Applications of Natural Language Processing in Clinical Research and Practice

Yanshan Wang
Mayo Clinic
Rochester, MN

Ahmad P. Tafti
Mayo Clinic
Rochester, MN

Sunghwan Sohn
Mayo Clinic
Rochester, MN

Rui Zhang
University of Minnesota
Minneapolis, MN

1 Tutorial Overview

Rapid growth in adoption of electronic health records (EHRs) has led to an unprecedented expansion in the availability of large longitudinal datasets. Large initiatives such as the Electronic Medical Records and Genomics (eMERGE) Network (Lemke et al., 2010), the Patient-Centered Outcomes Research Network (PCORNet) (Fleurence et al., 2014), and the Observational Health Data Science and Informatics (OHDSI) consortium (Hripcsak et al., 2015), have been established and have reported successful applications of secondary use of EHRs in clinical research and practice. In these applications, natural language processing (NLP) technologies have played a crucial role as much of detailed patient information in EHRs is embedded in narrative clinical documents. Meanwhile, a number of clinical NLP systems, such as MedLEE (Friedman et al., 1994), MetaMap/MetaMap Lite (Aronson and Lang, 2010), cTAKES (Savova et al., 2010), and MedTagger (Liu et al., 2013) have been developed and utilized to extract useful information from diverse types of clinical text, such as clinical notes, radiology reports, and pathology reports. Success stories in applying these tools have been reported widely (Wang et al., 2017).

Despite the demonstrated success of NLP in the clinical domain, methodologies and tools developed for the clinical NLP are still underknown and underutilized by students and experts in the general NLP domain, mainly due to the limited exposure to EHR data. Through this tutorial, we would like to introduce NLP methodologies and tools developed in the clinical domain, and showcase the real-world NLP applications in clinical research and practice at Mayo Clinic (the No. 1 national hospital ranked by the U.S. News & World Report and the No.1 hospital in the world by the Newsweek) and the University of Minnesota (the No. 41 best global universities ranked by the U.S. News & World Report). We will review NLP techniques in solving clinical problems and facilitating clinical research, the state-of-the art clinical NLP tools, and share collaboration experience with clinicians, as well as publicly available EHR data and medical resources, and finally conclude the tutorial with vast opportunities and challenges of clinical NLP. The tutorial will provide an overview of clinical backgrounds, and does not presume knowledge in medicine or health care. The goal of this tutorial is to encourage NLP researchers in the general domain (as opposed to the specialized clinical domain) to contribute to this burgeoning area.

In this tutorial, we will first present an overview of clinical NLP. We will then dive into two subareas of clinical NLP in clinical research, including big data infrastructure for large-scale clinical NLP and advances of NLP in clinical research, and two subareas in clinical practice, including clinical information extraction and patient cohort retrieval using EHRs. Around 70% of the tutorial will review clinical problems, cutting-edge methodologies, and public clinical NLP tools while another 30% introduce real-world clinical use cases at Mayo Clinic and the University of Minnesota. Finally, we will conclude the tutorial with challenges and opportunities in this rapidly developing domain.

2 Type of the tutorial

Introductory.

3 Outline

1. Introduction: Overview of Clinical NLP (10 minutes, Dr. Wang)
2. Big Data Infrastructure for Large-scale Clinical NLP (40 minutes, Dr. Tafti)
    - Motivation

Proceedings of NAACL HLT 2019: Tutorial Abstracts, pages 22–25
Minneapolis, Minnesota, June 2, 2019.

- Big data NLP: hope and hype
- Tools for big data NLP
- Case study: indexing Tweets data and health-related social media blog posts to trend analysis of cancer treatment strategies

3. Advances of NLP in Clinical Research (40 minutes, Dr. Zhang)
    - Motivation
    - Background of NLP to support clinical research
    - NLP Methodologies and tools for clinical research
    - Case study 1: family history information extraction
    - Case study 2: identifying use status of dietary supplements
4. Clinical Information Extraction: Methodologies and Tools (40 minutes, Dr. Sohn)
    - Motivation
    - Background of clinical information extraction
    - Methodology review: rule-based or machine learning/deep learning?
    - Tools and frameworks: UIMA framework, cTAKES, and MedTagger
    - Case study: ascertainment of asthma status using free-text EHRs
5. Patient Cohort Retrieval using EHRs (40 minutes, Dr. Wang)
    - Motivation
    - Background of patient cohort retrieval
    - Methodology: extraction of medical concepts, information retrieval for cohort identification
    - Case study 1: Patient cohort retrieval for epidemiology study
    - Case study 2: Patient cohort retrieval for clinical trials accrual
6. Clinical NLP: Challenges and Opportunities (10 minutes, Dr. Wang)
    - Challenges in methodology and practical applications
    - Opportunities for NLP in clinical research and practice

