Biomedical/clinical NLP

Ozlem Uzuner
Information Studies
University of Albany, SUNY
Albany, NY
ouzuner@albany.edu

Meliha Yetişgen
Biomedical and
Health Informatics
University of Washington
Seattle, WA
melihay@uw.edu

Amber Stubbs
Library and
Information Science
Simmons College
Boston, MA
stubbs@simmons.edu

Introduction

Recent years have seen a rapid growth in the use of biomedical documents and narrative clinical records for applications outside of direct patient care. Accordingly, recent years have also seen an increase in the development of NLP technologies for concept and relation extraction, summarization, and question answering on these data.

This tutorial will present an overview of the biomedical and clinical NLP data, tools, and methods with the intent of providing the researchers with a jump-start into these domains. We will focus on the demand for NLP in biomedical and clinical domains, the potential for impact, and the required NLP tasks. We will introduce this information in the following categories:

Overview of biomedical/clinical NLP

Biomedical narratives are often dense with domain-specific jargon. Clinical narratives, in addition to being dense with domain-specific jargon, exhibit the complexities of a specialized sub-language. They are written by the domain experts and for the domain experts. Their primary purpose is to assist in informing future decisions about the care of the patients. As a result, both biomedical and clinical narratives present challenges for existing open-domain NLP technologies and require special considerations for their accurate understanding and interpretation.

In this section, we will discuss the data sources currently available to researchers, as well as provide an overview of the research questions both domains. On the clinical side, this includes using EHRs for phenotyping and decision support systems. The biomedical side uses NLP to explore fields such as literature-based discovery and literature searches.

Current research questions in biomedical and clinical NLP

NLP applications are generally built to answer specific questions about data. In this section, we will provide examples of the types of questions researchers are asking in the clinical and biomedical domains. Additionally, we will discuss how different linguistic aspects of these data are addressed by looking at existing syntactic (part of speech tagging, parsing) and semantic (concept extraction, temporal information extraction, coreference resolution) systems.

Datasets and the annotation process

Building annotated corpora for any task can be challenging, but the biomedical and clinical domains have additional barriers that make creating these corpora difficult. In this section, we will discuss available annotated resources in both domains, and discuss challenges in biomedical and clinical corpus building, such as restrictions on data access and the need for domain experts to be part of the annotation process.

Clinical Annotation Case Study

The 2014 i2b2 NLP Shared Task[1] involved two NLP challenges: 1) de-identification of medical records, and 2) identification of risk factors for coronary artery disease in diabetic patients. Each of these tracks required a separate annotation effort, and in this portion of the tutorial we will describe the end-to-end process of creating this annotated resource, from data selection to writing the annotation guidelines to creating the final gold standards.

NLP Methods

Just as there are many research questions in the biomedical and clinical domains, there are many existing NLP systems that address these questions. In this portion of the tutorial, we will describe the three main approaches (rule-based, statistical, and hybrid) commonly used to process biomedical and clinical text. Additionally, we will present on-going research projects from our research labs including (1) extracting structure and semantics from clinical text through section segmentation and assertion analysis and (2) clinical applications such as phenotype modeling and specific examples of information extraction in the radiology domain.

Open questions and future directions

Research in the fields of biomedical and clinical NLP is far from complete; the end of this tutorial will look at current unsolved problems in these fields, as well as look ahead towards potential future research questions.

Acknowledgements

This tutorial was supported in part by Informatics for Integrating Biology and the Bedside (i2b2) award No. 2U54LM008748 from the National Institutes of Health (NIH)/National Q5 Library of Medicine (NLM), by award No. 1R13LM01141101 from the NIH NLM, by award No. 1R21EB016872 from NIH NIBIB, and by Institute of Translational Health Sciences award No. UL1TR000423 from NIH NCATS.

[1] `https://www.i2b2.org/NLP/HeartDisease/`