Saturday, November 4

8:30 am – 4:30 pm W08: Natural Language Processing Working Group Graduate Student Consortium, Highlight, and Codeathon

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

Hongfang Liu, Rong Xu, Stephane Meystre, Sivaram Arabandi, Kavishwar Wagholikar, Dina Demner-Fushman, Jon Patrick, Guergana Savova, Ozlem Uzuner, Chunhua Weng, Hua Xu, Pierre Zweigenbaum

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.

Monday, November 6

7:00am – 8:15am National Library of Medicine (NLM) Update
Patricia F. Brennan, PhD, RN

8:30am – 10:00am S14: Oral Presentations - Evaluating SNOMED-CT and Its Use to Achieve Semantic Interoperability

Achieving Logical Equivalence between SNOMED CT and ICD-10-PCS Surgical Procedures
Kin Wah Fung*, Julia Xu, Filip Ameye, Arturo Romero-Gutiérrez, Arabella D'Havé (8:30am – 8:48am)

Abstract: Surgical procedures are coded in SNOMED CT in the electronic health record and in ICD-10-PCS in administrative systems. We compared the logical definitions of SNOMED CT concepts to the ICD-10-PCS axial components to identify overlap and gaps. The biggest discrepancy was in the surgical approach which was specified in all ICD-10-PCS codes but only in 8.7% of SNOMED CT surgical procedures. Among the top 100 commonly used ICD-10-PCS codes, 25% could be matched fully in meaning and logical definition to pre-coordinated SNOMED CT concepts. Using post-coordination, it was possible to represent the full meaning of 86% of ICD-10-PCS codes. Logical mapping between SNOMED CT and ICD-10-PCS is feasible but will be more productive if more SNOMED CT concepts can become fully-defined. Short of full logical matching, partial logical matches can also be useful in suggesting candidate maps for expert review and to support interactive post-coordination.

Identifying Potentially Missing Hierarchical Relations in SNOMED CT based on Lexical Features – Impact of Synonyms and Lexico-syntactic Constraints
Satyajeet Raje, Olivier Bodenreider* (9:06 am – 9:24 am)

Abstract: We have evaluated the effects of adding synonyms and lexico-syntactic constraints on the identification of potentially missing hierarchical relations in a subset of SNOMED CT. Adding lexico-syntactic constraints alone increases precision, but both enhancements together degrade performance.

1:45pm – 3:15pm S35: Oral Presentations - Privacy & Deidentification

Modes of De-identification
Mehmet Kayaalp* (2:03pm–2:21pm)

Abstract: De-identification of protected health information is an essential method for protecting patient privacy. Most institutes require de-identification of patient data prior to conducting scientific studies; therefore, it is important for clinical scientists to be cognizant of all modes of de-identification and all services provided by their de-identification tools. In this article, we discuss eight different modes of de-identification that yield de-identified data at different levels of quality. Most of these modes can be used in combination to achieve the best performance.

3:30 pm – 5:00pm S46: Oral Presentations – Omics

Mining the literature for genes associated with placenta-mediated maternal diseases
Laritza Rodriguez*, Stephanie Morrison, Kathleen Greenberg, Dina Demner-Fushman (4:42pm-5:00pm)

Abstract: Automated literature analysis could significantly speed up understanding of the role of the placenta and the impact of its development and functions on the health of the mother and the child. To facilitate automatic extraction of information about placenta-mediated disorders from the literature, we manually annotated genes and proteins, the associated diseases, and the functions and processes involved in the development and function of placenta in a collection of PubMed/MEDLINE abstracts. We developed three baseline approaches to finding sentences containing this information: one based on supervised machine learning (ML) and two based on distant supervision: 1) using automated detection of named entities and 2) using MeSH. We compare the performance of several well-known supervised ML algorithms and identify two approaches, Support Vector Machines (SVM) and Generalized Linear Models (GLM), which yield up to 98% recall precision and F1 score. We demonstrate that distant supervision approaches could be used at the expense of missing up to 15% of relevant documents.

