

**EMNLP 2015**

**SIXTH INTERNATIONAL WORKSHOP ON  
HEALTH TEXT MINING AND  
INFORMATION ANALYSIS  
(LOUHI)**

Proceedings of the workshop

17 September 2015

Lisbon, Portugal

Order print-on-demand copies from:

Curran Associates  
57 Morehouse Lane  
Red Hook, New York 12571 USA  
Tel: +1-845-758-0400  
Fax: +1-845-758-2633  
curran@proceedings.com

©2015 The Association for Computational Linguistics  
ISBN: 978-1-941643-32-7

## Preface

The Sixth International Workshop on Health Text Mining and Information Analysis provides an interdisciplinary forum for researchers interested in automated processing of health documents. Health documents encompass electronic health records, clinical guidelines, spontaneous reports for pharmacovigilance, biomedical literature, health forums/blogs or any other type of health related documents. The Louhi workshop series fosters interactions between the Computational Linguistics, Medical Informatics and Artificial Intelligence communities. It started in 2008 in Turku, Finland and has been organized five times: Louhi 2010 was co-located with NAACL in Los Angeles, CA; Louhi 2011 was co-located with Artificial Intelligence in Medicine (AIME) in Bled, Slovenia; Louhi 2013 was held in Sydney, Australia during NICTA Techfest; and Louhi 2014 was co-located with EACL in Gothenburg, Sweden.

The aim of the Louhi 2015 workshop is to bring together research work on topics related to text mining of health documents, particularly emphasizing multidisciplinary aspects of health documentation and the interplay between nursing and medical sciences, information systems, computational linguistics and computer science. The topics include, but are not limited to, the following Natural Language Processing techniques and related areas:

- Techniques supporting information extraction, e.g. named entity recognition, negation and uncertainty detection
- Classification and text mining applications (e.g. diagnostic classifications such as ICD-10 and nursing intensity scores) and problems (e.g. handling of unbalanced data sets)
- Text representation, including dealing with data sparsity and dimensionality issues
- Domain adaptation, e.g. adaptation of standard NLP tools (incl. tokenizers, PoS-taggers, etc) to the medical domain
- Information fusion, i.e. integrating data from various sources, e.g. structured and narrative documentation
- Unsupervised methods, including distributional semantics
- Evaluation, gold/reference standard construction and annotation
- Syntactic, semantic and pragmatic analysis of health documents
- Anonymization / de-identification of health records and ethics
- Supporting the development of medical terminologies and ontologies
- Individualization of content, consumer health vocabularies, summarization and simplification of text

- NLP for supporting documentation and decision making practices
- Predictive modeling of adverse events, e.g. adverse drug events and hospital acquired infections

The call for papers encouraged authors to submit papers describing substantial and completed work but also focus on a contribution, a negative result, a software package or work in progress. We also encouraged to report work on low-resourced languages, addressing the challenges of data sparsity and language characteristic diversity.

We received 39 submissions, an unprecedented high number for the LOUHI series. Each submission went through a double-blind review process which involved three program committee members. Based on comments and rankings supplied by the reviewers, we accepted 19 papers (11 long papers and 8 short papers). The overall acceptance rate is 49% and the acceptance rate for long papers is 50%. During the workshop, 8 papers have been presented orally, and 11 papers have been presented as posters.

Finally, we would like to thank the members of the program committee for the quality of their reviews in a very short period. We are very grateful to Marie-Francine Moens for accepting to give an invited talk. We would also like to thank the authors for their submissions and the quality of their work.

Cyril Grouin, Thierry Hamon, Aurélie Névéol, Pierre Zweigenbaum

**Organizers:**

Cyril Grouin, LIMSI-CNRS, Orsay, France  
Thierry Hamon, LIMSI-CNRS, Orsay, France & Université Paris 13  
Aurélie Névéol, LIMSI-CNRS, Orsay, France  
Pierre Zweigenbaum, LIMSI-CNRS, Orsay, France

