NAACL HLT 2010

Second Louhi Workshop on Text and Data Mining of Health Documents (Louhi-10)

Proceedings of the Workshop

June 5, 2010
Los Angeles, California

USB memory sticks produced by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

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

Preface

Welcome to the Second Louhi Workshop on Text and Data Mining of Health Documents, Louhi 2010, in Los Angeles, California, USA. We aim to bring together researchers and practitioners in a multidisciplinary conference on new uses of computer systems for data mining in the health care area. The very first Louhi Conference 2008 in Turku, Finland, showed the diversity and complexity of these issues. The increasing access to clinical data from health care systems and the progress of new methods and approaches in text and data mining results in a growing attention from the perspective of patient care as well as biomedical research and education. Improving performance in medical information retrieval is a challenge of complex nature requiring several approaches. Automatic indexing of clinical findings, observations, diseases and treatments is of high relevance for clinical work and research in general. Implementing decision support and guidelines to reach evidence-based practice is a specific area of interest (Fiszman et al. 2000). Detecting adverse event is another field that needs a high-quality system to detect findings and events in clinical documents (Griffin and Resar 2009). In several other areas indexing and mining in clinical text are of utmost importance – from everyday clinical work to translational biomedical research (Meystre et al. 2008). Health care consumer web sites as well as news web sites contain important information worthwhile monitoring to extract both information on specific diseases directed to the layman as well as epidemiological information. The papers presented in this workshop aim at exploring computational methods and tools to improve and support the work in these different fields. The Second Louhi workshop will continue to focus and reflect on computer use in every-day clinical work in hospitals and clinics such as electronic record systems, computer aided summaries, clinical coding, computerized clinical guidelines, computer decision systems, as well as related ethical concerns and security. Much of this work concerns itself by necessity with computer aided language use, and as such Louhi aims at providing an arena for report on development in a diversity of languages. In the papers presented at Louhi 2010 we can read about many of the challenges identified above.

A short description of each paper follows in order of appearance. In the first paper, *Friberg Heppin* describes the Swedish MedEval corpus, which has been annotated for the needs of both physicians and patients. The corpus has then been indexed with two different methods for an information retrieval experiment that aims to satisfy the requirements of both user groups. *Bhatia et al.* extracted information from English electronic health records containing diabetes information from a large number of patients, with the aim to detect populations at high risk of diabetes. *Skeppstedt* has ported the negation detection system NegEx, which is written for English clinical text, to Swedish, and describes the porting process in detail and finally evaluates the Swedish version of NegEx. *Schreitter et al.* describe a system for automatic speech recognition of dictated medical records in English. The aim of their work was to reduce errors in recognizing medication names, trademarks, dosages and strengths, and the authors use the Unified Medical Language System (UMLS) as a knowledge base for the recognition. *von Etter et al.* present an approach on monitoring epidemic information from online news articles, with epidemic intelligence officers as the intended user group. They have defined guidelines based on correctness and reliability together with the medical users and further annotated 1 000 articles that were then utilized in a machine learning based classification experiment. The paper by *Martin* is closely related to the work by *von Etter et al.* described above, but focuses on English web pages containing health care information directed to health care consumers. The author describes an annotation scheme based on type of information, applied by two students annotating 200 pages of health information documents.

Roque et al. present an overview of five open source visualization tools for electronic health records for medical practitioners. They describe the tools in the context of users, goals and tasks focusing on the temporal aspects of the visual presentation. *Melton et al.* present a system for identifying the long form of acronyms and abbreviations in biomedical text, using MetaMap applied on the UMLS to identify the long forms by expanding the acronyms. In *Allvin et al.* the authors have carried out both a qualitative and quantitative comparative study of Finnish and Swedish nursing narratives from two intensive care units. As the Swedish and Finnish languages belong to different language groups, while the countries are culturally closely related, this study explores how this might influence what is expressed in the narratives. *Halgrim et al.* describe a hybrid system for medical extraction based on both rule based and statistical classifiers. The system is applied on English narrative clinical records from the i2b2 challenge and uses several rule based processing steps where field detection is one significant step detecting if a medication occurs in narrative text or in a list of medications. *Kokkinakis and Toporowska Gronostaj* have carried out a pilot study to extract events from Swedish medical and clinical text, based on Frame Semantics and their methodology Swedish FrameNet++ (SFN++). *Hirschman and Aberdeen* have developed new metrics for the de-identification and the re-identification problem of clinical text. They emphasize that traditional information extraction metrics are not enough to address the real-world questions on "how good are current de-identification systems?". *Medori and Fairon* present a semi-automatic system for assigning ICD-9-CM codes to discharge summaries in French, and shows that stemmed and extracted specific encoding information gives better classification results than without pre-processing. Finally, *Lin et al.* present ongoing work using Conditional Random Fields to extract important information from clinical research articles, focusing on extracting formulaic information, metadata about the authors, longitudinal data and medical intervention methods.