3:30pm – 5:00pm S54: System Demonstrations - Systems for Knowledge Representation

RxMix – Use of NLM drug APIs by non-programmers
Olivier Bodenreider*, Lee Peters (3:30pm-4:00pm)

Abstract: Subject matter experts, including pharmacists, pharmacy benefit managers, health researchers and health data analytics specialists, are generally knowledgeable in drug resources, such as RxNorm, and have developed interesting use cases for it. However, processing datasets often requires programming skills, e.g., making calls to a service, such as the RxNorm Application Programming Interface (API). To enable users who do not have programming skills to leverage National Library of Medicine (NLM) drug APIs, we developed RxMix, a web application designed for non-programmers to create queries for complex use cases and execute batch queries against our APIs.

5:00 pm – 6:30 pm Poster Session 1 (Columbia Hall)

RxNorm Concept History Service
Olivier Bodenreider*, Lee Peters

Abstract: RxNorm contains concepts which represent the drugs currently marketed in the U.S. With each monthly RxNorm release, new drug products are added and obsolete drug products are removed. While this curation supports e-prescribing and drug information exchange use cases, it is detrimental to analytics use cases. The issue in this case is that some of the RxNorm drug identifiers (RxCUIs) stored in clinical data warehouses have become obsolete and can no longer be interpreted in reference to the current RxNorm dataset. To address this issue, we started developing an RxNorm concept history service as part of the NLM drug APIs (https://rxnav.nlm.nih.gov/. This service is similar to what we developed for managing obsolete identifiers from the National Drug Code (NDC). For each obsolete RxCUI, the service returns a canonical representation of the drug concept, making it possible to relate an obsolete drug product to a class through its ingredient or to find a similar active drug product based on ingredient, strength and dose form information.

TNF-alpha Use as a Risk Factor for Lymphoma in Rheumatoid Arthritis Patients Using the Virtual Research Data Center
Gregory Brown*, Fabricio Kury, Clement McDonald

Abstract: Large anonymized data sets based on patient information can assess associations that are difficult to research with clinical trials. We used this approach to explore the association of TNF-alpha inhibitor use with lymphoma.

CTB: A Custom Taxonomy Builder for Named Entity Extraction
Dina Demner-Fushman*, Willie Rogers

Abstract: Clinical researchers are often interested in finding specific disease phenotypes in clinical text. When using dictionary-based named entity recognition tools, they might need to add an important local terminology or limit the scope of named entities provided in the publicly available resources. In this poster, we present an online, as well as downloadable, open-source tool that allows constructing custom-built subsets of the UMLS Metathesaurus.

Clinicians' Perceptions of Usefulness of the PubMed4Hh App for Clinical Decision-Making at the Point of Care
Kyungsook Gartrell*, Caitlin Brennan, Fang Liu, Gwenyth Wallen, Paul Fontelo

Abstract: Evidence-based medicine in healthcare has been facilitated Internet access through wireless mobile devices. This study was intended to test how evidence as abstracts and the bottom-line summaries, accessed with mobile devices affected clinicians' decision-making at the point of care or office. Our results show that retrieving relevant health information from biomedical literature using the PubMed4Hh was useful at the point of care and in the office.

Trend Analysis of EHR Events to Facilitate Hypothesis Generation and Data Quality Assessment
Yohan Sumathipala, Vojtech Huser*

Abstract: Common Data Models (CDMs) for Electronic Health Record (EHR) data facilitated the creation of software tools that help with data characterization and quality assessment. Our study develops methods for analyzing EHR data on the prevalence of various clinical events to identify temporal trends that lead to hypothesis generation, epidemiologic analysis, or a data quality inquiry. Clinical events included prescription drug ingredients, medical procedures, and diagnostic conditions. Our methods identified which events experienced significant rises or falls in prevalence.

Evolution of Research Topics in MEDLINE
Andrej Kastrin*, Thomas Rindflesch, Dimitar Hristovski

Abstract: This study identifies the major research focuses and the current status and trends in the life sciences. It provides a description of the intellectual structure and dynamics of the entire field of biomedicine from the perspective of frequently appearing MeSH descriptors. This study show that (i) using MeSH terms is plausible for tracking historical events in the biomedical domain; (ii) the evolution of MEDLINE occurs in an incremental fashion; (iii) over the years more and more diverse research disciplines are involved in the complex process of scientific evolution, and links among them become stronger; and (iv) different research areas have different dynamic evolution patterns.

