AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

Towards Effective Representation of Clinical Documents for Search and Retrieval

Anthony R. Davis, Michael Nossal, N. Stephen Ober
3M Health Information Systems, 7500 Old Georgetown Rd., Bethesda, MD 20814
{tdavis, mnossal, sober}@coderyte.com

Abstract

Recent studies have demonstrated the advantages of structured search of PubMed abstracts when compared with unstructured key word search. We explore whether search on clinical text is similarly enhanced by representing domain specific structures, information, and knowledge. Examples include representations of document structure and sections, local context such as negation, and appropriate modeling of scalar quantities. We examine tasks ranging from recruitment of suitable patients for studies, to chronic disease prevention and management, to longitudinal studies of individual patients or groups, as well as comparative experiments performed on an NLP enhanced clinical search tool that operates on large corpora of clinical text.

Aspects of Clinical Search

We examine here issues in processing, indexing, and representing information in clinical records, with goals including acute and chronic care management, patient complications, patient safety, provider profiling, and revenue cycle support and efficiency.

Health care professionals are trained to evaluate patients in a standard manner. The resulting medical records are intended not only to accurately document a patient's exam, diagnosis, and treatment, but to capture the context of the clinical situation, which can impact potential treatment strategies and care plans. Moreover, immediate access to a patient's past medical history can provide critical information necessary to properly evaluate a patient. Failure to do so can result in medical errors and patient harm, a major focus in today's health care environment. In addition, rather than focusing on a single point in time, clinicians may think about patients and their conditions using a continuum-of-care approach, with each clinical encounter providing additional data points that add to the full context of a patient's health care "story."

As a simple example, the difference between positive (patient has a heart attack) and negative (patient has no history of heart attack) polarity is critical to a patient's evaluation. Interpretation of scalar variables is critical as well. Differences as small as 1% in two lab values—especially from two time frames in the same patient—can dramatically change the care strategy and clinical outcome.

Related Work in Medical Search

It has been noted previously (e.g., Demner-Fushman and Lin 2007, Boudin, Nie, and Dawes 2010) that unstructured search queries or questions yield inferior performance on clinical search tasks, compared with explicitly structured query interfaces. The PICO ("Patient/Problem, Intervention, Comparison, Outcome") framework (Richardson, et al. 1995) formulated within the evidence-based medicine paradigm guided these researchers in constructing query interfaces that reflect a clinician's model of patient encounters. In addition, this approach demands techniques that can identify elements of each type in clinical text. For example, (Demner-Fushman and Lin 2007) describe extractors they developed for each element, which can be hand-crafted rules (for "Population") or term extraction based on UMLS semantic categories (for "Problem"). Both studies report improved performance on retrieval tasks when PICO structure is taken into account.

However, it is important to note that both studies used a corpus of PubMed abstracts, rather than the "raw material" of individual patient reports. Clearly, the differences between the two types of documents are significant, and while improving search and retrieval techniques on clinical literature promises much, it may not address many of the issues noted in the first paragraph, which pertain to individual patients, healthcare providers, and institutions, in a way that published clinical literature typically does not. Here, we examine some of the elements of search and retrieval on this primary clinical data, drawing on our experience with a large corpus of physician reports from various academic medical centers and other healthcare providers. We suggest that the PICO elements need to be augmented with other sorts of representation for effective search and retrieval on patient reports that healthcare professionals will derive benefit from. We then describe the features of a tool we have developed, DataScout, that provides some of this desired functionality.

Representing and Extracting Information from Patient Records

There are multiple levels of structure inherent in clinical documents. For various search scenarios, it can be advantageous to encode low and high level annotations into the same query.

The most fine-grained structures lie within individual “medical facts”. For instance, a “left ventricular ejection fraction of 35%” expresses laterality, a body part, an attribute being measured, and a scalar value measurement. Representation at this conceptual level expands the range of search, retrieval and analytic possibilities beyond what can be easily achieved with a term-based search. For example, once ejection fraction is represented as a scalar quantity, queries patients with an ejection fraction of less than 40% can be meaningfully processed.

Sentence-level analysis can reveal the negation or temporal context for the medical fact, as patients who are “suspicious for carcinoma” are very different from those who are “negative for carcinoma”. Note also that there are long distance relationships within medical text, such that a radiology impression section stating “unremarkable chest” may in fact indicate that the patient does not have pneumonia, if that was the stated purpose of the exam in the clinical indication section.

Most medical documents are “semi-structured”, in that they have labeled regions. These vary across different specialties; however, the diverse surface forms can usually be mapped to a “SOAP Note” structure, which affects search. Thus a condition documented in the “assessment” portion of the document typically has more weight than the subjective portion of the note. A good search tool can recognize that a patient with pneumonia documented in the clinical indication section but not in the impression section does not have pneumonia.

Larger-scale structures include annotations that reveal cross document relationships, such as temporal order between documents, or the correlation of documents with an individual patient or provider. For example, one might wish to find patients with a given test result from a pathology report after an initial indication from a radiology report, in order to validate or contradict a radiology report.

One example of the kind of clinical documentation obtainable with this type of analysis is the interaction of diabetes and smoking. Out of approximately 20,000 diabetic patients in one repository of patient records we have processed with DataScout, about 7,200 records mention smoking. However, the great majority of these mentions—over 90%—turn out not to indicate that the patient currently smokes. Information on smoking and diabetes is vital not only to guide patient education, but because it is a CMS Accountable Care Organization quality measure. Another example is the administration of aspirin following the onset of MI symptoms. We have found that an unstructured keyword-based search for 'MI' or 'myocardial infarction' and 'aspirin' in DataScout, on a corpus of ED clinical reports, yields a misleading picture of how frequently the aspirin protocol is followed, while a search using the ICD-9 codes for MI and a search for the term 'aspirin' provides a more realistic (and lower) frequency estimate. Again, this is a revealing indicator both for patient outcomes and as a care quality measure.

Challenges for the Future

The advent of U.S. health care reform has brought a plethora of performance measures propagated by health care entities (mostly payers) to assure that providers focus on clinical quality and cost reduction. Successfully implementing initiatives such as Meaningful Use, PQRS, ACO quality metrics, and the Joint Commission for the Accreditation of Healthcare Organizations’ clinical accreditation criteria can translate to millions of dollars in additional revenues and the assurance of continuing as a preferred provider in governmental and commercial health care networks. DataScout and similar NLP systems must therefore incorporate clinically sophisticated techniques, to assure that the implemented guidelines are as up-to-date as possible, and that their analytical capabilities provide meaningful and accurate information.

References

Boudin, F., Nie, J-Y., and Dawes, M. 2010. Clinical Information Retrieval using Document and PICO Structure. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 822–830. Association for Computational Linguistics.

Demner-Fushman, D., and Lin, J., 2007. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics* 35(1):63-103.

Richardson, W., Wilson, M., Nishikawa J., and Hayward, R. 1995. The well-built clinical question: A key to evidence-based decisions. *American College of Physicians Journal Club*, 123(3):A12–A13.