BioNLP 2023

BioNLP and BioNLP-ST

Proceedings of the Workshop

July 13, 2023

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 978-1-959429-85-2

BioNLP Year in Review: The Year of LLMs in the News

[11pt,a4paper]article [utf8]inputenc [T1]fontenc times

BioNLP Year in Review: The Year of LLMs in the News

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

While biomedical language processing dates back to the 1960th, and pretrained language models have been around for at least a decade, this year, Large Language Models (LLMs) applied to complex Biomedical Language Processing tasks have started to make the news regularly. The opportunities and threats of General Purpose Artificial Intelligence (AGI) have been discussed worldwide[1], and the potential pitfalls have been acknowledged in the technical reports that present the newer and larger models[2,3]. The discussions on the role of LLMs in health have followed suit. Assumptions were made that LLMs will engage with patients, which will subsequently improve outcomes, or, to quote the authors, LLM will save lives.The article explains that while traditional models predicted disease deterioration or helped to detect cancer, LLMs can be used for those tasks formerly done by more traditional models and yes, those foundation models might have better predictive accuracy and work with multimodal data, but in healthcare the major impact will be that we are about to change the way human-AI interaction will work."[4].

Two observations in the above statement warrant further discussion: 1) that traditional models might be better suited for some tasks, e.g., predicting patients' trajectories; and 2) that chatbots powered by LLMs might finally engage with patients. The first assumption warrants further exploration of different approaches to traditional BioNLP topics. We are therefore happy to present in this volume a range of traditional tasks approached using both traditional and new models. The second assumption is supported both by the well-documented tendency of users to engage with and trust online resources, and by the fundamental qualitative improvement in the text generated by LLMs, which meets all but one evaluation criteria. Although this text is generally grammatical, cohesive, topically relevant, and plausible, its Achilles heel is factuality, or rather lack thereof, and the potential harm that this can cause to patients. These pitfalls must be considered before providing chatbot-powered services[5,6].

> National Eating Disorders Association Phases Out Human Helpline, Pivots to Chatbot
> `May 31, 2023`
>
> `(NPR) - For more than 20 years, the National Eating Disorders Association (NEDA) has operated a phone line and online platform for people seeking help with anorexia, bulimia, and other eating disorders. Last year, nearly 70,000 individuals used the helpline. NEDA shuttered that service in May. Instead, the non-profit will use a chatbot called Tessa that was designed by eating disorder experts, with funding from NEDA.`
>
> https://bioethics.com/archives/70493

> Eating disorder helpline shuts down AI chatbot that gave bad advice
> JUNE 1, 2023
>
> An AI-powered chatbot that replaced employees at an eating disorder hotline has been shut down after it provided harmful advice to people seeking help. Every single thing Tessa suggested were things that led to the development of my eating disorder,wrote Sharon Maxwell, who describes herself as a weight inclusive consultant and fat activist, on Instagram. This robot causes harm."
> In an Instagram post on Wednesday, NEDA announced it was shutting down Tessa, at least temporarily.
> https://www.cbsnews.com/news/eating-disorder-helpline-chatbot-disabled/

An example of the many media discussions surrounding the tremendous promise and potential harms associated with use of LLMs for health-related purposes is shown above.

Although ChatGPT and other LLMs outperform traditional neural networks in many tasks, and in certain cases provide reliable reasoning and explanations, evaluation of their performance is needed in different scenarios. Prompt engineering and new evaluation metrics are needed to fully understand LLMs limitations. There are several potential harms associated with LLMs (still to be fully examined); they include, but are not limited to discrimination, misinformation, and harms from human-computer interaction, e.g., conversational agents engaging with patients in mental health care. The weaknesses of LLMs underline the pressing need to carry out further research into fact-checking, misinformation detection, bias detection, and explainability, as well as the requirement to develop reliable and feasible evaluation approaches. A number of results pertaining to these important areas of ongoing research will be presented at BioNLP 2023.

