An Information Extraction Framework for Cohort Identification Using Electronic Health Records

Hongfang Liu PhD[1], Suzette J. Bielinski PhD[1], Sunghwan Sohn PhD[1], Sean Murphy[1], Kavishwar B. Wagholikar MBBS PhD[1], Siddhartha R. Jonnalagadda PhD[1], Ravikumar K.E. PhD[1], Stephen T. Wu PhD[1], Iftikhar J. Kullo MD[2], Christopher G Chute MD PhD[1]
[1]Department of Health Sciences Research, [2]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN

Abstract

Information extraction (IE), a natural language processing (NLP) task that automatically extracts structured or semi-structured information from free text, has become popular in the clinical domain for supporting automated systems at point-of-care and enabling secondary use of electronic health records (EHRs) for clinical and translational research. However, a high performance IE system can be very challenging to construct due to the complexity and dynamic nature of human language. In this paper, we report an IE framework for cohort identification using EHRs that is a knowledge-driven framework developed under the Unstructured Information Management Architecture (UIMA). A system to extract specific information can be developed by subject matter experts through expert knowledge engineering of the externalized knowledge resources used in the framework.

Introduction

With the rapid adoption of Electronic Health Records (EHRs), it is desirable to harvest the information and knowledge in EHRs to support automated systems at point-of-care and to enable secondary use of EHRs for clinical and translational research. Much of the EHR data is in free text form. Comparing to structured data, free text is a more conventional way in the health care environment to express concepts and events. The free text records are generated by automated or manual transcription of dictation recordings and direct entry by the care providers. However, free text is very challenging for searching, summarization, decision-support, or statistical analysis. To reduce medical errors, improve health care quality, and enable secondary use of EHRs, information extraction (IE), which structures and encodes clinical information stored in free text, is necessary. Approaches to IE are based on either symbolic techniques (e.g., NegEx[1]) or statistical machine learning. Clinical IE applications using symbolic techniques can be cumbersome to implement and may lack portability. IE applications based on statistical NLP techniques require annotated examples and are easy to apply but may not accurately capture the relationships among words in a document such as negation. As pointed out by Wagholikar et al.[2] and Chapman et al.[3], when a task involves a specific subdomain (as in preventive care decision support of cervical cancer) or a limited number of named entities (as in detection of influenza), sublanguage analysis detecting subdomain semantics combined with contextual information detection and expert knowledge engineering is a viable approach.

Through a collaboration with IBM, Mayo Clinic has developed a system called Mayo Clinic Information Extraction system used to process all clinical notes available in the Enterprise Data Trust (EDT)[4]. A variation of this system named clinical Text Analysis and Knowledge Extraction System (cTAKES) has been released under an Apache open source license through the open health NLP consortium (OHNLP)[5]. Additional modules have been implemented within cTAKES recently including a smoking status identification module[6], a refined drug information extraction module, and a side effect extraction module [7]. Here, we report a new IE framework for extracting named entities and their corresponding contextual information under cTAKES that is purely knowledge-driven. We will assume named entities are handcrafted based on expert knowledge with or without sublanguage analysis. We test the framework through implementing a couple of NLP components for cohort identification.

In the following, we describe background information of the modules in the IE framework. We then describe the system in detail followed by case studies on two eMERGE phenotypes to evaluate the IE framework.

Background

In this section, we describe the background information about section and contextual information detection. We then describe two Electronic Medical Records and Genomics (eMERGE) phenotypes used in evaluating the IE framework.

Section Detection - Clinical notes are often divided into sections, or segments, such as "history of present illness" or "past medical history." These sections may have subsections as well, such as the "cardiovascular exam" section of the "physical exam." One can gain greater understanding of clinical notes by recognition of the section in which a

name entity locates. For instance, both "past medical history" and "family medical history" sections can contain a list of diseases, but the context information is very different. Section tagging is an early step in NLP applications for clinical notes. One such system for section detection is SecTag which recognizes section headers through terminology lookup, machine learning, spelling correction, and scoring techniques[8,9]. The terminology used by SecTag provides a list of concepts that represent particular section headings by extending Logical Observation Identifiers Names and Codes (LOINC®). Each concept in SecTag has one or more synonyms that may be used to specify a section in an actual note.

