Natural Language Processing Working Group Pre-Symposium: Graduate Student Consortium, Year-in-Review, and Community Shared Tasks

Hongfang Liu, PhD[1], Ozlem Uzuner, PhD[2], Sivaram Arabandi, MD, MS[3], Dina Demner-Fushman, MD, PhD[4], Stephane Meystre, MD, PhD[5], Jon Patrick, PhD[6], Guergana Savova, PhD[7], Kavishwar Wagholikar, MBBS, PhD[8], Chunhua Weng, PhD[9], Hua Xu, PhD[10], Rong Xu, PhD[11], Meliha Yetisgen, PhD[12], Pierre Zweigenbaum, PhD[13]

[1]Mayo Clinic, Rochester, MN, [2]George Mason University, Fairfax, Virginia, [3]Ontopro, Houston, TX, [4]National Library of Medicine, Bethesda, MD, [5]Medical University of South Carolina, Charleston, SC, [6]Health Language Analytics, Sydney, Australia, [7]Boston Children's Hospital, Harvard Medical School, Boston, MA, [8]Massachusetts General Hospital, Boston, MA, [9]Columbia University, New York, NY, [10]University of Texas Health Science Center at Houston, TX, [11]Case Western Reserve University, Cleveland, OH, [12]University of Washington, Seattle, WA, [13]LIMSI, CNRS, Université Paris-Saclay, Orsay, France

Abstract:
The application of Natural Language Processing (NLP) methods and resources to clinical and biomedical text has received growing attention over the past years, but progress has been constrained by difficulties to access shared tools and resources, partially caused by patient privacy and data confidentiality constraints. Efforts to increase sharing and interoperability of the few existing resources are needed to facilitate the progress observed in the general NLP domain. To address this need, the AMIA NLP working group pre-symposium continues the tradition since its inception in 2012 to provide a unique platform for close interactions among students, scholars, and industry professionals who are interested in biomedical NLP. The event will consist of three sections: 1) a graduate student consortium, where students can present their work and get feedback from experienced researchers in the field; 2) a year-in-review session, where significant NLP articles in the biomedical domain will be presented followed by a panel discussion; and 3) a NLP community challenge session. This session will invite organizers and participants of NLP challenges in the biomedical domain to present the design, implementation, and results of the challenges.

Introduction:
The application of Natural Language Processing (NLP) in the general English domain has seen enormous progress over the past decades, progress that was enabled by the large availability of tools and resources that could be shared, reused, and improved in multiple projects and collaborations. Applying NLP to the textual content of patient electronic health records (i.e., clinical text) is constrained by strict patient privacy and confidentiality laws and regulations. These domain-specific peculiarities for access and sharing of resources (e.g., annotated text corpora) and tools (e.g., trained machine learning algorithms) require creative solutions. Despite these privacy restrictions, many research teams have succeeded in creating and then sharing resources based on clinical text in a thoughtful and sustainable manner. Despite the legal specifics surrounding patient data, NLP-based technologies have been permeating clinical and translational research. Our general objective with this pre-symposium is (1) to provide a platform for the next generation of biomedical NLP scientists to get focused feedback on their in-progress graduate work from a panel of senior academicians, (2) to demonstrate the latest achievements, resources and tools within the biomedical NLP community along with their reusability, portability, and interoperability.

The pre-symposium will be divided into three 2-hour sessions, with a graduate student consortium (i.e., 'doctoral' consortium also open to Masters students), where students will present their in-progress graduate work and get feedback from experienced researchers in the field, followed by NLP year-in-review session, where notable NLP

publications, events, and methodology breakthroughs within the past year will be highlighted and discussed, and then a community shared task session where organizers of various NLP shared tasks will provide summary followed by the practical implementation and testing of systems used in those events. The allocation of time will be adjusted based on the number of attendees and the number of submissions received. We also plan to invite poster submissions to provide an additional platform to a larger group of participants to present their research work and receive feedback.

Session 1 – Graduate student consortium and poster session (120 minutes)
A selection of students will present their work and get feedback from experienced researchers in the field. A call for submissions will invite graduate students to submit applications for a podium presentation of their graduate research work (in the biomedical NLP fields). Four selected students will each have 15 minutes for presentation and 15 minutes for discussion with a panel of ten established experts and researchers in biomedical NLP (Hongfang Liu, Rong Xu, Jon Patrick, Guergana Savova, Chunhua Weng, Pierre Zweigenbaum, Dina Demner-Fushman, Ozlem Uzuner, Hua Xu, Stephane Meystre).

