BioNLP 2020

The 19th SIGBioMed Workshop on Biomedical Language Processing

Proceedings of the Workshop

July 9, 2020

©2020 The Association for Computational Linguistics

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 978-1-952148-09-5

BioNLP 2020: Research unscathed by COVID-19

Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, Junichi Tsujii

The past year has been more than exciting for natural language processing in general, and for biomedical natural language processing in particular. A gradual accretion of studies of reproducibility and replicability in natural language processing, biomedical and otherwise had been making it clear that the reproducibility crisis that has hit most of the rest of science is not going to spare text mining or its related fields. Then, in March of 2020, much of the world ground to a sudden halt.

The outbreak of the COVID-19 disease caused by the novel coronavirus SARS-CoV-2 made computational work more obviously relevant than it had perhaps ever been before. Suddenly, newscasters were arguing about viral clades, the daily news was full of stories about modelling, and your neighbor had heard of PCR. But, some of us did not really see a role for natural language processing in the brave new world of computational instant reactions to an international pandemic.

That was wrong.

In mid-late March of 2020, a joint project between the Allen Artificial Intelligence Institute (Ai2), the National Library of Medicine (NLM), and the White House Office of Science and Technology Policy (OSTP) released CORD-19, a corpus of work on the SARS-CoV-2 virus, on COVID-19 disease, and on related coronavirus research. It was immediately notable for its inclusion of "gray literature" from preprint servers, which mostly have been neglected in text mining research, as well as for its flexibility with regards to licensing of content types. Perhaps most importantly, it was released in conjunction with a number of task types, including one related to ethics–although the value of medical ethics has been widely obvious since the Nazi "medical" experimentation horrors of the Second World War, the worldwide pandemic has made the value of medical **ethicists** more apparent to the general public than at any time since. Those task type definitions enabled the broader natural language processing community to jump into the fray quite quickly, and initial results have been quick to arrive.

Meanwhile, the pandemic did nothing to slow research in biomedical natural language processing on any other topic, either. That can be seen in the fact that this year the Association for Computational Linguistics SIGBIOMED workshop on biomedical natural language processing received 73 submissions. The unfortunate effect of the pandemic was the cancellation of the physical workshop, which would have allowed acceptance of all high-quality submissions as posters, if not for podium presentations. Indeed, the poster sessions at BioNLP have been continuously growing in size, due to the large number of high-quality submissions that the workshop receives annually. Unfortunately, because this year the Association for Computational Linguistics annual meeting will take place online only, there will be no poster session for the workshop. Consequently, only a handful of submissions could be accepted for presentation.

Transitioning of the traditional conferences to online presentations at the beginning of the COVID-19 pandemic showed that the traditional presentation formats are not as engaging remotely as they are in the context of in-person sessions. We are therefore exploring a new form of presentation, hoping it will be more engaging, interactive, and informative: 22 papers (about 30% of the submissions) will be presented in panel-like sessions. Papers will be grouped by similarity of topic, meaning that participants with related interests will be able to interact regarding their papers with a hopefully optimal number of people on line at the same time. As we write this introduction, the conference plans and platform are still evolving, as are the daily lives of much of the planet, so we hope that you will join us in planning for the worst, while hoping for the best.

Panel Discussions

papers referenced in this section are included in this volume, unless otherwise indicated

Session 1: High accuracy information retrieval, spin and bias

Invited talk and discussion lead: Kirk Roberts

Presentations: The exploration of Information Retrieval approaches enhanced with linguistic knowledge continues in the work that allows life-science researchers to search PubMed and the CORD-19 collection using patterns over dependency graphs (Taub-Tabib et al.) Representing biomedical relationships in the literature by encoding dependency structure with word embeddings promises to improve retrieval of relationships and literature-based discovery (Paullada et al.) Word embeddings trained on biomedical research articles and the tests based on their associations and coherence, among others, allow detecting and quantifying gender bias over time (Rios et al.) A BioBERT model fine-tuned for relation extraction might assist in detecting spin in reporting the results of randomized clinical trials (Koroleva et al.) Finally, a novel sequence-to-set approach to generating terms for pseudo-relevance feedback is evaluated (Das et al.)