New MetaMap Features for Processing Numerical Tables
Francois-Michel Lang*, Dina Demner-Fushman

Abstract: A recent project using MetaMap to identify adverse reactions in drug labels encountered two serious problems: long runtimes and low precision. Analysis revealed that tables containing many numbers caused both problems; we present two MetaMap enhancements that substantially improved MetaMap's runtime and precision with no loss in recall.

Enhancing LexSynonym Features in the Lexical Tools
Chris Lu*, Destinee Tormey, Lynn McCreedy, Allen Browne

Abstract: Synonym features in the Lexical Tools are used as element synonyms for subterm substitution in query expansion to improve recall in concept mapping in various NLP projects, such as UMLS-CORE, Sophia, STMT, MMTx. This paper describes the implementation and evaluation of enhanced synonym features in the Lexical Tools (2017). The results show an improvement (~5%) on both recall and F1 with similar precision using enhanced synonym features in query expansion for UMLS concept mapping.

RxNav 2.0 – A web-based, mobile-responsive RxNorm browser
Lee Peters*, Olivier Bodenreider

Abstract: In 2016 RxNav was redesigned to conform to modern web application development principles, with the goal of simplifying the user experience and supporting mobile devices. This poster describes the changes from the original application and highlights some of the new features.

MeSHgram: An Open Source Tool to Visually Browse Co-occurrence of MeSH Terms in PubMed
Satyajeet Raje*, Ravi Teja Bhupatiraju, Abdelrahman Hosny, Ben Busby

Abstract: MeSHgram is an interactive tool to browse cooccurrences of MeSH terms within the PubMed corpus. It allows users to search for MeSH terms and quickly visually inspect the evolution of cooccurrences over time as well as other MeSH terms in context in real-time. The tool can be used for quantification of known research patterns as well as potentially aid novel hypotheses generation. MeSHgram is available at www.meshgram.org.

Detecting Adverse Drug Event Safety Signals from MEDLINE Reports: Challenges in Employing Cross-terminology Mapping of MeSH to MedDRA
Abhivyakti Sawarkar*, Alfred Sorbello, Anna Ripple, Olivier Bodenreider

Abstract: The US Food and Drug Administration developed a web-based prototype software tool to automate detection of adverse drug events (ADE) from MEDLINE through quantitative data mining of Medical Subject Heading (MeSH) indexing terms. To facilitate interoperability between MEDLINE and FDA Adverse Event Reporting System, we render MeSH descriptors for ADEs to Medical Dictionary for Regulatory Activities preferred terms leveraging the Unified Medical Language System. We describe challenges in cross-terminology mapping in a pharmacovigilance use case.

Tuesday, November 7

10:30am-12:00pm S70: Oral Presentations - Knowledge Modeling, Management, and Assessment

Information Retrieval for Biomedical Datasets: The 2016 bioCADDIE Challenge
Kirk Roberts*, Anupama Gururaj, Xiaoling Chen, Saeid Pournejati, Trevor Cohen, William Hersh, Dina Demner-Fushman, Lucila Ohno-Machado, Hua Xu (10:30am – 10:48am)

Abstract: This abstract describes the 2016 bioCADDIE Dataset Retrieval Challenge, an information retrieval (IR) shared task focusing on retrieving biomedical datasets for researchers. The information associated with biomedical datasets includes structured codes, unstructured descriptions, and linked scientific articles. Due to this complexity, evaluating dataset IR methods on a common benchmark is vital. This presentation will cover the motivation and mechanics of the challenge, an overview of the participating systems, and directions for future research.

Cardioprotective Drugs and Incident Dementias in Medicare's Big Data
Fabricio Kury*, Seo Baik, Clement McDonald

Abstract: Mixed evidence has been published about the possibility that antihypertensives and statins might protect against incident dementias, but these studies are challenging because they require big cohorts, long follow-up, and detailed data on drug use. We leveraged the vast Medicare data to comprehensively study this topic. We were unable to identify a protective effect for statins, while antihypertensives were mostly protective albeit not in agreement with previous studies.