The purpose of the graduate student consortium is to provide opportunities for direct interactions between students and researchers in biomedical NLP, so that students can 1) refine their research focus; 2) discuss specific questions about study design, algorithm development, or evaluation plan; 3) receive constructive feedback and suggestions about their dissertation work; and 4) establish possible collaborations. Informal feedback from students at previous years' doctoral consortiums was very positive; they appreciated the help offered in these four topics.

Session 2 – Year-in-review (120 minutes)
A selection of existing and significant NLP related efforts will be highlighted in this session. A call for submissions will invite researchers to submit their significant research effort including those published, in press or under development projects in the past 12 months. The presenter should identify themselves as the corresponding author during the submission process, and accepted presenters are required to make the presentation themselves.

All NLP year-in-review submissions will be evaluated by according to the following criteria:

- Relevance, interest, and value of the topic to NLP-WG,
- Impact of the paper(s) on informatics/medicine/biology (while the impact of papers on science is not fully reflected by ISI/Google-like impact factors or high number of downloads, high values in such factors will clearly stand as a strong argument for acceptance),
- "Presentability" of the work to a large, diverse audience,
- Quality of oral presentations by the submitter (if known),
- Submissions that permit the presentation of related interesting unpublished new results will be viewed favorably.

These "soft" criteria attempt to capture the underlying goal of the Highlight session, namely the presentation of exciting and thought-provoking efforts in advancing NLP in the biomedical domain.

Session 3 – Recent Community Shared Tasks (120 minutes)
The history of text mining and NLP in the general domain shows that community shared tasks based on carefully curated resources, such as those organized in the MUC, TREC, SemEval and ACE events, have significantly contributed to the progress of their respective fields. In the last decade, we have observed an increasing number of NLP shared tasks organized in the biomedical domain. This session will invite organizers and participants of NLP shared tasks in the biomedical domain to present the design, implementation, and results of these events. The session aims to provide a forum for the biomedical NLP community to be informed about the-state-of-the art approaches for the topics covered by the shared tasks. The organizers and participants of biomedical NLP shared tasks are

encouraged to have their systems available for audiences to gain hands-on experience of using the systems. A sample list of NLP shared tasks is shown in the following:

Title	Organizer	Website
TREC Precision Medicine Track	Kirk Roberts	http://www.trec-cds.org
FDA TAC Challenge	Dina Demner-Fushman	https://bionlp.nlm.nih.gov/tac2018druginetractions/
SemEval Clinical TempEval	Guergana Savova	http://alt.qcri.org/semeval2017/task12 https://competitions.codalab.org/competitions/17286
CLEF eHealth	Aurélie Névéol	https://sites.google.com/view/clef-ehealth-2018/task-1-multilingual-information-extraction-icd10-coding
WMT Biomedical Translation Task	Aurélie Névéol	http://www.statmt.org/wmt18/
n2c2 Shared Tasks	Ozlem Uzuner	https://n2c2.dbmi.hms.harvard.edu

The specific objectives of the session include:

1. Familiarize the audience with existing biomedical NLP community shared tasks and state-of-the-art approaches.
2. Explore the development and installation of NLP systems addressing the topics of these shared tasks

We will make efforts to coordinate the content of our pre-symposium to complement specific talks and/or panels at AMIA on similar topics, once we know which presentations have been accepted.

Learning objectives:

After participating in this session, attendees should be better able to:

1. Implement constructive feedback in their graduate research efforts
2. Establish awareness of the latest advancement in the field
3. Establish awareness and understanding of existing and available biomedical NLP tools, resources and shared tasks

Intended audience:

Anyone interested in designing or adopting biomedical NLP methods and tools is welcome. Participants working on biomedical NLP projects and researchers wishing to learn and share NLP knowledge and resources are strongly encouraged to attend. The AMIA NLP working group is a large community of over 400 members. In the last three years of the NLP pre-symposium (2013-2017), we had about 40-60 participants each year and the event was very successful. We would expect a similar or higher number of participants this year.

We will distribute information about the pre-symposium and approach potential participants using the following methods: 1) creating a **website** to communicate about the pre-symposium; 2) sending messages via mailing lists of research communities, such as the AMIA NLP working group list and LinkedIn page, the BioNLP list, and the AMIA member list; 3) contacting training programs in biomedical informatics to involve more students; and 4) promoting the event through personal contacts of organizers and active NLP working group members.

Levels of difficulty:

40% intermediate, and 60% advanced content

Pre-requisites:

None

Length of the pre-symposium: 6 hours