Session 2: Clinical Language Processing

Invited talk and discussion lead: Tim Miller

Presentations: Not surprisingly, much of the potentially reproducible work in the clinical domain is based on the Medical Information Mart for Intensive Care (MIMIC) data (Johnson et al., 2016). Kovaleva et al. used the MIMIC-CXR data to explore Visual Dialog for radiology and prepare the first publicly available silver- and gold-standard datasets for this task. Searle et al. present a MIMIC-based silver standard for automatic clinical coding and warn that frequently assigned codes in MIMIC-III might be undercoded. Mascio et al. used MIMIC and the Shared Annotated Resources (ShARe)/CLEF dataset in four classification tasks: disease status, temporality, negation, and uncertainty. Temporality is explored in-depth by Lin et al., and Wang et al. explore approaches to a clinical Semantic Textual Similarity (STS) task. Xu et al. apply reinforcement learning to deal with noise in clinical text for readmission prediction after kidney transplant.

Session 3: Language Understanding

Invited talk and discussion lead: Anna Rumshisky

Presentations: Bringing clinical problems and poetry together, this creative work seeks to better understand dyslexia through a self-attention transformer and Shakespearean sonnets (Bleiweiss). Detection of early stages of Alzheimer's disease using unsupervised clustering is explored with 10 years of President Ronald Reagan's speeches (Wang et al.). Stavropoulos et al. introduce BIOMRC, a cloze-style dataset for biomedical machine reading comprehension, along with new publicly available models, and provide a leaderboard for the task. Another type of question answering – answering questions that can be answered by electronic medical records–is explored by Rawat et al. Hur et al. study veterinary records to identify reasons for administration of antibiotics. DeYoung et al. expand the Evidence Inference dataset and evaluate BERT-based models for the evidence inference task.

Session 4: Named Entity Recognition and Knowledge Representation

Invited talk and discussion lead: Hoifung Poon

Invited talk: Machine Reading for Precision Medicine

The advent of big data promises to revolutionize medicine by making it more personalized and effective, but big data also presents a grand challenge of information overload. For example, tumor sequencing has become routine in cancer treatment, yet interpreting the genomic data requires painstakingly curating knowledge from a vast biomedical literature, which grows by thousands of papers every day. Electronic medical records contain valuable information to speed up clinical trial recruitment and drug development, but curating such real-world evidence from clinical notes can take hours for a single patient. Natural language processing (NLP) can play a key role in interpreting big data for precision medicine. In particular, machine reading can help unlock knowledge from text by substantially improving curation efficiency. However, standard supervised methods require labeled examples, which are expensive and time-consuming to produce at scale. In this talk, Dr. Poon presents Project Hanover, where the team overcomes the annotation bottleneck by combining deep learning with probabilistic logic, and by exploiting self-supervision from readily available resources such as ontologies and databases. This enables the researchers to extract knowledge from millions of publications, reason efficiently with the resulting knowledge graph by learning neural embeddings of biomedical entities and relations, and apply the extracted knowledge and learned embeddings to supporting precision oncology.

Hoifung Poon is the Senior Director of Precision Health NLP at Microsoft Research and an affiliated professor at the University of Washington Medical School. He leads Project Hanover, with the overarching goal of structuring medical data for precision medicine. He has given tutorials on this topic at top conferences such as the Association for Computational Linguistics (ACL) and the Association for the Advancement of Artificial Intelligence (AAAI). His research spans a wide range of problems in machine learning and natural language processing (NLP), and his prior work has been recognized with Best Paper Awards from premier venues such as the North American Chapter of the Association for Computational Linguistics (NAACL), Empirical Methods in Natural Language Processing (EMNLP), and Uncertainty in AI (UAI). He received his PhD in Computer Science and Engineering from University of Washington, specializing in machine learning and NLP.