**Program Committee:**

Sophia Ananiadou, University of Manchester, U.K.  
Sabine Bergler, Concordia University, Canada  
Thomas Brox Røst, Norwegian University of Science and Technology, Norway  
Kevin B. Cohen, University of Colorado/School of Medicine, USA  
Francisco Couto, University of Lisbon, Portugal  
Hercules Dalianis, Stockholm University, Sweden  
Louise Deléger, INRA, France  
Gaël Dias, Normandie University, France  
Martin Duneld/Hassel, Stockholm University, Sweden  
Richárd Farkas, Institute of Informatics, Hungary  
Filip Ginter, University of Turku, Finland  
Natalia Grabar, CNRS UMR 8163, STL Université de Lille3, France  
Gintarė Grigonytė, Stockholm University, Sweden  
Aron Henriksson, Stockholm University, Sweden  
Rezarta Islamaj, NIH/NLM/NCBI, USA  
Antonion Jimeno Yepes, IBM Research, Australia  
Jussi Karlgren, KTH, Royal Institute of Technology, Sweden  
Dimitrios Kokkinakis, University of Gothenburg, Sweden  
Maria Kvist, Stockholm University, Sweden  
Alberto Lavelli, Fondazione Bruno Kessler, Italy  
David Martínez, University of Melbourne and MedWhat.com, Australia  
Beáta Megyesi, Uppsala University, Sweden  
Marie-Jean Meurs, UQAM & Concordia University, QC, Canada  
Fleur Mougin, Université de Bordeaux, ERIAS, Centre INSERM U897, ISPED, France  
Danielle L Mowery, University of Utah, USA  
Henning Müller, University of Applied Sciences Western Switzerland, Switzerland  
Mariana Neves, Hasso-Plattner-Institute at the University of Potsdam, Germany  
Jong C. Park, KAIST Computer Science, Korea  
Jon D. Patrick, Health Language Laboratories, Australia  
Sampo Pyysalo, University of Turku, Finland  
Stefan Schulz, Graz General Hospital and University Clinics, Austria  
Tapio Salakoski, University of Turku, Finland  
Sanna Salanterä, University of Turku, Finland

Isabel Segura-Bedmar, Universidad Carlos III de Madrid, Spain  
Maria Skeppstedt, Gavagai and Linnaeus University, Sweden  
Hanna Suominen, NICTA, Australia  
Suzanne Tamang, Stanford University School of Medicine, USA  
Özlem Uzuner, MIT, U.S.A.  
Sumithra Velupillai, Stockholm University, Sweden  
Karin Verspoor, University of Melbourne, Australia  
Mats Wirén, Stockholm University, Stockholm, Sweden

**Invited Speaker:**

Marie-Francine Moens, Department of computer Science, Katholieke Universiteit Leuven

## Table of Contents

<i>In-depth annotation for patient level liver cancer staging</i> Wen-wai Yim, Sharon Kwan and Meliha Yetisgen .....	1
<i>Predicting Continued Participation in Online Health Forums</i> Farig Sadeque, Thamar Solorio, Ted Pedersen, Prasha Shrestha and Steven Bethard .....	12
<i>Redundancy in French Electronic Health Records: A preliminary study</i> Eva D'hondt, Xavier Tannier and Aurélie Névéol .....	21
<i>Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs?</i> Cyril Grouin, Nicolas Griffon and Aurélie Névéol .....	31
<i>An Analysis of Biomedical Tokenization: Problems and Strategies</i> Noa P. Cruz Diaz and Manuel Maña López .....	40
<i>Annotation of Clinically Important Follow-up Recommendations in Radiology Reports</i> Meliha Yetisgen, Prescott Klassen, Lucas McCarthy, Elena Pellicer, Tom Payne and Martin Gunn	50
<i>On the Impact of Twitter-based Health Campaigns: A Cross-Country Analysis of Movember</i> Nugroho Dwi Prasetyo, Claudia Hauff, Dong Nguyen, Tjts van den Broek and Djoerd Hiemstra	55
<i>Exploring Word Embedding for Drug Name Recognition</i> Isabel Segura-Bedmar, Víctor Suárez-Paniagua and Paloma Martínez .....	64
<i>Creating a rule based system for text mining of Norwegian breast cancer pathology reports</i> Rebecka Weegar and Hercules Dalianis .....	73
<i>Parser Adaptation to the Biomedical Domain without Re-Training</i> Jeff Mitchell and Mark Steedman .....	79
<i>Expanding a dictionary of marker words for uncertainty and negation using distributional semantics</i> Alyaa Alfalahi, Maria Skeppstedt, Rickard Ahlbom, Roza Baskalayci, Aron Henriksson, Lars Asker, Carita Paradis and Andreas Kerren .....	90
<i>Held-out versus Gold Standard: Comparison of Evaluation Strategies for Distantly Supervised Relation Extraction from Medline abstracts</i> Roland Roller and Mark Stevenson .....	97
<i>Checking a structured pathology report for completeness of content using terminological knowledge</i> Sebastian Busse .....	103
<i>Effectively Crowdsourcing Radiology Report Annotations</i> Anne Cocos, Aaron Masino, Ting Qian, Ellie Pavlick and Chris Callison-Burch .....	109
<i>Identifying Key Concepts from EHR Notes Using Domain Adaptation</i> Jiaping Zheng and Hong Yu .....	115
<i>Information Extraction from Biomedical Texts: Learning Models with Limited Supervision</i> Marie-Francine Moens .....	120

