JOURNAL OF BIOMEDICAL SEMANTICS

INTRODUCTION Open Access

Louhi 2010: Special issue on Text and Data Mining of Health Documents

Hercules Dalianis*, Martin Hassel, Sumithra Velupillai

From Second Louhi Workshop on Text and Data Mining of Health Documents
Los Angeles, CA, USA.

* Correspondence: hercules@dsv.su.se
Department of Computer and System Sciences, (DSV) Stockholm University Forum 100, 164 40 Kista, Sweden

Abstract

The papers presented in this supplement focus and reflect on computer use in every-day clinical work in hospitals and clinics such as electronic health record systems, pre-processing for computer aided summaries, clinical coding, computer decision systems, as well as related ethical concerns and security. Much of this work concerns itself by necessity with incorporation and development of language processing tools and methods, and as such this supplement aims at providing an arena for reporting on development in a diversity of languages. In the supplement we can read about some of the challenges identified above.

Introduction

An increasing production of clinical textual data from electronic health record systems has pushed on the development of new methods and approaches for text and data mining. Since the nature of clinical data is complex and relate to many sources of human knowledge it is a challenging task that requires several approaches. Clinical text is noisy with numerous spelling errors and non-standard abbreviations, as well as incomplete sentences [1]. In parallel, people working in both patient care as well as biomedical research and education has discovered the potential of the combination of this abundant source of clinical data and advanced tools. One example is the automatic indexing of clinical findings, observations, diseases and treatments that is of high relevance for clinical work and research in general. The indexed data needs to be retrieved and this poses new challenges in the clinical and medical domain in which there is a lot of implicit and tacit knowledge.

Implementing decision support and guidelines to reach evidence-based practice is a specific area of interest [2]. Detecting adverse events is another field that needs a high-quality system to detect findings and events in clinical documents [3]. In several other areas, information retrieval and access of clinical text are of outmost importance – from everyday clinical work to translational biomedical research. Meystre et al. [4] present an elaborate and extensive overview of the research area. Health care consumer Web sites as well as news Web sites contain important information worthwhile monitoring to extract both information about specific diseases directed to laymen as well as epidemiological information. The papers presented in this supplement aim at exploring computational methods and tools to improve and support the work in these different fields. This

supplement focuses and reflects on computer use in every-day clinical work in hospitals and clinics such as electronic health record systems, pre-processing for computer aided summaries, clinical coding, computer decision systems, as well as related ethical concerns and security. Much of this work concerns itself by necessity with incorporation and development of language processing tools and methods, and as such this supplement aims at providing an arena for reporting on development in a diversity of languages. In the supplement we can read about some of the challenges identified above.

The workshop that preceded the outcome of this supplement was the Second Louhi Workshop on Text and Data Mining of Health Documents, Louhi 2010, that took place in Los Angeles, California, USA in June 2010. For more information on the workshop, see http://dsv.su.se/en/louhi10/.

Summary of selected papers

In the paper by Allvin et al. [5] 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. The conclusions are that the domain specific attributes as for example non-standard abbreviations, missing subject and spelling errors are similar over languages, but the electronic records systems differ in that the headings are defined by the nurses while writing in Finland but in Sweden they are predefined in templates.

The paper by Halgrim et al. [6] describes 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. The medication is divided into six different sub classes as medication name, dosage, frequency, duration, mode and reason for the medication. The authors obtained comparable results to the top systems in the 2009 i2b2 challenge.

To interpret clinical text and specifically to detect when a diagnosis is affirmed, one needs to perform negation detection. Skeppstedt [7] has ported the negation detection system NegEx by Chapman et al. [8] which is written for English clinical text, to SwedishThe paper describes the porting process in detail and evaluates the Swedish version of NegEx on 558 manually classified sentences containing negation triggers. Even though the precision was lower for the Swedish NegEx than for the English NegEx, it could still be concluded that the same trigger phrase approach is viable in a Swedish context since many negated propositions were identified through a limited set of trigger phrases.

Building resources for information retrieval purposes with a clinical focus is very important for the research community. With these, it may be possible to understand different user perspectives and to develop methods for tailored requirements. Friberg Heppin [9] describes the Swedish MedEval, a test collection for information retrieval research, where assessments have been made both for topical relevance and target reader groups (physicians and patients). This collection has been used for experiments with different user scenarios. Moreover, the collection has two indexes, one containing

split compounds, which is very useful for collections in morphologically complex languages such as Swedish.

