Research Opportunity

Preceptor Name: Cui Tao, PhD
Campus: UTHealth Houston

Overall grant or other program name:
Cancer Patient Medical History Extraction, Modeling, and Reasoning

Briefly describe your overall research program:
A substantial portion of the information in medical records is historical in nature. Past patient history in Electronic Health Records (EHR) can be long, especially for patients with chronic diseases. Accurate representation and efficient analysis of the temporal aspects in patient medical history could boost an array of clinical and translational research such as disease progression studies, decision support systems, and personalized medicine.

To automatically untangle and linearize the temporal constraints of clinical events embedded in EHR data, however, is quite challenging due to the following reasons: (1) EHR systems are by nature highly diverse. EHR data typically reside within disparate databases, registries, data collections, and departmental systems, in different formats. (2) Approximately 80% of the EHR data consists of unstructured data, which is not directly machine-queriable. (3) Useful temporal constraints and relations may not be stated explicitly in unstructured documents, but rather need to be computed or inferred.

Barriers to temporal data modeling, normalization, extraction, and reasoning have precluded the efficient use of EHR data sources for event history evaluation and trending analysis: (1) The current federally-supported open-source EHR data normalization tools do not focus on the time aspect of unstructured data yet; (2) Existing time models focus only on structured data with absolute time, lack supporting reasoning systems, or only offer application-specific partial solutions which cannot be adapted to complex EHR data; (3) Current temporal information extraction and relation resolution approaches are either difficult to adapt to EHR data, not scalable, or only offer application-specific partial solution.

This project addresses current gaps among ontologies, Natural Language Processing (NLP), and EHR-based clinical research for temporal data.
representation, normalization, extractions, and reasoning. We propose to expand our TIMER (Temporal Information Modeling, Extracting, and Reasoning) framework, an open source software package, for automatic temporal information normalization and reasoning for large, diverse, and heterogeneous EHR data sets and structure the integrated data for further analysis.

We will use Colon Cancer (CRC) as an evaluation use case to test our tool. For CRC, we will identify the patient cohort using diagnoses codes which usually can provide high specificity for cancer cases. We will use TIMER to follow important events in the history including a family history of colon polyps or colorectal cancer, a personal history of colon polyps or inflammatory bowel disease, a personal history of obesity, evidence of anemia or blood loss per rectum and prior screening or surveillance activities aimed at preventing development of colorectal cancer; as well as structured relevant events in the EHR including fecal blood tests, proctoscopies, colonoscopies, and diagnosed inflammatory bowel disease or colon polyps.

**Name of specific project in which a student will participate:**
CRC patient medical history annotation

**Specific duties of student:**
Student will help annotating the patient history of a set of CRC patients. We will provide students with trainings on using our annotation tool, ontologies, and reasoners. Students will also be involved in ontology improvement and manuscript preparation.

**This would be a good fit for someone with what major or interests?**
Students have biology/medical background and/or IT background who are interested in EHR and informatics research.

NB: you may submit as many research opportunities as you like, but to be competitive you should probably limit yourself to no more than three.

E-mail your finished form(s) to cpritsummer@uth.tmc.edu