Contextual Information – Contextual information of a condition includes: negation (is the condition negated or not), temporality (historical or current), and experiencer (who has the condition). ConText is a system that determines the values for the above three contextual properties of a clinical condition [10]. The contextual property negation specifies the status of the clinical existence of a condition. The default value of this property is affirmed. If a clinical condition occurs within the scope of a trigger term for negation, ConText will change the default value to negative. For example, in the sentence ‘‘The patient denies any nausea,” the value of negation for the condition ‘‘nausea” will be negated.

The eMERGE consortium - The eMERGE consortium was organized and funded by the National Human Genome Research Institute (NHGRI) and the National Institute of General Medical Sciences (NIGMS) to develop, disseminate and apply approaches for combining DNA biorepositories with the EHR for large-scale, high-throughput genetic research [11]. The eMERGE consortium has demonstrated the applicability and portability of EHR derived phenotype algorithms using different types and modalities of clinical data for algorithm execution including billing and diagnoses codes, NLP, laboratory measurements, patient procedure encounters, and medication data[12,13]. We implemented the NLP component for the following two eMERGE phenotypes where Mayo has been the primary site of developing the algorithms:

Peripheral arterial disease (PAD) – PAD is a highly prevalent disease affecting about 8 million individuals aged 40 years or older in the US with nearly 20% of the elderly (>70y) patients seen in general medical practice affected by the disease. It is associated with significant mortality and morbidity, underscoring the necessity of a rigorous investigation of factors that influence susceptibility to PAD. An NLP system has already been developed under UIMA for detecting PAD cases from radiology notes in 2010 [14]. An evaluation on a test data of manually annotated 455 Mayo cases indicated that the accuracy agreement between the 2010 system and the gold standard was 0.93. However, other eMERGE sites require a substantial amount of time to deploy the 2010 system. Here, we re-implemented the algorithm by referring to the official PAD algorithm available publicly aiming to demonstrate that a customized NLP engine can be developed efficiently.

Heart failure (HF) – HF is a complex disease in which the heart is unable to supply sufficient blood flow to the body and is diagnosed based on the presence of clinical symptoms and further characterized by cardiac ejection fraction (i.e. reduced or preserved). In 2010, HF affected 6.6 million Americans at a cost of 34.4 billion[15,16]. However, the syndromic nature of HF presents challenges to identify HF cases and controls from EHR data for research given that the diagnosis relies on clinical evaluation. Mayo is in the process of developing the HF phenotyping algorithm. We have conducted sublanguage analysis and derived knowledge-based rules for HF that can be executed on clinical notes to identify HF patients.

System Description

As mentioned, our IE framework is knowledge-driven and developed under Unstructured Information Management Architecture (UIMA) which has been widely adopted to implement systems for processing unstructured content[17]. Different UIMA components can be combined to create a pipeline of modular tools, and all components use the same data structure, the Common Analysis Structure (CAS). In general, a UIMA pipeline consists of three types of components, a Collection Reader for accessing the documents from a source and initializing a CAS object for each document. The analysis of the documents is performed by Analysis Engines that add annotations to the CAS objects. Finally, CAS Consumers are used for final processing, e.g., for exporting the annotated information to a format that can be used for downstream analysis (e.g., building machine learning classifiers).

Figure 1 shows an overall architecture of the system as well as an example of knowledge needed. After initializing a document to a CAS object, sentences in the document are detected, tokens and chunks are generated. Section detection is then performed so that we can specify sections to extract information. The default section dictionary is the SecTag terminology supplemented with section synonyms acquired from i2b2 2010 NLP corpora[18]. The information extracted can be of two types: concept mention (CM) or matching (MATCH) where the context information detection (adapted from ConText) is performed only on all CM instances. There are three knowledge