Reliability, or trustworthiness, of information is also an important aspect in information retrieval for medical purposes. Martin [10] describes work on annotating Web documents containing health care information directed to health care consumers for reliability and type. Inter-rater reliability agreement results in this study are promising and opens up the possibility for automatic classification of medical information on the Web, which is very important for both professionals and lay people.

Dedication

This supplement is dedicated to professor Hans Åhlfeldt, Linköping University, Sweden, a colleague and also reviewer of Louhi 2010 workshop that passed away suddenly in September 2010.

Acknowledgments

We would like to thank Nordforsk – the Nordic council, for partly funding this work through HEXAnord-HEalth teXt Analysis network in the Nordic and Baltic countries. Our gratitude is also extended to the reviewers of both the Louhi 2010 workshop and this supplement:

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
Richrd Farkas, University of Szeged, Hungary
Filip Ginter, University of Turku, Finland
Anette Hulth, Swedish Institute for Infectious Disease Control, Sweden
Sabine Koch, Karolinska Institutet, Sweden
Gunnar H. Nilsson, 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
Sampo Pyysalo, University of Tokyo
Tapio Salakoski, University of Turku, Finland
Sanna Salanterä, University of Turku, Finland
Hanna Suominen, National ICT Australia (NICTA) and Australian National University, Australia
György Szarvas, UKP Lab, Technical University of Darmstadt, Germany
Jaak Vilo, University of Tartu, Estonia
Pierre Zweigenbaum, LIMSI, France

This article has been published as part of *Journal of Biomedical Semantics* Volume 2 Supplement 2, 2011: Proceedings of the Second Louhi Workshop on Text and Data Mining of Health Documents. The full contents of the supplement are available online at http://www.jbiomedsem.com/supplements/2/S3.

Competing interests

The authors declare that they have no competing interests.

Published: 14 July 2011

References

1. Levin MA, Krol M, Doshi AM, Reich DL: **"Extraction and Mapping of Drug Names from Free Text to a Standardized Nomenclature".** *Proceedings of AMIA Annual Symposium* 2007, 438-442.
2. Fiszman M, Haug PJ: **"Using medical language processing to support real-time evaluation of pneumonia guidelines".** *Proceedings of AMIA Annual Symposium* 2000, 235-9.
3. Griffin FA, Resar RK: **"IHI Global Trigger Tool for Measuring Adverse Events".** *IHI Innovation Series white paper.* Second edition. Cambridge, MA: Institute for Healthcare Improvement; 2009.
4. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JE: **"Extracting information from textual documents in the electronic health record: a review of recent research".** *Yearbook of Medical Informatics* 2008, 128-144.
5. Allvin H, Carlsson E, Dalianis H, Danielsson-Ojala R, Daudaravicius V, Hassel M, Kokkinakis D, Lundgren-Laine H, Nilsson GH, Nytrø Ø, Salanterä S, Skeppstedt M, Suominen H, Velupillai S: **"Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies".** *Journal of Biomedical Semantics* 2011, **2(Suppl 3)**:S1.
6. Halgrim S, Xia F, Solti I, Cadag E, Uzuner Ö: **"A cascade of classifiers for extracting medication information from discharge summaries".** *Journal of Biomedical Semantics* 2011, **2(Suppl 3)**:S2.
7. Skeppstedt M: **"Negation Detection in Swedish Clinical Text: An adaption of NegEx to Swedish".** *Journal of Biomedical Semantics* 2011, **2(Suppl 3)**:S3.

8. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: **"A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries".** In *Journal of iomedical nformatics. Volume 34.* Elsevier; 2001:(5):301-310.
9. Friberg Heppin K: **"MedEval- A Swedish Medical Test Collection with Doctors and Patients User Groups".** *Journal of Biomedical Semantics* 2011, **2(Suppl 3)**:S4.
10. Martin M: **"Reliability and type of consumer health documents on the World Wide Web: an annotation study".** *Journal of Biomedical Semantics* 2011, **2(Suppl 3)**:S5.

doi:10.1186/2041-1480-2-S3-I1
Cite this article as: Dalianis *et al.*: Louhi 2010: Special issue on Text and Data Mining of Health Documents. *Journal of Biomedical Semantics* 2011 **2**(Suppl 3):I1.