A Hands-on Introduction to Natural Language Processing in Healthcare

Scott L. DuVall\textsuperscript{a,b}, Olga V. Patterson\textsuperscript{a,b}

\textsuperscript{a} VA Salt Lake City Health Care System, Salt Lake City, UT, USA
\textsuperscript{b} University of Utah, Salt Lake City, UT, USA

Abstract

This tutorial is an update to the natural language processing (NLP) tutorial given at the Medinfo conference in 2010. It will provide attendees with a general overview of NLP tools and methods used in health research and patient care. Participants will be introduced to NLP, the types of problems that can be addressed with NLP, and how to effectively plan and execute a NLP task using patient medical records. Synthetic clinical notes will be provided along with open-source tools that will allow participants to implement a working NLP system. The eHOST annotation application and the Leo NLP libraries, tools developed and used in the Department of Veterans Affairs (VA) and built on existing community standards, will be introduced and used to illustrate the complete life cycle of an NLP project, from design to human annotation / chart review to NLP system creation to evaluation. The tutorial will be presented by two instructors involved in the design and development of these NLP tools who have completed more than 100 NLP tasks in the VA and other health care institutions. Attendees will experience the process of completing an NLP task and leave the tutorial with concrete examples of how NLP can be used at their institutions to benefit research studies or patient care.

Keywords:
Natural Language Processing; Information Processing.

General Topics

While some patient health information is stored in structured, coded format in electronic medical records, the majority of clinical information available about a patient’s state of health is stored in clinical text notes and reports. Diagnosis codes, for example, can indicate that a patient has a particular condition such as diabetes. But it is in the progress notes and other clinical text where details about disease progression and control, rationale for treatment decisions, and summary and inference of clinical findings are documented. Care providers create narrative text through data entry or dictation, because writing is a more natural form of communication and can provide much richer detail than searching a limited number of options in drop-down boxes or using other structured methods of information collection. Clinical text also serves as a medical-legal record of interaction with the patient. Natural language processing (NLP) describes a family of methods aimed at deriving meaning from narrative text. NLP is used in healthcare in order to automatically obtain information contained in clinical text in electronic medical records.

To effectively use the pertinent information contained in clinical text for decision support, research, and other tasks, the information must be captured in a more formal representation that computers can manipulate. NLP can make this possible by automatically identifying, extracting, and structuring the information – a common task in the NLP family called information extraction. Information extraction identifies and assigns values to terms in narrative text that represent concepts like symptoms, procedures, or medications. Extracted information can be mapped to ontologies, terminologies, and other formal representations of health information to feed decision support and research.

With the growing adoption of electronic medical records, and the push for care to be paperless, the amount of patient information available in clinical text and the number of these documents continues to increase. Meystre et al [1] reviewed methods using NLP for automated encoding of electronic clinical documents that have been evaluated by several groups. Although NLP processes are only beginning to be used in day-to-day healthcare operations, they have a history of success in being used to answer specific clinical questions [2-6]. These include identifying diagnoses, medical problems, and signs and symptoms [7-11]. Successes like these show the value of using NLP and continue to push NLP methods into the mainstream.

Recently, a broad array of NLP applications have been developed and applied to specific clinical tasks or research questions, and interest in NLP has spread beyond the circle of NLP researchers to mainstream informaticians and health practitioners: clinicians, public health practitioners, clinical researchers, and others are seeking out information on how to implement and apply NLP tools. Only within the last five years have suites of NLP tools become publicly available for use by a broader audience than the original developers. In extracting information from clinical notes using NLP, most systems employ a pipeline of individual processes in combination. This pipeline of processes is cumulative, moving from basic tasks like separating words or paragraphs through higher-level tasks that may depend on the output of upstream processes. Frameworks exist that help facilitate the creation of pipeline-based NLP projects. The advantage of adopting a framework is the general architecture that provides a solid start to an NLP project and the ability to re-use modules created for other projects or by other teams.
The Apache Unstructured Information Management Architecture (UIMA) project is one such framework [12]. UIMA provides a data model for storing text and layering structure information, known as annotations, on the text and an infrastructure for creating new modules that can be strung together in pipelines with other new or existing modules. As such, UIMA NLP modules developed by many different researchers for many different types of projects designed to handle many different tasks are available for incorporating in new NLP projects [13-14]. The community development and collaboration supported by these frameworks have the potential to substantially advanced NLP use in healthcare.