Presentations: Nejadgholi et al. analyze errors in NER and introduce an F-score that models a forgiving user experience. Peng et al. study NER, relation extraction, and other tasks with a multi-tasking learning approach. Amin et al. explore multi-instance learning for relation extraction. ShafieiBavani et al. also explore relation and event extraction, but in the context of simultaneously predicting relationships between all mention pairs in a text. Chang et al. provide a benchmark for knowledge graph embedding models on the SNOMED-CT knowledge graph and emphasize the importance of knowledge graphs for learning biomedical knowledge representation.

Acknowledging the community

As always, we are profoundly grateful to the authors who chose BioNLP for presenting their innovative research. The authors' willingness to continue sharing their work through BioNLP consistently makes the workshop noteworthy and stimulating. We are equally indebted to the program committee members (listed elsewhere in this volume) who produced thorough reviews on a tight review schedule and with an admirable level of insight, despite the timeline being even shorter than usual and the workload higher, while at the same time handling the unprecedented changes in their work and life caused by the COVID-19 pandemic.

References

Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35

Wang, L.L., Lo,K., Chandrasekhar, Y. et al. Cord-19: The covid-19 open research dataset. ArXiv, abs/2004.10706, 2020.

Organizers:

Dina Demner-Fushman, US National Library of Medicine
Kevin Bretonnel Cohen, University of Colorado School of Medicine, USA
Sophia Ananiadou, National Centre for Text Mining and University of Manchester, UK
Junichi Tsujii, National Institute of Advanced Industrial Science and Technology, Japan

Program Committee:

Sophia Ananiadou, National Centre for Text Mining and University of Manchester, UK
Emilia Apostolova, Language.ai, USA
Eiji Aramaki, University of Tokyo, Japan
Asma Ben Abacha, US National Library of Medicine
Siamak Barzegar, Barcelona Supercomputing Center, Spain
Olivier Bodenreider, US National Library of Medicine
Leonardo Campillos Llanos, Universidad Autonoma de Madrid, Spain
Qingyu Chen, US National Library of Medicine
Fenia Christopoulou, National Centre for Text Mining and University of Manchester, UK
Aaron Cohen, Oregon Health & Science University, USA
Kevin Bretonnel Cohen, University of Colorado School of Medicine, USA
Brian Connolly, Kroger Digital, USA
Viviana Cotik, University of Buenos Aires, Argentina
Manirupa Das, Amazon Search, Seattle, WA, USA
Dina Demner-Fushman, US National Library of Medicine
Bart Desmet, Clinical Center, National Institutes of Health, USA
Travis Goodwin, , US National Library of Medicine
Natalia Grabar, CNRS, France
Cyril Grouin, LIMSI - CNRS, France
Tudor Groza, The Garvan Institute of Medical Research, Australia
Antonio Jimeno Yepes, IBM, Melbourne Area, Australia
Halil Kilicoglu, University of Illinois at Urbana-Champaign, USA
Ari Klein, University of Pennsylvania, USA
Andre Lamurias, University of Lisbon, Portugal
Majid Latifi, Trinity College Dublin, Ireland
Alberto Lavelli, FBK-ICT, Italy
Robert Leaman, US National Library of Medicine
Ulf Leser, Humboldt-Universität zu Berlin, Germany
Maolin Li, National Centre for Text Mining and University of Manchester, UK
Zhiyong Lu, US National Library of Medicine
Timothy Miller, Children's Hospital Boston, USA
Claire Nedellec, INRA, France
Aurelie Neveol, LIMSI - CNRS, France
Mariana Neves, German Federal Institute for Risk Assessment, Germany
Denis Newman-Griffis, Clinical Center, National Institutes of Health, USA
Nhung Nguyen, The University of Manchester, UK
Karen O'Connor, University of Pennsylvania, USA
Naoaki Okazaki, Tokyo Institute of Technology, Japan
Yifan Peng, US National Library of Medicine
Laura Plaza, UNED, Madrid, Spain
Francisco J. Ribadas-Pena, University of Vigo, Spain
Angus Roberts, The University of Sheffield, UK
Kirk Roberts, The University of Texas Health Science Center at Houston, USA