1:45pm – 3:15pm S87: System Demonstrations - Systems for Knowledge Management

Systems Demonstration: NIH Common Data Elements Repository (1:45 pm – 2:15 pm)
Liz Amos*, Christophe Ludet, Lisa Lang, Vojtech Huser

Abstract: The NIH Common Data Elements Repository has been designed to provide access to structured human and machine-readable definitions of data elements and measures that have been recommended or required by NIH Institutes and Centers. While large repositories exist (most notably NCI), the ability to search across all CDEs and link to clinical data standards, seemed critical. The CDE Repository contains approximately 42,000 data elements and 1700 forms from 12 different NIH initiatives. This systems demonstration highlights the tool's ability to readily search across all collections from initiatives in a centralized place, annotate with clinical terminology(ies), and variety of form creation and export capabilities. The demonstration will feature collaboration functionality, ability to link to variable level datasets, and interaction with other NLM terminology tools and resources. Discussion of two pilot use cases will demonstrate the tool in context while highlighting feedback that guided improvements as well as describe challenges faced during implementation.

5:00 pm – 6:30 pm Poster Session 2 (Columbia Hall)

Testing an Explainer Animation for Public Health Education
Jeffrey Day*, Anne Altemus

Abstract: We are experimenting with online explainer animations to improve the outreach mission of the National Library of Medicine. A short animation on histamine will accompany a food allergy article in NIH MedlinePlus magazine and we will see if it can help enrich the magazine articles, attract new audiences, and expand awareness of the NIH. If successful, this project can inspire more multimedia communication strategies to convey public health messages.

Crowdsourcing Evidence-Based Medicine (EBM)
Paul Fontelo*, Fang Liu

Abstract: We developed a new crowdsourcing method to search, review, and rate recent evidence from PubMed citations. It provides an open environment for users to not only discover and rate the usefulness of a citation, but also read relevant citations previously rated by other clinicians. This tool harnesses the wisdom of the crowd and may help healthcare professionals find high quality evidence from PubMed, more efficiently.

Annotation of Research Common Data Elements Using Clinical Terminologies
Vojtech Huser*, Cynthia Burke, Minh-Diep Nguyen, Liz Amos

Abstract: Common Data Elements (CDEs) aim to standardize how research data is captured in clinical trials. We used 22,705 PhenX CDEs as input for a pilot study in annotation of CDEs using a framework that is inspired by the SNOMEDCT compositional grammar. Annotation of CDEs with SNOMEDCT terms can (1) aid in discovery of clinical trials data for re-use by researchers, (2) integration of research and EHR data and (3) discovering overlap across CDE initiatives.

Mapping U.S. FDA National Drug Codes to Anatomical-Therapeutic-Chemical Classes using RxNorm
Fabricio Kury*, Olivier Bodenreider

Abstract: For analyzing prescription datasets, usually one wishes to identify drugs by classes rather than U.S. FDA National Drug Codes. We demonstrate how to utilize the NLM RxNorm online API to map NDCs to WHO Anatomical-Therapeutic-Chemical classes, and present statistics and caveats of mapping 71,309 NDCs found in Medicare Part D claims from 2006 to 2013 and 134,580 NDCs found in a commercial all-payer claims dataset from Partners Healthcare from 2011 to 2012.

Number of Citations of Structured and Unstructured Abstracts in PubMed
Fang Liu*, Paul Fontelo

Abstract: Structured abstracts are more informative so some may rely solely on abstracts to inform decisions. We wanted to know whether articles with structured abstracts would be cited more often than unstructured abstracts. We reviewed citations data for 11 terms in PubMed. We found that unstructured abstracts outnumber structured abstracts by more than 2:1. However, we found no significant difference in the number of citation between structured and unstructured abstract.

