EMNLP-IJCNLP 2019

# Tenth International Workshop on Health Text Mining and Information Analysis LOUHI 2019

## Proceedings of the Workshop

November 3, 2019
Hong Kong

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 978-1-950737-77-2

# Introduction (TBD)

The International Workshop on Health Text Mining and Information Analysis (LOUHI) 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. The eight previous editions of the workshop were co-located with SMBM 2008 in Turku, Finland, with NAACL 2010 in Los Angeles, California, with Artificial Intelligence in Medicine (AIME 2011) in Bled, Slovenia, during NICTA Techfest 2013 in Sydney, Australia, co-located with EACL 2014 in Gothenburg, Sweden, with EMNLP 2015 in Lisbon, Portugal, with EMNLP 2016 in Austin, Texas; in 2017 was held in Sydney, Australia; and in 2018 was co-located with EMNLP 2018 in Brussels, Belgium. This year the workshop is co-located with EMNLP 2019 in Hong Kong.

The aim of the LOUHI 2019 workshop is to bring together research work on topics related to 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
- Terminology and information model standards (SNOMED CT, FHIR) for health text mining
- Bridging gaps between formal ontology and biomedical NLP

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.

This year we received a high number of submissions (50), therefore the selection process was very competitive. Due to time and space limitations, we could only choose a small number of the submitted papers to appear in the program.

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 23 papers. Although the selection was entirely based on the scores provided by the reviewers, we regretfully had to set a relatively high threshold for acceptance. The overall acceptance rate is 46%. After the decision about acceptance, 2 papers were withdrawn by the authors. During the workshop, 11 papers will be presented orally, and 10 papers will be presented as posters.

Finally, we would like to thank the members of the program committee for providing balanced reviews in a very short period of time, and the authors for their submissions and the quality of their work.

**Organizers:**

Eben Holderness, McLean Hospital, Harvard Medical School & Brandeis University, USA
Antonio Jimeno Yepes, IBM Research, Australia
Alberto Lavelli, FBK, Trento, Italy
Anne-Lyse Minard, LLL-CNRS, Université d'Orléans, France
James Pustejovsky, Brandeis University, USA
Fabio Rinaldi, Dalle Molle Institute for Artificial Intelligence Research - IDSIA, University of Zurich, Switzerland & FBK, Trento, Italy

**Program Committee:**

Mohammad Akbari, National University of Singapore, Singapore
Rafael Berlanga Llavori, Universitat Jaume I, Spain
Georgeta Bordea, Université de Bordeaux, France
Leonardo Campillos Llanos, LIMSI, CNRS, France
Kevin B Cohen, University of Colorado/School of Medicine, USA
Francisco Couto, University of Lisbon, Portugal
Hercules Dalianis, Stockholm University, Sweden
Graciela Gonzalez-Hernandez, University of Pennsylvania, USA
Natalia Grabar, CNRS UMR 8163, STL Université de Lille3, France
Cyril Grouin, LIMSI, CNRS, Université Paris-Saclay, Orsay, France
Thierry Hamon, LIMSI, CNRS, Université Paris-Saclay, Orsay, France & Université Paris 13, Villetaneuse, France
Aron Henriksson, Stockholm University, Sweden
Eben Holderness, McLean Hospital, Harvard Medical School & Brandeis University, USA
Rezarta Islamaj-Dogan, NIH/NLM/NCBI, USA
Hyeju Jang, University of British Columbia, Canada
Yoshinobu Kano, Shizuoka University, Japan
Jin-Dong Kim, Research Organization of Information and Systems, Japan
Dimitrios Kokkinakis, University of Gothenburg, Sweden
Analia Lourenco, Universidade de Vigo, Spain
David Martinez, University of Melbourne and MedWhat.com, Australia
Sérgio Matos, University of Aveiro, Portugal
Marie-Jean Meurs, UQAM & Concordia University, QC, Canada
Timothy Miller, Harvard Medical School, USA
Hans Moen, University of Turku
Diego Molla, Maquaire University, Australia
Danielle L Mowery, University of Utah, USA
Henning Müller, University of Applied Sciences Western Switzerland, Switzerland
Goran Nenadic, University of Manchester, UK
Aurélie Névéol, LIMSI, CNRS, Université Paris-Saclay, Orsay, France
Mariana Lara Neves, German Federal Institute for Risk Assessment, Germany
Jong C. Park, KAIST Computer Science, Korea
Piotr Pęzik, University of Lodz, Poland
Laura Plaza, Universidad Complutense de Madrid, Spain
James Pustejovsky, Brandeis University, USA
Thomas Brox Røst, Norwegian University of Science and Technology, Norway
Tapio Salakoski, University of Turku, Finland