We received in total 16 submissions from eleven countries and three continents, and after a rigorous double-blind peer-review process we could accept 14 of these submissions to be published in the Louhi 2010 proceedings.

Dear reader, most welcome to study this proceeding, which we hope will raise interest and open new perspectives in text and data mining of health documents.

Stockholm, April 2010

Hercules Dalianis, Martin Hassel and Gunnar Nilsson

References

Marcelo Fiszman and Peter J. Haug. 2000. Using medical language processing to support real-time evaluation of pneumonia guidelines. *Proceedings of AMIA Annual Symposium 2000*; 235-9.

Frances A. Griffin and Roger K.Resar. 2009. IHI Global Trigger Tool for Measuring Adverse Events (Second Edition). *IHI Innovation Series white paper*. Cambridge, MA: Institute for Healthcare Improvement; 2009.

Stéphane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and. John E. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. In: *IMIA Year-book of Medical Informatics 2008*. Methods Inf Med 2008; 47 Suppl 1:138-154.

Chair:

Hercules Dalianis, DSV/Stockholm University

Program co-chairs:

Martin Hassel, DSV/Stockholm University
Gunnar Nilsson, Karolinska Institutet

Organizers:

Hercules Dalianis, email: hercules@dsv.su.se
Martin Hassel, email: xmartin@dsv.su.se
Sumithra Velupillai, email: sumithra@dsv.su.se
Address: DSV, Stockholm University, Forum 100, 164 40 Kista, Sweden

Program Committee:

Sophia Ananiadou, University of Manchester, UK
Stephen Anthony, University of New South Wales, Australia
Henrik Boström, Stockholm University
Søren Brunak, Technical University of Denmark, DTU
Wendy Chapman, University of Pittsburgh
Aaron Cohen, Oregon Health & Science University
Richárd Farkas, University of Szeged, Hungary
Filip Ginter, University of Turku, Finland
Helena Karsten, Åbo Akademi, Finland
Dimitrios Kokkinakis, University of Gothenburg
Anette Hulth, Swedish Institute for Infectious Disease Control, Sweden
Sabine Koch, Karolinska institutet, Sweden
Jong C. Park, KAIST, South Korea
Tapio Pahikkala, University of Turku, Finland
Serguei Pakhomov, Center for Clinical and Cognitive Neuropharmacology, University of Minnesota, USA
Jon D. Patrick, University of Sydney, Australia
Sampo Pyysalo, University of Tokyo
Tapio Salakoski, University of Turku, Finland
Sanna Salanterä, University of Turku, Finland
Laura Slaughter, NTNU, Norway
Hanna Suominen, University of Turku, Finland
György Szarvas, UKP Lab, Technical University of Darmstadt, Germany
Özlem Uzuner, University at Albany, State University of New York, USA
Jaak Vilo, University of Tartu, Estonia
Pierre Zweigenbaum, LIMSI, France
Hans Åhlfeldt, Linköping University, Sweden

Publication chair:

Hercules Dalianis, DSV/ Stockholm University

Local organizing chair:

Sumithra Velupillai, DSV/ Stockholm University

Invited Speaker:

Eduard Hovy, Information Sciences Institute of the University of Southern California

Table of Contents

MedEval- A Swedish Medical Test Collection with Doctors and Patients User Groups
Karin Friberg Heppin 1

Extracting Information for Generating A Diabetes Report Card from Free Text in Physicians Notes
Ramanjot Singh Bhatia, Amber Graystone, Ross A Davies, Susan McClinton, Jason Morin and Richard F Davies 8

Negation Detection in Swedish Clinical Text
Maria Skeppstedt 15

Using Domain Knowledge about Medications to Correct Recognition Errors in Medical Report Creation
Stephanie Schreitter, Alexandra Klein, Johannes Matiasek and Harald Trost 22

Assessment of Utility in Web Mining for the Domain of Public Health
Peter von Etter, Silja Huttunen, Arto Vihavainen, Matti Vuorinen and Roman Yangarber 29

Reliability and Type of Consumer Health Documents on the World Wide Web: an Annotation Study
Melanie Martin 38

Automated Identification of Synonyms in Biomedical Acronym Sense Inventories
Genevieve B. Melton, SungRim Moon, Bridget McInnes and Serguei Pakhomov 46