Roland Roller, DFKI GmbH, Berlin, Germany
Diana Sousa, University of Lisbon, Portugal
Karin Verspoor, The University of Melbourne, Australia
Davy Weissenbacher, University of Pennsylvania, USA
W John Wilbur, US National Library of Medicine
Shankai Yan, US National Library of Medicine
Chrysoula Zerva, National Centre for Text Mining and University of Manchester, UK
Ayah Zirikly, Clinical Center, National Institutes of Health, USA
Pierre Zweigenbaum, LIMSI - CNRS, France

Additional Reviewers:

Jingcheng Du, School of Biomedical Informatics, UTHealth

Invited Speakers:

Hoifung Poon, Microsoft Research
Tim Miller, Boston Childrens Hospital and Harvard Medical School
Kirk Roberts, School of Biomedical Informatics, UTHealth
Anna Rumshisky, University of Massachusetts Lowell

Table of Contents

Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings
Anthony Rios, Reenam Joshi and Hejin Shin 1

Sequence-to-Set Semantic Tagging for Complex Query Reformulation and Automated Text Categorization in Biomedical IR using Self-Attention
Manirupa Das, Juanxi Li, Eric Fosler-Lussier, Simon Lin, Steve Rust, Yungui Huang and Rajiv Ramnath 14

Interactive Extractive Search over Biomedical Corpora
Hillel Taub Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen and Yoav Goldberg 28

Improving Biomedical Analogical Retrieval with Embedding of Structural Dependencies
Amandalynne Paullada, Bethany Percha and Trevor Cohen 38

DeSpin: a prototype system for detecting spin in biomedical publications
Anna Koroleva, Sanjay Kamath, Patrick Bossuyt and Patrick Paroubek 49

Towards Visual Dialog for Radiology
Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer Beymer, Anna Rumshisky and Vandana Mukherjee Mukherjee 60

A BERT-based One-Pass Multi-Task Model for Clinical Temporal Relation Extraction
Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard and Guergana Savova 70

Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset
Thomas Searle, Zina Ibrahim and Richard Dobson 76

Comparative Analysis of Text Classification Approaches in Electronic Health Records
Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan and Angus Roberts 86

Noise Pollution in Hospital Readmission Prediction: Long Document Classification with Reinforcement Learning
Liyan Xu, Julien Hogan, Rachel E. Patzer and Jinho D. Choi 95

Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity
Yuxia Wang, Fei Liu, Karin Verspoor and Timothy Baldwin 105

Entity-Enriched Neural Models for Clinical Question Answering
Bhanu Pratap Singh Rawat, Wei-Hung Weng, So Yeon Min, Preethi Raghavan and Peter Szolovits 112

Evidence Inference 2.0: More Data, Better Models
Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall and Byron C. Wallace 123

Personalized Early Stage Alzheimer's Disease Detection: A Case Study of President Reagan's Speeches
Ning Wang, Fan Luo, Vishal Peddagangireddy, Koduvayur Subbalakshmi and Rajarathnam Chandramouli 133

BioMRC: A Dataset for Biomedical Machine Reading Comprehension
Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos and Ryan McDonald 140

Neural Transduction of Letter Position Dyslexia using an Anagram Matrix Representation
Avi Bleiweiss 150

Domain Adaptation and Instance Selection for Disease Syndrome Classification over Veterinary Clinical Notes
Brian Hur, Timothy Baldwin, Karin Verspoor, Laura Hardefeldt and James Gilkerson 156

Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings
David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt and Andrew Taylor 167

Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience
Isar Nejadgholi, Kathleen C. Fraser and Berry de Bruijn 177

A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction
Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva and Guenter Neumann 187

Global Locality in Biomedical Relation and Event Extraction
Elaheh ShafieiBavani, Antonio Jimeno Yepes, Xu Zhong and David Martinez Iraola 195

An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining
Yifan Peng, Qingyu Chen and Zhiyong Lu 205

Conference Program

Thursday July 9, 2020

08:30–08:40 Opening remarks

08:40–10:30 Session 1: High accuracy information retrieval, spin and bias

08:40–09:10 ***Invited Talk – Kirk Roberts***

09:10–09:20 *Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings*
Anthony Rios, Reenam Joshi and Hejin Shin