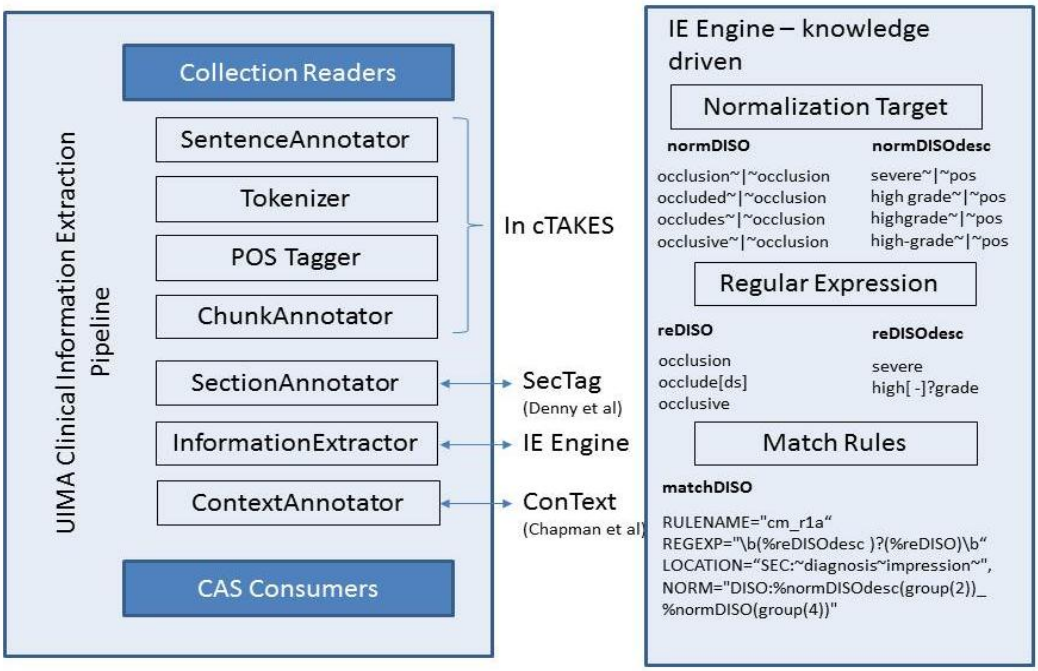

Figure 1. System architecture of the IE framework under cTAKES.

components: regular expression, normalization and match rule, for executing the IE engine where the regular expression component specifies patterns used in the match rule components and normalization target is to specify the target form of a regular expression. For example, "severe occlusion" and "high-grade occlusion" in diagnosis or impression sections will be matched instances and they are normalized to "pos_occlusion". Those knowledge components are externalized from the IE engine to facilitate customizability and maintenance.

Experiment

To evaluate the described IE framework, we implemented two eMERGE phenotyping NLP algorithms where one (PAD) has been defined in eMERGE I and the other (HF) is in the process of development in eMERGE II. The following describes the experiments in detail.

Peripheral arterial disease (PAD) **–** We downloaded the eMERGE PAD algorithm [19] and crafted three knowledge components used in the IE framework. We evaluated the performance of the implementation by classifying each of the 455 documents (i.e., the gold standard data set used to evaluate the PAD algorithm deployed back in 2010) into two classes: Class I - positive and probable documents, and Class II - negative and unknown documents. The knowledge components include two groups of concepts: anatomical concepts and disorder/procedure concepts. Each radiology note was processed through the IE Engine and then classified using the following rule: if a document has a sentence containing one positive mention from anatomical terms and one positive mention from disorder/procedure terms, we classify the document as Class I, otherwise, Class II. We reported the error matrix where various performance metrics can be derived.

Heart failure (HF) – To develop HF algorithms, we started with sublanguage analysis to acquire terms and their corresponding context information for HF patients. The data set used for sublanguage analysis includes 706 HF patients from the Heart Failure in the Community Cohort (HL72435), a gold standard cohort of manually abstracted cases defined according to Framingham Heart Failure Criteria [20]. Structured EHR data (e.g., billing and diagnoses codes, echocardiography measurement, and lab values) was combined with analyses of unstructured data (clinical notes) to identify the set of parameters needed to capture all the cases. This preliminary version of the algorithm was executed in 6,307 subjects in the Mayo Genome Consortia (MayoGC[21]), a large cohort of Mayo Clinic patients with EHR-linked genotype data and 616 eMERGE participants.