Characteristics and Analysis of Finnish and Swedish Clinical Intensive Care Nursing Narratives
Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravicius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgren-Laine, Gunnar Nilsson, Øystein Nytrø, Sanna Salanterä, Maria Skeppstedt, Hanna Suominen and Sumithra Velupillai 53

Extracting Medication Information from Discharge Summaries
Scott Halgrim, Fei Xia, Imre Solti, Eithon Cadag and Özlem Uzuner 61

Linking SweFN++ with Medical Resources, towards a MedFrameNet for Swedish
Dimitrios Kokkinakis and Maria Toporowska Gronostaj 68

Measuring Risk and Information Preservation: Toward New Metrics for De-identification of Clinical Texts
Lynette Hirschman and John Aberdeen 72

A Comparison of Several Key Information Visualization Systems for Secondary Use of Electronic Health Record Content
Francisco Roque, Laura Slaughter and Aleksandr Tkatšenko 76

Machine learning and features selection for semi-automatic ICD-9-CM encoding
Julia Medori and Cédrick Fairon 84

Extracting Formulaic and Free Text Clinical Research Articles Metadata using Conditional Random Fields
Sein Lin, Jun-Ping Ng, Shreyasee Pradhan, Jatin Shah, Ricardo Pietrobon and Min-Yen Kan . . 90

Workshop Program

Saturday, June 5, 2010

Session I

8:45–9:00 Opening Remarks

9:00–10:00 Invited Talk: Creating Training Material for Health Informatics: Toward a Science of Annotation, Eduard Hovy

10:00–10:30 *MedEval- A Swedish Medical Test Collection with Doctors and Patients User Groups*
Karin Friberg Heppin

10:30–11:00 **Morning break**

Session II: Paper presentations

11:00–11:30 *Extracting Information for Generating A Diabetes Report Card from Free Text in Physicians Notes*
Ramanjot Singh Bhatia, Amber Graystone, Ross A Davies, Susan McClinton, Jason Morin and Richard F Davies

11:30–12:00 *Negation Detection in Swedish Clinical Text*
Maria Skeppstedt

12:00–12:30 *Using Domain Knowledge about Medications to Correct Recognition Errors in Medical Report Creation*
Stephanie Schreitter, Alexandra Klein, Johannes Matiasek and Harald Trost

12:30–2:00 **Lunch break**

Saturday, June 5, 2010 (continued)

Session III: Paper presentations

2:00–2:30 *Assessment of Utility in Web Mining for the Domain of Public Health*
Peter von Etter, Silja Huttunen, Arto Vihavainen, Matti Vuorinen and Roman Yangarber

2:30–3:00 *Reliability and Type of Consumer Health Documents on the World Wide Web: an Annotation Study*
Melanie Martin

3:00–3:30 **Afternoon break**

Session IV: Poster presentations

3:00–4:00 *Automated Identification of Synonyms in Biomedical Acronym Sense Inventories*
Genevieve B. Melton, SungRim Moon, Bridget McInnes and Serguei Pakhomov

Characteristics and Analysis of Finnish and Swedish Clinical Intensive Care Nursing Narratives
Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravicius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgren-Laine, Gunnar Nilsson, Øystein Nytrø, Sanna Salanterä, Maria Skeppstedt, Hanna Suominen and Sumithra Velupillai

Extracting Medication Information from Discharge Summaries
Scott Halgrim, Fei Xia, Imre Solti, Eithon Cadag and Özlem Uzuner

Linking SweFN++ with Medical Resources, towards a MedFrameNet for Swedish
Dimitrios Kokkinakis and Maria Toporowska Gronostaj

Measuring Risk and Information Preservation: Toward New Metrics for De-identification of Clinical Texts
Lynette Hirschman and John Aberdeen

A Comparison of Several Key Information Visualization Systems for Secondary Use of Electronic Health Record Content
Francisco Roque, Laura Slaughter and Aleksandr Tkatšenko

Saturday, June 5, 2010 (continued)

Session V: Paper presentations

4:00–4:30 *Machine learning and features selection for semi-automatic ICD-9-CM encoding*
Julia Medori and Cédrick Fairon

4:30–5:00 *Extracting Formulaic and Free Text Clinical Research Articles Metadata using Conditional Random Fields*
Sein Lin, Jun-Ping Ng, Shreyasee Pradhan, Jatin Shah, Ricardo Pietrobon and Min-Yen Kan

5:00–5:15 Closing Remarks and information about the next Louhi workshop