Health Informatics

C-C1-01:

Strategic Plan for Developing Natural Language Processing Capacity at HMORN Sites

David Carrell, PhD, *Group Health, Center for Health Studies*; Virginia Immanuel, *Group Health, Center for Health Studies*

For the specific purpose of crafting a strategic plan for developing scientific and technical capacity for applying natural language processing (NLP) techniques to clinical text in HMO Research Network (HMORN) sites we convened a working conference in Seattle, December 17–18, 2008. Expert consultation in the strategic planning process was provided by six NLP experts from three institutions considered national leaders in the development and application of NLP technologies: the Mayo Clinic, the University of Pittsburgh Medical Center, and Vanderbilt University. Conference participants included representatives of five HMORN sites, Northwestern University, and the University of Washington. Specific aims of the conference were: 1) To develop a strategic plan for acquiring natural language processing capacity including, potentially, development of in-house capacity, subcontracting NLP tasks to one or more external collaborators/contractors, some combination of these approaches, or none at all; and 2) To educate investigators and technical staff about the processes of developing, testing, and applying NLP algorithms for extracting data from clinical text in the context of typical HMORN-style research programs. Traditional conference presentations addressed "What is natural language processing," "Illustrations of potential NLP projects," "Considerations when developing and applying NLP algorithms," and "Comprehensive, large-scale information extraction systems." A round-table-style discussion sessions focused on issues of human capital requirements for NLP projects, subcontracting NLP tasks to outside entities, multi-site collaboration on NLP projects, and a strategic plan for developing NLP capacity. This presentation summarizes the strategic planning products of the conference, currently in draft form and undergoing edits by a working group of the expert consultants and conference participants. Preliminarily, the strategic plan recommends incremental development of NLP capacity in HMORN-like sites through sequential execution of thoughtfully-selected small-scale NLP and NLP-like projects that represent key technical aspects of what can become a mature program of NLP capacity within 3–5 years. Initially, small-scale projects would be conducted in close collaboration with outside informatics experts to provide scientific and technical guidance to the immediate project and to assure fidelity with a longer-term locally-adapted strategic plan.

C-C1-02:

Data Extraction From Text, Step 1: Preparing Text for Machine Processing

David Carrell, PhD, *Group Health, Center for Health Studies*

Background: Natural language processing (NLP) uses software to assist in the extraction of information from clinical text, a process usually performed entirely by chart abstractors. Before NLP can be applied the text in question must be prepared for machine processing. In research settings this pre-processing work often involves several successive and related tasks, requiring substantial amounts of time and attention from people representing various types of clinical, scientific and technical expertise. Appreciating the tasks and participants involved in pre-processing clinical text can make the work more manageable, efficient, and effective. **Methods:** The information presented here comes from case study analyses of three small-scale projects involving preparation of clinical text (pathology reports, radiology reports, and progress notes) for processing by the Cancer Text Information Extraction System. Supplementing these experiences is information from anecdotal conversations with natural language processing experts. **Results:** Ten separate pre-processing tasks were identified: 1) obtaining source feeds, 2) assessing completeness, 3) de-duplication, 4) universe description, 5) cleaning and formatting, 6) de-identification, 7) database loading, 8) sampling, 9) preparation of the NLP system input feed, and 10) quality assurance. Nine types of expertise or task participants required for pre-processing were identified: 1) IRB representative, 2) source-system manager, 3) network/dbase administrator, 4) programmer, 5) statistician, 6) investigator, 7) informaticist, 8) clinical domain expert, and 9) manual chart abstractor.

Conclusions: Pre-processing clinical text is an important phase and potentially challenging aspect of extracting information from clinical text using NLP. Because researchers require accurate information about the larger universe of documents or patients represented by the sampled and processed text, pre-processing can present numerous challenges, the solutions to which draw on many areas of expertise in a multi-step and sometimes iterative process.

C-C1-03:

Identifying Respiratory-Related Clinical Conditions From ED Reports With Topaz

Wendy W Chapman, PhD, *University of Pittsburgh*; Henk Harkema, PhD, *University of Pittsburgh*

Background: Identifying patients for research studies often requires timely and expensive manual chart review. Automated searching on textual records results in many false positive cases, because a large proportion of clinical conditions described in reports occurred in the past history (past history of pneumonia), were experienced by a relative and not the patient (father suffered from CHF), or are described as being absent (patient denies chest pain). Natural Language Processing (NLP) promises to automatically encode clinical conditions and modifiers necessary for accurate identification of patients matching a given case definition. **Methods:** We developed an NLP system called Topaz that currently identifies 55 respiratory related clinical conditions from clinical reports, assigning the value acute, chronic, or absent to each of the conditions in a three-step process. Module 1: Find instances of the conditions in the text. Module 2: For each condition marked by Topaz, assign the following values: Existence (present, absent); Temporality (recent, historical, hypothetical); Experiencer (patient, other). Module 3: Integrate information from individual annotations to assign a single value to each of the 55 conditions. **Results:** We evaluated each module on a set of 60 emergency department reports by comparing against physician annotations. Sensitivity and positive predictive value for Modules 1 and 2 are as follows: Module 1: Mark conditions: 99%, 89%; Module 2: Negation (98%, 96%), Historical (76%, 50%), Hypothetical (100%, 81%), Experiencer (100%, 33%). Weighted kappa score between values assigned to the 55 conditions by Topaz (Module 3) compared to physician annotations was 0.85, indicating high agreement between Topaz and a physician. **Conclusion:** In spite of mistakes in marking conditions and assigning contextual features, Topaz performed similarly to a physician in assigning 55 conditions the values acute, chronic, or absent based on the ED report. Topaz could improve sensitivity of case retrieval by addressing linguistic variation common in clinical reports (e.g., synonymy, abbreviations, etc). Topaz could also improve precision of case retrieval by addressing temporality, experiencer, and negation of clinical conditions.

C-C1-04:

Building a Health Services Information Technology Research Environment

David W Gehrum, BS, *Geisinger Information Technology*; JB Jones, PhD, MBA, *Geisinger Center for Health Research*; Gregory J Romania, MS, *Geisinger Information Security Office*; David L Young, BS, *Geisinger Information Technology*; Virginia R Lerch, MPH, *Geisinger Center for Health Research*; Christa A Bruce, MPH, *Geisinger Center for Health Research*; Diane Donkochik, BA, *Geisinger Information Technology*; Walter F Stewart, PhD, MPH, *Geisinger Center for Health Research*

Background: The electronic health record (EHR) has opened a new era for health services research (HSR) where information technology (IT) is used to re-engineer care processes. While the EHR provides one means of advancing novel solutions, a promising strategy is to develop tools (e.g., online questionnaires, visual display tools, decision support) distinct from, but which interact with, the EHR. Development of such software tools outside the EHR offers an advantage in flexibility, sophistication, and ultimately in portability to other settings. However, institutional IT departments have an imperative to protect patient data and to standardize IT processes to ensure system-level security and support traditional business needs. Such imperatives usually present formidable process barriers to testing novel software solutions. We describe how, in collaboration with our IT department, we are creating an environment and a process that allows for routine and rapid