Results and Discussion

The implementation of the PAD algorithm under the new IE framework took less than one hour. Table 1 shows the contingency matrix for this new PAD implementation. The accuracy of the algorithm is 88% (399=263+136 divided by 455) which is acceptable but less than 93% achieved by the 2010 PAD system as reported previously [14]. Note that we could not define the classification the same way as previously reported [14] since there are only three classes

(positive, probable, and negative) in the current version of the gold standard while there were four classes (positive, probably, negative, and unknown) in the previous version of the gold standard.

Utilizing an existing terminology source, the UMLS, over 100 HF terms were picked by domain experts. Sublanguage analyses of the clinical notes identified six common terms present in 90% of cases from the HF Cohort: *multi-organ failure*, *cardiac failure*, *heart failure*, *CHF*, *LVF*, *ventricular failure*, under major problem or chief complaint sections. The NLP algorithm for HF is to identify the presence of those six common terms in major problem and chief complaint sections with contextual information as non-negative and non-probable. Table 2 shows the preliminary result when comparing the NLP algorithm and the usual cohort identification based on primary HF diagnosis codes (428.X). Among 706 patients in an existing HF cohort (i.e., the patients are all HF patients), there are 586 (83%) of them identified by both methods, 72 (10%) of them are identifiable only by ICD codes and 41 (6%) identifiable by only the NLP algorithm. There are 7 cases not identifiable by both methods. Combining the NLP algorithm with ICD9 codes has a recall of 99%. In the MayoGC/eMERGE Cohort, we identified 535 patients agreed by both as HF cases while additional 94 patients are identified by HF diagnosis codes as cases and 684 patients are identified only by the NLP algorithm as cases. The remaining 5,610 patients in MayoGC/eMERGE cohort are identified as non-HF cases. Abstraction of the 94 code-only cases revealed that 79% were due to coding errors with the remaining missing an NLP hit due to the lack of electronic notes (i.e. referral patients, HF predated start of EHR, or uncommon HF terms used). Preliminary abstraction of 50 of the "NLP only" patients demonstrated that about 40% were true HF cases.

Table 1. Statistics of contingency matrix on the implementation of PAD algorithm.

	System Output		
Gold Standard		Class I (Positive or Probable)	Class II (Negative or Unknown)
	Class I (Positive or Probable)	263	10
	Class II (Negative or UnKnown)	46	136

Table 2. Statistics of patients identified by the HF algorithm.

Cohort	**N (subjects)**	**Code and NLP**	**Code Only**	**NLP Only**	**Neither**
HF Case Cohort	706	586 (83%)	72 (10%)	41 (6%)	7 (1%)
MayoGC/eMERGE Cohort	6923	535 (8%)	94 (1%)	684 (10%)	5610 (81%)

Note that our downstream classification of PAD documents is based on one simple rule (i.e., the co-occurrence of a positive anatomical mention and a positive disorder/procedure mention) compared to the implementation reported previously. The rationale behind this is that manually crafted complex decision rules generally have poor portability across different institutions and we can always tend to statistical machine learning for automated inference.

High throughput phenotyping using EHR data is challenging. Our results on HF phenotyping have shown that reliance on structured data such as ICD9 diagnosis codes is insufficient to accurately identify cases. The preliminary data reported for HF cases suggests that a multi-modal approach to phenotyping can substantially improve our ability to accurately identify HF cases and non-cases. The proposed framework facilitates the construction of IE systems by subject matter experts or study coordinators who may not be familiar with NLP. Further studies are required to test the cross-institution portability of the systems using the IE framework. The reported IE framework will be open source to the research community.

Conclusion

In this paper, we have presented a knowledge-driven IE framework for cohort identification through externalizing knowledge components needed for section and contextual information extraction as well as the specific IE tasks in hand. Through implementing two eMERGE phenotyping NLP algorithms, we demonstrate the framework can be used for fast implementation of IE systems for extracting specific clinical information taking sections and contextual information into consideration.

