# Natural Language Processing Working Group Pre-Symposium: Graduate Student Consortium, Highlight, and Codeathon

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

**[1]Section of Medical Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, [2]Case Western Reserve University, Cleveland, OH, [3]Medical University of South Carolina, Charleston, SC, [4]Ontopro, Houston, TX, [5]Massachusetts General Hospital, Boston, MA, [6]National Library of Medicine, Bethesda, MD, [7]Health Language Analytics, Sydney, Australia, [8]Boston Children's Hospital, Harvard Medical School, Boston, MA, [9]SUNY, Albany, NY, [10]Columbia University, New York, NY, [11]University of Texas Health Science Center at Houston, TX, [12]LIMSI, C NRS, 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 limited 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 answer 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 clinical 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 highlight session, where significant NLP articles in clinical and biomedical domains will be presented followed by a panel discussion; and 3) a 'codeathon' of NLP tools, where user developers of NLP tools will interact with tool developers to implement tools on practical NLP tasks in groups.*

**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 limited by strict patient privacy and confidentiality laws and regulations. These limitations render access and sharing of resources (e.g., annotated text corpora) and tools based on this clinical text (e.g., trained machine learning algorithms) very difficult. Despite these difficulties, several research teams have succeeded in creating and then sharing resources based on clinical text. As a consequence, the impact of NLP for clinical and translational research is limited. Our general objective with this pre-symposium is to enhance the awareness of these resources and tools within the biomedical and clinical NLP communities and improve the reusability, portability, and interoperability of these tools and resources.

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 can present their work and get feedback from experienced researchers in the field, followed by highlight session, where significant articles will be highlighted and discussed, and then practical implementation and testing of NLP tools for practical NLP tasks in a 'codeathon'. 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 so a larger group of participants can present their research work.

**Session 1 – Graduate students 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 and clinical 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 and clinical 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 the biomedical and clinical NLP field, 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 – NLP highlights (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 highlight 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 clinical and biomedical domain.

**Session 3 – NLP tooling "Codeathon" (120 minutes)**

*Ease of use* is a very important aspect greatly influencing the adoption and ultimately the success of any tool. In the case of NLP, while many of the currently available tools are feature rich, their usability remains a problem as can be seen by the numerous questions on the forums regarding installation of the tool itself and usage of modules. Therefore, this session aims to improve the field of NLP tooling by bringing together the community of tool developers and users to provide hands-on experience of using the tools to solve specific tasks and explore areas of improvement in usability.

The specific objectives of the 'codeathon' include:

1. Installation of the NLP tool(s).
2. Familiar with NLP processing pipelines.
3. Explore the use of the tools to develop baseline systems for specific tasks, e.g., adverse event detection or family history information extraction.

We will adopt a 'codeathon' format for this session. The following process will be used:

1. We will send out a Call for Proposals (CFP) to participate as a *provider* or a *user*. Providers are typically expected to be teams with mature NLP tools. The providers will propose a task in the CFP and will agree to provide directions to accomplish the task. The symposium committee will review the proposal to approve the usefulness of the proposed task for the community, and on approving will include the task. The tasks will be grouped under tracks based on the degree of expertise required and the theme of the task.
2. Groups will be formed by the submission to implement actual NLP systems for specific tasks in one month period where providers will serve as mentors for the users to develop the system.
3. In the 'codeathon' session, the user team will demo the system and the duos will discuss how they achieve their goals and any lessons learnt to share with the audiences.

We will make efforts to coordinate the content of our pre-symposium so that we complement but do not overlap with specific talks 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 on their graduate research efforts
2. Discover and understand existing and available clinical and biomedical NLP tools and resources as well as the latest advancement in the field
3. Users will gain a hands on experience of using the tools with expert guidance
4. Providers (tool developers) will have a better understanding of the usability and robustness of their tools and areas of improvement.

**Intended audience:**

Anyone interested in designing or adopting clinical and biomedical NLP methods and tools is welcome. Participants working on biomedical and clinical NLP projects and researchers who want to learn and share NLP knowledge and resources are strongly encouraged to attend. The AMIA NLP working group is a large community with over 500 members. In the last three years of the NLP pre-symposium (2012-2016), 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