A challenge for research teams in adopting frameworks like UIMA can be the lack of enterprise programming expertise. UIMA has capability to support systems that analyze and annotate images, video, and sound in addition to text. This flexibility means understanding documentation not specific to NLP development, the need to override and specialize functionality, and the curation of XML descriptor files. As an attempt to bring this powerful functionality and interoperability to more research teams, our team developed a set of libraries that encapsulate the most common NLP functionality and configuration in a simple programmatic interface. This set of libraries is called Leo and supports the rapid development of new NLP functions and seamless incorporation of existing UIMA modules [15].

For NLP systems to be trained and evaluated, human annotation or chart review can be used to create reference standards. The eHOST Annotation Tool allows researchers to define specific annotation tasks and manually identify concepts and relationships that can then be used to train an NLP tool to automatically identify [16]. Because human annotation is slow and costly, eHOST was designed to boost efficiency.

This tutorial addresses the need in the community to learn through hands-on experience how to apply and experiment with NLP tools that are publicly available. We will introduce NLP in healthcare, demonstrate Leo and eHOST, and guide participants in the implementation of an NLP project from initial design stages to evaluation.

**Tutorial Structure**

**NLP Overview**

An overview of what NLP is and how it is used in healthcare will be presented. A description of current challenges in NLP, which problems can be addressed with NLP, and how to conceptually frame a project in a way that NLP can help solve it will be provided.

**Manual Annotation**

The creation of an annotation guideline that explicitly defines the NLP task and the translation of the guideline into an annotation schema will be demonstrated. Participants will be assisted in using eHOST to annotate records, perform a consensus review, and measure inter-annotator agreement.

Records annotated by participants, along with a set of records annotated prior to the conference, will be used as the reference standard for evaluating the NLP tools.

**NLP Development Overview**

UIMA and Leo will be introduced with a description of how they facilitate NLP pipeline creation. An NLP pipeline will be designed and created using examples.

**Creation of NLP Modules**

Participants will be walked through the creation of four NLP modules: identify concepts with regular expression, dictionary-lookup, concept-value pair creation, and classification. Each step in the implementation process will be explained from defining the types of annotations that will be made to programming the actual annotation of the documents to reviewing the module output. Participants will be assisted in the actual implementation of the modules.

**Creation of Multi-Module NLP Pipeline**

The creation of an NLP pipeline and incorporation of modules developed in the community will be illustrated. Participants will combine the modules they created along with other modules that will be provided to form a complete pipeline. Defining the input and output of the system along with how the modules interact will be demonstrated.

**Evaluation of System Performance**

All documents will be processed through the completed pipeline. Participants will be shown how to set up the pipeline output for scoring against the manually annotated reference standard and each NLP system will be evaluated. Ensuring that the score truly reflects the accuracy of the task and how to translate those results into practice will be addressed.

**Specific Educational Goals**

At the end of the workshop, participants will be able to:

- Understand the current uses of NLP in healthcare
- Describe how to frame problems that can be addressed by NLP
- Describe the tools available for NLP projects
- Understand the role of manual annotation in NLP
- Translate NLP tasks into annotation guidelines
- Manually annotate records using eHOST
- Create an NLP project using Leo
- Create NLP modules for concept lookup, concept-value pair linking, and classification
- Create an NLP pipeline that utilizes modules created for other projects or by other NLP researchers
- Understand how to evaluate NLP output

**Expected Attendees**

This workshop is intended for informaticians, application programmers in clinical settings, and clinicians with an interest in implementing NLP tools. Familiarity with the Java programming language is helpful, but not required. Code samples, step-by-step instructions for creating NLP modules, and assistance with building pipelines will be provided.
Tutorial Speakers

This tutorial will be taught by two instructors experienced as researchers, developers, and users of NLP and annotation tools for specific clinical domains.

Scott L. DuVall is the VA Informatics and Computing Infrastructure Associate Director in the VA Salt Lake City Health Care System in Salt Lake City, Utah responsible for development and application of natural language processing and annotation systems. He leads the VINCI Services team, which has completed more than 100 NLP tasks for the VA and other healthcare institutions. He received his PhD in biomedical informatics from the University of Utah and is a Research Assistant Professor with a joint appointment in the University of Utah School of Medicine and the College of Pharmacy.

Olga V. Patterson is a Research Associate at the VA Salt Lake City Health Care System in Salt Lake City, Utah. She received her PhD in the Department of Biomedical Informatics, University of Utah. Her research experience includes information extraction and document classification of clinical text.

References