09:20–09:30 *Sequence-to-Set Semantic Tagging for Complex Query Reformulation and Automated Text Categorization in Biomedical IR using Self-Attention*
Manirupa Das, Juanxi Li, Eric Fosler-Lussier, Simon Lin, Steve Rust, Yungui Huang and Rajiv Ramnath

09:30–09:40 *Interactive Extractive Search over Biomedical Corpora*
Hillel Taub Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen and Yoav Goldberg

09:40–09:50 *Improving Biomedical Analogical Retrieval with Embedding of Structural Dependencies*
Amandalynne Paullada, Bethany Percha and Trevor Cohen

09:50–10:00 *DeSpin: a prototype system for detecting spin in biomedical publications*
Anna Koroleva, Sanjay Kamath, Patrick Bossuyt and Patrick Paroubek

10:00–10:30 ***Discussion***

10:30–10:45 ***Coffee Break***

Thursday July 9, 2020 (continued)

10:45–13:00 Session 2: Clinical Language Processing

10:45–11:15 ***Invited Talk – Tim Miller***

11:15–11:25 *Towards Visual Dialog for Radiology*
Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer Beymer, Anna Rumshisky and Vandana Mukherjee Mukherjee

11:25–11:35 *A BERT-based One-Pass Multi-Task Model for Clinical Temporal Relation Extraction*
Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard and Guergana Savova

11:35–11:45 *Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset*
Thomas Searle, Zina Ibrahim and Richard Dobson

11:45–11:55 *Comparative Analysis of Text Classification Approaches in Electronic Health Records*
Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan and Angus Roberts

11:55–12:05 *Noise Pollution in Hospital Readmission Prediction: Long Document Classification with Reinforcement Learning*
Liyan Xu, Julien Hogan, Rachel E. Patzer and Jinho D. Choi

12:05–12:15 *Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity*
Yuxia Wang, Fei Liu, Karin Verspoor and Timothy Baldwin

12:15–12:45 ***Discussion***

12:45–13:30 ***Lunch***

Thursday July 9, 2020 (continued)

13:30–15:30 Session 3: Language Understanding

13:30–14:00 ***Invited Talk – Anna Rumshisky***

14:00–14:10 *Entity-Enriched Neural Models for Clinical Question Answering*
Bhanu Pratap Singh Rawat, Wei-Hung Weng, So Yeon Min, Preethi Raghavan and Peter Szolovits

14:10–14:20 *Evidence Inference 2.0: More Data, Better Models*
Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall and Byron C. Wallace

14:20–14:30 *Personalized Early Stage Alzheimer's Disease Detection: A Case Study of President Reagan's Speeches*
Ning Wang, Fan Luo, Vishal Peddagangireddy, Koduvayur Subbalakshmi and Rajarathnam Chandramouli

14:30–14:40 *BioMRC: A Dataset for Biomedical Machine Reading Comprehension*
Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos and Ryan McDonald

14:40–14:50 *Neural Transduction of Letter Position Dyslexia using an Anagram Matrix Representation*
Avi Bleiweiss

14:50–15:00 *Domain Adaptation and Instance Selection for Disease Syndrome Classification over Veterinary Clinical Notes*
Brian Hur, Timothy Baldwin, Karin Verspoor, Laura Hardefeldt and James Gilkerson

15:00–15:30 ***Discussion***

15:30–15:45 ***Coffee Break***

Thursday July 9, 2020 (continued)

15:45–17:45 Session 4: Named Entity Recognition and Knowledge Representation

15:45–16:25 ***Invited Talk: Machine Reading for Precision Medicine, Hoifung Poon***

16:25–16:35 *Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings*
David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt and Andrew Taylor

16:35–16:45 *Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience*
Isar Nejadgholi, Kathleen C. Fraser and Berry de Bruijn

16:45–16:55 *A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction*
Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva and Guenter Neumann

16:55–17:05 *Global Locality in Biomedical Relation and Event Extraction*
Elaheh ShafieiBavani, Antonio Jimeno Yepes, Xu Zhong and David Martinez Iraola

17:05–17:15 *An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining*
Yifan Peng, Qingyu Chen and Zhiyong Lu

17:15–17:45 ***Discussion***

17:45–18:00 Closing remarks