Acknowledgement

This work was supported by National Institutes of Health: eMERGE II (HG06379) and Heart Failure in the Community (R01 HL72435).

References

1. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. *Evaluation of negation phrases in narrative clinical reports.* Proc AMIA Symp. 2001:105-109.
2. Wagholikar KB, Maclaughlin KL, Henry MR, et al. *Clinical decision support with automated text processing for cervical cancer screening*. Journal of the American Medical Informatics Association: JAMIA. Sep 1 2012;19(5):833-839.
3. Chapman WW, Gundlapalli AV, South BR, Dowling JN. *Natural language processing for biosurveillance.* Infectious Disease Informatics and Biosurveillance. 2011:279-310.
4. Chute CG, Beck SA, Fisk TB, Mohr DN. *The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data.* Journal of American Medical Informatics Association: JAMIA. Mar-Apr 2010;17(2):131-135.
5. Savova GK, Masanz JJ, Ogren PV, et al. *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*. Journal of the American Medical Informatics Association : JAMIA. Sep-Oct 2010;17(5):507-513.
6. Sohn S, Savova GK. *Mayo clinic smoking status classification system: extensions and improvements*. Proc AMIA Symp. 2009::619-623.
7. Sohn S, Kocher JP, Chute CG, Savova GK. *Drug side effect extraction from clinical narratives of psychiatry and psychology patients.* Journal of the American Medical Informatics Association : JAMIA. Dec 2011;18 Suppl 1:i144-149.
8. Denny JC, Spickard A, 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. *Evaluation of a method to identify and categorize section headers in clinical documents*. Journal of the American Medical Informatics Association : JAMIA. Nov-Dec 2009;16(6):806-815.
9. Denny JC, Miller RA, Johnson KB, Spickard A, 3rd. *Development and evaluation of a clinical note section header terminology*. ProcAMIA Symp. 2008:156-160.
10. Harkema H, Dowling JN, Thornblade T, Chapman WW. *ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports*. J Biomed Inform. Oct 2009;42(5):839-851.
11. Kho AN, Pacheco JA, Peissig PL, et al. *Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium.* Science Translational Medicine. April 20, 2011;3(79):79re71.
12. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. *A genome-wide association study of red blood cell traits using the electronic medical record.* PLoS ONE. 2010;5(9).
13. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. *Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease.* Journal of the American Medical Informatics Association: JAMIA. Sep-Oct 2010;17(5):568-574.
14. Savova GK, Fan J, Ye Z, et al. *Discovering peripheral arterial disease cases from radiology notes using natural language processing.* Proc AMIA Symp. 2010:722-726.
15. Roger VL, Go AS, Lloyd-Jones DM, et al. *Heart Disease and Stroke Statistics--2012 Update: A Report From the American Heart Association.* Circulation. Jan 3 2012;125(1):e2-e220.
16. Heidenreich PA, Trogdon JG, Khavjou OA, et al. *Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association.* Circulation. Mar 1 2011;123(8):933-944.
17. Selkoe DJ. *Alzheimer's disease: genes, proteins, and therapy*. Physiol. Rev. 2001;81:741-766 %L 2482.
18. Uzuner O, South BR, Shen S, DuVall SL. *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.* Journal of the American Medical Informatics Association: JAMIA. Sep-Oct 2011;18(5):552-556.
19. *eMERGE Library of Phenotype Algorithms* https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Library_of_Phenotype_Algorithms.
20. Roger VL, Weston SA, Redfield MM, et al. *Trends in heart failure incidence and survival in a community-based population*. JAMA. Jul 21 2004;292(3):344-350.
21. Bielinski SJ, Chai HS, Pathak J, et al. *Mayo Genome Consortia: A Genotype-Phenotype Resource for Genome-Wide Association Studies With an Application to the Analysis of Circulating Bilirubin Levels*. Mayo Clin Proc. 2011 Jul; 86(7): 606-14