4 Instructors

Yanshan Wang is a Research Associate at Mayo Clinic. His current work is centered on developing novel NLP and artificial intelligence (AI) methodologies for facilitating clinical research and solving real-world clinical problems. Since he joined Mayo Clinic in 2015, he has been leading several NIH-funded projects, which aims to leverage and develop novel NLP techniques to automatically retrieve cohorts from clinical data repository using free-text EHR data. Dr. Wang has extensive collaborative research experience with physicians, epidemiology researchers, statisticians, NLP researchers, and IT technicians. He collaborated with rheumatologists and developed a NLP system for automatic identification of skeletal site-specific fractures from radiology reports for osteoporosis patients. He has had ongoing collaboration with epidemiologists and clinical neurologists on developing novel AI solutions to provide better care for elders. Dr. Wang has published over 40 peer-reviewed articles at referred computational linguistic conferences (e.g., NAACL), and medical informatics journals and conference (e.g., JBI, JAMIA, JMIR and AMIA). He has served on program committees for EMNLP, NAACL, IEEE-ICHI, IEEE-BIBM, etc. (wang.yanshan@mayo.edu)

Ahmad P. Tafti is a Research Associate at Mayo Clinic, with a deep passion for improving health informatics using diverse medical data sources combined with advanced computational methods. Dr. Tafti's major interests are AI, machine learning, and computational health informatics. He completed his PhD in Computer Science at University of Wisconsin-Milwaukee, and some part of his international studies were carried out at Oracle Education Center, Technical University of Vienna, and Medical University of Vienna, Austria. He won the General Electric Honorable Mention Award and received the 3rd place in the Larry Hause Student Poster Competition at an IEEE conference as part of his PhD project. Dr. Tafti has published over 20 first-author peer-reviewed publications in prestigious journals and conferences (e.g., CVPR, AMIA, ISVC, JMIR, PLOS, IEEE Big Data), addressing medical text and medical image analysis and understanding using advanced computational strategies. In addition, Dr. Tafti has served as a workshop organizer, steering committee member, technical reviewer, and a

program committee member for several reputable conferences and journals, including KDD 2017, AMIA, IEEE ICHI, ISMCO, ISVC, IEEE Journal of Biomedical and Health Informatics, and International Journal of Computer Vision and Image Processing. He was awarded a NVIDIA GPU Grant for his accomplishments in deep learning community. (tafti.ahmad@mayo.edu)

Sunghwan Sohn is an Associate Professor of Biomedical Informatics at Mayo Clinic. He has expertise in mining large-scale EHRs to unlock unstructured and hidden information using natural language processing and machine learning, thus creating new capacities for clinical research and practice in order to achieve better patient solutions. He has been involved in the development of cTAKES, the most popular NLP tool in the clinical domain. Dr. Sohns research facilitates the best use of EHRs to solve clinical problems and improve public health. His work provides biomedical scientists and clinicians access to unstructured information from clinical narratives and clinical text analytics necessary for clinical research and patient care. Dr. Sohns research goal is the best utilization of informatics to facilitate translational research and precision medicine across heterogeneous EHR data and systems in a large population. (sohn.sunghwan@mayo.edu)

Rui Zhang is an Assistant Professor in the College of Pharmacy and the Institute for Health Informatics (IHI), and also graduate faculty in Data Science at the University of Minnesota (UMN). He is the Leader of NLP Services in Clinical and Transnational Science Institution (CTSI) at the UMN. Dr. Zhangs research focuses on health and biomedical informatics, especially biomedical NLP and text mining. His research interests include the secondly analysis of EHR data for patient care as well as pharmacovigilance knowledge discovery through mining biomedical literature. His researcher program is funded by federal agencies with over 3.5 million dollars including National Institutes of Health, the Agency for Health and Research Quality (AHRQ), and a medical device industry - Medtronic Inc. He also a co-investigator of a 42.6 million of CTSI grant. His work has been recognized on a national scale including Journal of Biomedical Informatics Editors Choice, nominated for Distinguished paper in AMIA Annual Symposium and Marco Ramoni Distinguished Paper Award for Translational Bioinformatics, as well as highlighted by The Wall Street Journal. (zhan1386@umn.edu)

Audience, Previous Tutorials and Venue

Based on the recent upsurge of interest in applications of NLP in the clinical domain, we target an audience of 60 to 100 students and researchers from both academia and industry. We are not aware of any recent tutorial on the topic of clinical NLP. No technical equipment is required. Since Mayo Clinic is located at Rochester, MN and the University of Minnesota is located at Minneapolis, MN, our preference for the venue is NAACL 2019 at Minneapolis, MN.

Acknowledgement

This tutorial has been made possible by partial support from the National Center for Advancing Translational Sciences (NCATS) Open Health Natural Language Processing (OHNLP) Consortium (U01TR002062).

References

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17(3):229–236.

Rachael L Fleurence, Lesley H Curtis, Robert M Califf, Richard Platt, Joe V Selby, and Jeffrey S Brown. 2014. Launching pcornet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association* 21(4):578–582.

Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association* 1(2):161–174.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. 2015. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics* 216:574.

Amy A Lemke, Joel T Wu, Carol Waudby, Jill Pulley, Carol P Somkin, and Susan Brown Trinidad. 2010. Community engagement in biobanking: Experiences from the emerge network. *Genomics, Society and Policy* 6(3):50.

Hongfang Liu, Suzette J Bielinski, Sunghwan Sohn, Sean Murphy, Kavishwar B Wagholikar, Siddhartha R Jonnalagadda, KE Ravikumar, Stephen T

Wu, Iftikhar J Kullo, and Christopher G Chute. 2013. An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings* 2013:149.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17(5):507–513.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2017. Clinical information extraction applications: a literature review. *Journal of biomedical informatics* .