<i>Adverse Drug Event classification of health records using dictionary based pre-processing and machine learning</i>	
Stefanie Friedrich and Hercules Dalianis .....	121
<i>NLP-Based Readability Assessment of Health-Related Texts: a Case Study on Italian Informed Consent Forms</i>	
Giulia Venturi, Tommaso Bellandi, Felice Dell’Orletta and Simonetta Montemagni .....	131
<i>Mining and Ranking Biomedical Synonym Candidates from Wikipedia</i>	
Abhyuday Jagannatha, Jinying Chen and Hong Yu .....	142
<i>Representing Clinical Notes for Adverse Drug Event Detection</i>	
Aron Henriksson .....	152



## Program of the workshop

Thursday, September 17, 2015

### 09:00–10:30 Session I - Corpus creation

*In-depth annotation for patient level liver cancer staging*

Wen-wai Yim, Sharon Kwan and Meliha Yetisgen

*Predicting Continued Participation in Online Health Forums*

Farig Sadeque, Tamar Solorio, Ted Pedersen, Prasha Shrestha and Steven Bethard

*Redundancy in French Electronic Health Records: A preliminary study*

Eva D'hondt, Xavier Tannier and Aurélie Névéal

*Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs?*

Cyril Grouin, Nicolas Griffon and Aurélie Névéal

### 11:00–12:30 Session II - Poster

*An Analysis of Biomedical Tokenization: Problems and Strategies*

Noa P. Cruz Diaz and Manuel Maña López

*Annotation of Clinically Important Follow-up Recommendations in Radiology Reports*

Meliha Yetisgen, Prescott Klassen, Lucas McCarthy, Elena Pellicer, Tom Payne and Martin Gunn

*On the Impact of Twitter-based Health Campaigns: A Cross-Country Analysis of Movember*

Nugroho Dwi Prasetyo, Claudia Hauff, Dong Nguyen, Tijs van den Broek and Djord Hiemstra

*Exploring Word Embedding for Drug Name Recognition*

Isabel Segura-Bedmar, Víctor Suárez-Paniagua and Paloma Martínez

*Creating a rule based system for text mining of Norwegian breast cancer pathology reports*

Rebecka Weegar and Hercules Dalianis

*Parser Adaptation to the Biomedical Domain without Re-Training*

Jeff Mitchell and Mark Steedman

**Thursday, September 17, 2015 (continued)**

*Expanding a dictionary of marker words for uncertainty and negation using distributional semantics*

Alyaa Alfalahi, Maria Skeppstedt, Rickard Ahlbom, Roza Baskalayci, Aron Henriksson, Lars Asker, Carita Paradis and Andreas Kerren

*Held-out versus Gold Standard: Comparison of Evaluation Strategies for Distantly Supervised Relation Extraction from Medline abstracts*

Roland Roller and Mark Stevenson

*Checking a structured pathology report for completeness of content using terminological knowledge*

Sebastian Busse

*Effectively Crowdsourcing Radiology Report Annotations*

Anne Cocos, Aaron Masino, Ting Qian, Ellie Pavlick and Chris Callison-Burch

*Identifying Key Concepts from EHR Notes Using Domain Adaptation*

Jiaping Zheng and Hong Yu

**12:30–14:00** *Lunch break*

**14:00–15:30** **Session III - Invited talk**

*Information Extraction from Biomedical Texts: Learning Models with Limited Supervision*

Marie-Francine Moens

**16:00–17:30** **Session IV - Corpus processing**

*Adverse Drug Event classification of health records using dictionary based pre-processing and machine learning*

Stefanie Friedrich and Hercules Dalianis

*NLP-Based Readability Assessment of Health-Related Texts: a Case Study on Italian Informed Consent Forms*

Giulia Venturi, Tommaso Bellandi, Felice Dell'Orletta and Simonetta Montemagni

*Mining and Ranking Biomedical Synonym Candidates from Wikipedia*

Abhyuday Jagannatha, Jinying Chen and Hong Yu

*Representing Clinical Notes for Adverse Drug Event Detection*

Aron Henriksson

**Thursday, September 17, 2015 (continued)**

