Røst, Norwegian University of Science and Technology, Norway

Invited Speaker:

Sophia Ananiadou, University of Manchester, U.K.

Table of Contents

Workshop Program

Sunday 27 April 2014

08:45 Opening Remarks by Sumithra Velupillai, Department of Computer and Systems Sciences (DSV), Stockholm University

Keynote Talk

09:00 *Keynote: Supporting evidence-based medicine using text mining*
Sophia Ananiadou

Session I

10:05 *A Graph-Based Analysis of Medical Queries of a Swedish Health Care Portal*
Farnaz Moradi, Ann-Marie Eklund, Dimitrios Kokkinakis, Tomas Olovsson and Philippas Tsigas

10:30 Coffee

Session II

11:00 *The impact of near domain transfer on biomedical named entity recognition*
Nigel Collier, Mai-vu Tran and Ferdinand Paster

11:25 *Domain Adaptation with Active Learning for Coreference Resolution*
Shanheng Zhao and Hwee Tou Ng

11:55 *Towards Cross-Domain PDTB-Style Discourse Parsing*
Evgeny Stepanov and Giuseppe Riccardi

12:20 Lunch Break

Sunday 27 April 2014 (continued)

Session III

13:45 *Translating SNOMED CT Terminology into a Minor Language*
Olatz Perez-de-Viñaspre and Maite Oronoz

14:10 *A System for Building FrameNet-like Corpus for the Biomedical Domain*
He Tan

14:35 *Gene–disease association extraction by text mining and network analysis*
Changqin Quan and Fuji Ren

Poster Session (with coffee 15:30-16:00)

Negation scope and spelling variation for text-mining of Danish electronic patient records
Cecilia Engel Thomas, Peter Bjødstrup Jensen, Thomas Werge and Søren Brunak

Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature
Noha Alnazzawi, Paul Thompson and Sophia Ananiadou

Precise Medication Extraction using Agile Text Mining
Chaitanya Shivade, James Cormack and David Milward

Applying UMLS for Distantly Supervised Relation Detection
Roland Roller and Mark Stevenson

Adverse Drug Event prediction combining shallow analysis and machine learning
Sara Santiso, Arantza Casillas, Alicia Perez, Maite Oronoz and Koldo Gojenola

Reducing VSM data sparseness by generalizing contexts: application to health text mining
Amandine Périnet and Thierry Hamon

Disambiguation of Period Characters in Clinical Narratives
Markus Kreuzthaler and Stefan Schulz

Tuning HeidelTime for identifying time expressions in clinical texts in English and French
Thierry Hamon and Natalia Grabar

Sunday 27 April 2014 (continued)

Session IV

16:00 *Detecting drugs and adverse events from Spanish social media streams*
Isabel Segura-Bedmar, Ricardo Revert and Paloma Martínez

16:25 *Care Episode Retrieval*
Hans Moen, Erwin Marsi, Filip Ginter, Laura-Maria Murtola, Tapio Salakoski and Sanna Salanterä

16:50 Closing Remarks by Sumithra Velupillai, Department of Computer and Systems Sciences (DSV), Stockholm University