HALF-DAY TUTORIAL: Biomedical/Clinical Natural Language Processing

TUTORIAL LEADERS:
Ozlem Uzuner (Primary Contact)
Information Studies
University at Albany, SUNY
Albany, NY
Tel: Direct: 518)-442-4687; Dept: 518-442-5110
Email: ouzuner@albany.edu

Meliha Yetisgen
Biomedical and Health Informatics
Linguistics
University of Washington
Seattle, WA
Tel: Direct: 206-685-8224; Dept: 206-616-0369
Email: melihay@uw.edu

WHAT YOU CAN TAKE AWAY:
This tutorial will present an overview of the biomedical and clinical NLP as a field, and provide information on data, tools, and methods of this field with the intent of providing the researchers with a jump-start into these domains. The focus will be on the demand for NLP in biomedical and clinical domains, the potential for impact, and the required NLP tasks. The data and methods that support the NLP tasks will be introduced and a vision for future work will be shared.

CONTENTS:
1) Introduction to biomedical/clinical NLP (30 minutes)
We will open the tutorial with an introduction to biomedical/clinical NLP including motivation and significance of the domain in health care as well was challenges and current research focus. This section will include:
   a) Motivation and significance
      i) Text processing needs in biomedical research and clinical practice
      ii) General research problems in biomedical and clinical NLP
   b) Challenges to performing biomedical and clinical NLP

2) Current research focus (45 minutes)
The first part of this tutorial will discuss the current research focus in biomedical/clinical NLP. We will start with describing domain specific applications that require text processing and continue with main NLP research topic areas.

13 March 2014
a) Biomedical/clinical focused applications
   i) Biomedical: Examples include literature-based discovery, literature search, and protein interaction extraction.
   ii) Clinical: Examples include phenotyping, decision support systems, and billing systems.

b) NLP research
   i) Syntax: POS tagging, parsing
   ii) Semantics: medical concept extraction, acronym expansion, relation extraction, temporal information extraction in the clinical context, assertion analysis, and coreference resolution

3) Coffee Break (15 minutes)

4) Datasets and annotation process (30 minutes)
   In the second half of the tutorial, we will describe the nature of datasets and annotation process. We will highlight the important information that would get researchers started with obtaining biomedical and clinical data, and the sensitive nature of clinical data, then proceed to introducing the available (with data use agreements) datasets. Generation of annotated biomedical and clinical corpora require attention to domain expertise as well as linguistic knowledge. Therefore, many biomedical and clinical NLP tasks need to involve experts of both types. Other experiments show that biomedical and clinical NLP annotations can be high in cognitive load and can benefit from post-processing and consistency checks.
   a. Nature of clinical data
      i. Privacy considerations: HIPAA guidelines to access clinical data with personal health information
      ii. Data access and sharing process through institutional review boards (IRB)
   b. Available datasets
      ii. Clinical datasets: i2b2, Cincinnati Children’s Hospital Computational Medicine Center, MIMIC II.
   c. Annotation process
      i. Key aspects of biomedical and clinical annotation
         1. Annotators
         2. Tools
         3. Methods
      ii. Differences from traditional corpus annotation process
      iii. Lessons learned.

5) Methods (45 minutes)
   We will continue the second half of the tutorial by describing the commonly used methods to process biomedical and clinical text. We will present these methods in three groups:
   a. Rule-based approaches
   b. Statistical approaches
   c. Hybrid approaches
   Recently, the biomedical and clinical NLP communities have generated common repositories that can serve as a starting point for acquiring existing tools. We will include those systems in our presentation.
   d. Summary of publicly available tools
   e. Pointers to publicly available tools
6) Open questions and the future directions (15 minutes)
We will conclude the tutorial by going over the open questions and future directions of the field.

BIO OF LEADERS:
Ozlem Uzuner is Associate Professor in the Department of Information Studies in the College of Computing and Information at the University at Albany, and a research affiliate at the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology (MIT). Her research interests include applications of natural language processing to biomedical and clinical domain.

Meliha Yetisgen is Assistant Professor in the Department of Biomedical Informatics and an Adjunct Assistant Professor in the Department of Linguistics at the University of Washington (UW). Her research interests include statistical natural language processing and its application to biomedical literature and clinical text. Before joining to UW as a faculty, she worked in industry as a text-mining researcher and was a PhD student and post-doctoral researcher at UW.