Harmonizing User-defined Phenotypic Variables using Latent Semantic Analysis (LSA) to Improve Data Discoverability in dbGaP
Liz Amos, Satyajeet Raje*, Masato Kimura, Preeti Kochar, Victoria Wilder, Ben Busby

Abstract: Harmonizing phenotypic variables in dbGaP is critical for effective reuse of the data. Sustaining manual post-coordinated efforts is infeasible. The purpose of this study is to (semi-)automate the process harmonization of variables in dbGaP. Our initial results a method based on Latent Semantic Analysis (LSA) leveraging user-defined descriptions of these variables show promise in this direction.

Wednesday, November 8

8:30am – 10:00am S105: Oral Presentations - Enhanced Cohort Identification and Retrieval

Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval
Tracy Edinger*, Dina Demner-Fushman, Aaron Cohen, Steven Bedrick, William Hersh (8:30am-8:48am)

Abstract: Objective: Secondary use of electronic health record (EHR) data is enabled by accurate and complete retrieval of the relevant patient cohort, which requires searching both structured and unstructured data. Clinical text poses difficulties to searching, although chart notes incorporate structure that may facilitate accurate retrieval. Methods: We developed rules identifying clinical document sections, which can be indexed in search engines that allow faceted searches, such as Lucene or Essie, an NLM search engine. We developed 22 clinical cohorts and two queries for each cohort, one utilizing section headings and the other searching the whole document. We manually evaluated a subset of retrieved documents to compare query performance. Results: Querying by section had lower recall than whole-document queries (0.83 vs 0.95), higher precision (0.73 vs 0.54), and higher F1 (0.78 vs 0.69). Conclusion: This evaluation suggests that searching specific sections may improve precision under certain conditions and often with loss of recall.

8:30am – 10:00am S110: System Demonstrations - Tools for Clinical and Public Health Learning

NLM3D: an Open-Source Library of Medical-Imaging-derived 3D Polygonal Models for Use in Public Health, Medical and STEM Applications
Kristen Browne*, Anne Altemus (9:00am-9:30am)

Abstract: Medical and scientific communications professionals are increasingly utilizing animation, gaming platforms, and augmented/virtual reality to engage with their audiences. The NLM 3D project aims to capture and translate the wealth of high resolution 3D imaging data that is being produced by the scientific community such that it can be leveraged by communications efforts. As a starting point, the Visible Human image archives from the National Library of Medicine are being segmented and processed into high quality, production-ready 3D polygonal anatomical models. These models are uploaded to a custom Drupal-based archive along with comprehensive metadata such that they can readily searched for and retrieved through a web-based graphic user interface. Medical ontologies are being employed to aid model organization, while established standards from the 3D community are being used to ensure consistency and quality of model construction and display. This system can easily be adapted to include further dataset types including veterinary anatomy, cellular and subcellular structures, and molecular models. This holds promise that NLM3D: Anatomy could serve as the beginnings of an expansive and comprehensive scientific 3D model library.

10:30am – 12:00pm S115: Oral Presentations - Implementing and Using Interoperability/HIE for Population Health

The LOINC/RSNA Radiology Playbook: A unified terminology for radiology procedures
Daniel Vreeman*, Ken Wang, Chris Carr, Beverly Collins, Swapna Abhyankar, Jamalynne Deckard, Clement McDonald, Daniel Rubin, Curtis Langlotz (11:06am-11:24am)

Abstract: The Regenstrief Institute and the Radiological Society of North America are collaborating to produce a common information model and unified terminology with a single governance process for radiology procedure names. Here we describe our approach to unifying LOINC radiology content and the RadLex Playbook into the jointly developed LOINC/RSNA Radiology Playbook. The current version (June 2017) contains unified modeling of 5,500+ terms covering all imaging modalities.

Enabling Interoperability between Healthcare Devices and EHR Systems
Swapna Abhyankar*, Paul Schluter, Kathryn Bennett, Daniel Vreeman, Clement McDonald (11:24am – 11:42am)

Abstract: Regenstrief Institute and IEEE have created a mapping between IEEE 11073 and LOINC concepts that bridges the gap between healthcare devices and EHRs. Such a map allows information from a variety of devices to flow into and integrate with a patient's EHR record so that the data are interpretable, actionable, and accessible for research.