Sanna Salanterä, University of Turku, Finland
Stefan Schulz, Graz General Hospital and University Clinics, Austria
Maria Skeppstedt, Institute for Language and Folklore, Sweden
Amber Stubbs, Simmons College, USA
Hanna Suominen, CSIRO, Australia
Suzanne Tamang, Stanford University School of Medicine, USA
Sumithra Velupillai, KTH, Royal Institute of Technology, Sweden, and King's College London, UK
Özlem Uzuner, MIT, USA
Pierre Zweigenbaum, LIMSI, CNRS, Université Paris-Saclay, Orsay, France

**Additional Reviewers:**

Andre Lamurias, Universidade de Lisboa, Portugal
Bridget McInnes, Virginia Commonwealth University, USA
Meliha Yetisgen, University of Washington, USA

# Table of Contents

# Conference Program

**November 3, 2019**

**9:00–10:30 Session 1**

**9:00** ***Introduction***

9:05 *Cross-document coreference: An approach to capturing coreference without context*
Kristin Wright-Bettner, Martha Palmer, Guergana Savova, Piet de Groen and Timothy Miller

**09:30** ***Poster booster***

**09:45** ***Poster session***

*Comparing the Intrinsic Performance of Clinical Concept Embeddings by Their Field of Medicine*
John-Jose Nunez and Giuseppe Carenini

*On the Effectiveness of the Pooling Methods for Biomedical Relation Extraction with Deep Learning*
Tuan Ngo Nguyen, Franck Dernoncourt and Thien Huu Nguyen

*Syntax-aware Multi-task Graph Convolutional Networks for Biomedical Relation Extraction*
Diya Li and Heng Ji

*BioReddit: Word Embeddings for User-Generated Biomedical NLP*
Marco Basaldella and Nigel Collier

*Leveraging Hierarchical Category Knowledge for Data-Imbalanced Multi-Label Diagnostic Text Understanding*
Shang-Chi Tsai, Ting-Yun Chang and Yun-Nung Chen

*Experiments with ad hoc ambiguous abbreviation expansion*
Agnieszka Mykowiecka and Malgorzata Marciniak

*Multi-Task, Multi-Channel, Multi-Input Learning for Mental Illness Detection using Social Media Text*
Prasadith Kirinde Gamaarachchige and Diana Inkpen

**November 3, 2019 (continued)**

*Extracting relevant information from physician-patient dialogues for automated clinical note taking*
Serena Jeblee, Faiza Khan Khattak, Noah Crampton, Muhammad Mamdani and Frank Rudzicz

*Biomedical Relation Classification by single and multiple source domain adaptation*
Sinchani Chakraborty, Sudeshna Sarkar, Pawan Goyal and Mahanandeeshwar Gattu

*Assessing the Efficacy of Clinical Sentiment Analysis and Topic Extraction in Psychiatric Readmission Risk Prediction*
Elena Alvarez-Mellado, Eben Holderness, Nicholas Miller, Fyonn Dhang, Philip Cawkwell, Kirsten Bolton, James Pustejovsky and Mei-Hua Hall

**10:30–11:00 Break**

**11:00–12:30 Session 2**

11:00 *What does the language of foods say about us?*
Hoang Van, Ahmad Musa, Hang Chen, Stephen Kobourov and Mihai Surdeanu

11:25 *Dreaddit: A Reddit Dataset for Stress Analysis in Social Media*
Elsbeth Turcan and Kathy McKeown

11:50 *Towards Understanding of Medical Randomized Controlled Trials by Conclusion Generation*
Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su and Yun-Nung Chen

12:15 *Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text*
Hanna Berg, Taridzo Chomutare and Hercules Dalianis

**November 3, 2019 (continued)**

**12:40–14:00 Lunch**

**14:00–15:30 Session 3**

14:00 *Invited Talk*
TBA

14:40 *Automatic rubric-based content grading for clinical notes*
Wen-wai Yim, Ashley Mills, Harold Chun, Teresa Hashiguchi, Justin Yew and Bryan Lu

15:05 *Dilated LSTM with attention for Classification of Suicide Notes*
Annika M Schoene, George Lacey, Alexander P Turner and Nina Dethlefs

**15:30–16:00 Break**

**16:00–17:40 Session 4**

16:00 *Writing habits and telltale neighbors: analyzing clinical concept usage patterns with sublanguage embeddings*
Denis Newman-Griffis and Eric Fosler-Lussier

16:25 *Recognizing UMLS Semantic Types with Deep Learning*
Isar Nejadgholi, Kathleen C. Fraser, Berry De Bruijn, Muqun Li, Astha LaPlante and Khaldoun Zine El Abidine

16:50 *Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text*
Matus Falis, Maciej Pajak, Aneta Lisowska, Patrick Schrempf, Lucas Deckers, Shadia Mikhael, Sotirios Tsaftaris and Alison O'Neil

17:15 *Neural Token Representations and Negation and Speculation Scope Detection in Biomedical and General Domain Text*
Elena Sergeeva, Henghui Zhu, Amir Tahmasebi and Peter Szolovits