Work presented in this volume

BioNLP 2023 received 59 valid submissions, of which 11 were accepted as oral presentations and 34 as posters. Two shared tasks were co-located with the workshop, each focused on the summarization of a different text type, i.e., clinical notes and biomedical literature. Overviews of these tasks are presented in this volume. For clinical notes, subtasks consisted of generating lists of diagnoses and problems using patients' daily care plans, and generating summaries of radiology reports that highlight key observations and conclusions. For biomedical literature, the focus was on Readability-controlled Summarization", i.e., the ability to generate different summaries of research articles that are aimed at either a technical or a layaudience. Given the increased appearance of NLP in the news, as highlighted above, plain language summaries are becoming more important, to make research results accessible to a non-expert audience.

Keynote

Dementia Detection from Speech: New Developments and Future Directions

Presented by Kathleen Fraser

Abstract:
Diagnosing and treating dementia is a pressing concern as the global population ages. A growing number of publications in NLP tackle the question of whether we can use speech and language analysis to automatically detect signs of this devastating disease. However, the field of NLP has changed rapidly since the task was first proposed. In this talk, Dr. Kathleen Fraser will summarize the foundational approaches to dementia detection from speech, and then review how current approaches are building on and improving over the earlier work. Dr. Fraser will present several areas that she believes are promising future directions, and discuss preliminary work from her group specifically on the topic of multimodal machine learning for remote cognitive assessment.

Bio:
Dr. Kathleen Fraser is a computer scientist in the Digital Technologies Research Centre at the National Research Council Canada. Her research focuses on the use of natural language processing (NLP) in healthcare applications, as well as assessing and mitigating social bias in artificial intelligence systems. Dr. Fraser received her PhD in computer science from the University of Toronto in 2016, and subsequently completed a post-doc at the University of Gothenburg, Sweden. She was named an MIT Rising Star in Electrical Engineering and Computer Science, and was awarded the Governor General's Gold Academic Medal in 2017. She also co-founded the start-up Winterlight Labs, later acquired by Cambridge Cognition. She has been a research officer at the National Research Council since 2018 and also holds a position as adjunct professor at Carleton University.

Acknowledging the community

As always, we are deeply grateful to the authors of the submitted papers and to the reviewers (listed elsewhere in this volume) who produced three thorough and thoughtful reviews for each paper in a fairly short review period. The quality of submitted work continues to grow, and the organizers are truly grateful to the members of our amazing Program Committee, who helped us to determine which work was ready to be presented, and which would benefit from the additional experiments and analyses suggested by the reviewers.

Our special thanks to the Shared Tasks organizers and participants. The datasets and approaches generated in these community-wide evaluations are bound to advance the state-of-the-art for these essential tasks.

We are particularly grateful to Rich Gerber, Softconf/Start Conference Manager, for the extensive and timely support, and to Paul Thompson, National Centre for Text Mining, The University of Manchester, for supporting the production of this volume and extensive editing.

As in years past, we are looking forward to a productive workshop, which we hope will lead to new collaborations and research, thus allowing our community to continue developing their valuable contributions to public health and well-being.

References

1. EU Parliament news. Artificial intelligence: threats and opportunities

   https://www.europarl.europa.eu/news/en/headlines/society/20200918STO87404/artificial-intelligence-threats-and-opportunities

2. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375. 2023 Mar 20.

3. Singhal et al. Towards Expert-Level Medical Question Answering with Large Language Models. https://doi.org/10.48550/arXiv.2305.09617

4. Lutz Finger. Large Language Models & AI In Healthcare.

   https://www.forbes.com/sites/lutzfinger/2023/03/27/large-language-models–ai-in-healthcare/?sh=6de65700598b

5. Tristan Bove. Sam Altman and other technologists warn that A.I. poses a ‘risk of extinction’ on par with pandemics and nuclear warfare. https://fortune.com/2023/05/30/sam-altman-ai-risk-of-extinction-pandemics-nuclear-warfare/

6. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, Cheng M, Glaese M, Balle B, Kasirzadeh A, Kenton Z. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359. 2021 Dec 8.