Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models

Jay Urbain *

Milwaukee School of Engineering, Milwaukee, WI, United States
CTSI of SE Wisconsin/Medical College of Wisconsin, Milwaukee, WI, United States

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 4 August 2015
Accepted 7 August 2015
Available online xxxx

Keywords:
Biomedical text mining
Clinical informatics
Translational research
Natural language processing
Named entity recognition
Distributional semantic models
Heart disease risk factors
Diabetes

ABSTRACT

We present the design, and analyze the performance of a multi-stage natural language processing system employing named entity recognition, Bayesian statistics, and rule logic to identify and characterize heart disease risk factor events in diabetic patients over time. The system was originally developed for the 2014 i2b2 Challenges in Natural Language in Clinical Data. The system's strengths included a high level of accuracy for identifying named entities associated with heart disease risk factor events. The system's primary weakness was due to inaccuracies when characterizing the attributes of some events. For example, determining the relative time of an event with respect to the record date, whether an event is attributable to the patient's history or the patient's family history, and differentiating between current and prior smoking status. We believe these inaccuracies were due in large part to the lack of an effective approach for integrating context into our event detection model. To address these inaccuracies, we explore the addition of a distributional semantic model for characterizing contextual evidence of heart disease risk factor events. Using this semantic model, we raise our initial 2014 i2b2 Challenges in Natural Language of Clinical data F1 score of 0.838 to 0.890 and increased precision by 10.3% without use of *any* lexicons that might bias our results.

1. Introduction

Diabetes Mellitus is a common disease with cardiovascular complications. Complications such as an ST elevation myocardial infarction (STEMI) are associated with mortality, significant morbidity, and healthcare spending. The ability to identify patients likely to have a significant cardiovascular event within 1–3 years provides an opportunity for successful intervention. A significant challenge to developing models for predicting cardiac risk involves the identification of temporally related events and measurements in the unstructured text in electronic health records. The *2014 i2b2 Challenges in Natural Language Processing in Clinical Data* track for identifying risk factors for heart disease over time was created to facilitate development of natural language processing systems to address this challenge [1]. The details of the i2b2 Natural Language Challenge are documented in the annotation guidelines [2], and are summarized in the following *Annotation* section. Teams were provided with 521 de-identified medical record note text containing 64,035 risk factors. For evaluation, risk factor instances were rolled up to the document level for a total of 16,167 distinct record/document risk factors.

Accurate identification of risk factors requires proper characterization of time, and whether a risk factor is attributable to the patient or a family member. For example, a mention of hypertension could be a past or current condition, and the condition could be attributable to either the patient or a patient's family member. Capturing this information requires an effective approach for integrating contextual semantics within the event detection model. Distributional semantic models (DSM) can be used to quantify the semantic similarity between linguistic terms based on their distributional properties in large samples of text. The central assumption here is that the context surrounding a given word or phrase provides important information about its meaning [3–5]. DSMs provide a mechanism for representing terms, concepts, relations, or sentence meaning by using distributional statistics. The semantic properties of terms are captured in a multi-dimensional space by vectors that are constructed from large bodies of text by observing the distributional patterns of co-occurrence with their neighboring words. These vectors can then be used as measures of text similarity between words, phrases, concepts, relations, or snips of arbitrary text. Early work on use of distributional semantic modeling in EHRs (Electronic Health

* Address: Milwaukee School of Engineering, Milwaukee, WI, United States.
E-mail address: urbain@msoe.edu

http://dx.doi.org/10.1016/j.jbi.2015.08.009
1532-0464/

Records) has focused on providing vector-based representations of medical concepts, i.e., SNOMED [6], and for synonym recognition [7].

2. Annotation task

Given a set of medical records, the annotation task was to create a set of text annotations that track the progression of heart disease in diabetic patients. Multiple records were annotated for each patient, which provides a general timeline to be created from the set. Annotation tags and attributes were used to indicate the presence and progression of disease (diabetes, heart disease), associated risk factors (hypertension, hyperlipidemia, smoking status, obesity status, and family history), disease-related medications, and the time they were present in the patient's medical history. Each disease and risk factor associated with this task was assigned its own set of indicators that is used to identify whether or not the disease or risk factor is present for that patient, and when it is present. Annotations are summarized in Table 1.

Every tag except for SMOKER and FAMILY_HIST has a time attribute that is used to show when the indicator for each medical problem is known to have existed. These reflect when the indicator occurred/was active in relation to the date the medical record was written, i.e., document creation time (DCT): before DCT, during DCT, after DCT, not mentioned.

3. Methods

We developed our own NLP pipeline for this challenge. The processing pipeline consisted of the following process:

1. Preprocessing and dimensional indexing for distributional statistics.
2. Risk factor named entity recognition.
3. Attribute and measurement extraction.
4. Contextual measurement via distributional semantic model.
5. Risk factor event classification.
6. Record level aggregation of risk factor classification.

3.1. Preprocessing and dimensional indexing for distributional statistics

XML-formatted EHR data and training annotations were first imported into a relational database. Individual patient records are parsed into sentences; and sentences are parsed into words, noun phrases and candidate named entities. An inverted index is constructed using a data warehousing style dimensional data model [8,9]. We have scaled a variation of this model to several hundred Gigabytes for chemical patent retrieval [10]. The grain of the index is the individual word with attributes for position, part-of-speech, phrase and entity membership. Dimensional indexing facilitates efficient OLAP style SQL queries for aggregating distributional statistics of candidate risk events. Data can be efficiently aggregated by word, phrase, entity, sentence, or document to construct distributional co-occurrence vector representations of words, phrases, entities, or sentences.

3.2. Risk factor event recognition

Heart disease risk factor event recognition consisted of training conditional random fields (CRF) based named entity recognition (NER) models [11], and subsequent execution of the NER models on test data to identify candidate instances of risk factor events.

Table 1
Risk factor tags, attributes, and descriptions.

Risk factor	Indicator	Descriptions
Diabetes	Mention	Type 1 or Type 2 diabetes diagnosis, e.g., *HX DM*
	High A1c	A1c > 6.5; E.g., *A1c: 6.2*
	High glucose	2 fasting blood glucose measurements > 126; *e.g., SMBP 130*
CAD	Mention	Diagnosis or history of CAD
	Event	E.g., MI, STEMI, NSTEMI, bypass surgery, CABG, percutaneous, cardiac arrest, ischemic cardiomyopathy
	Test result	Exercise or pharmacologic stress test showing ischemia, abnormal cardiac catheterization showing coronary stenoses (narrowing)
	Symptom	Chest pain consistent with angina
Hyperlipidemia/Hyper-cholesterolemia	Mention	Diagnosis/history of Hyperlipidemia or Hypercholesterolemia
	High Cholesterol	Total cholesterol of over 240
	High LDL	LDL measurement of over 100 mg/dL
Hypertension	Mention	Diagnosis or preexisting condition of Hypertension.
	High BP	BP measurement of over 140/90 mm/hg
Obesity	Mention	Description of obesity
	BMI	>30
	Waist	Men $\geqslant 40''$; woman $\geqslant 35''$
Medications	Diabetes	Metformin, insulin, sulfonylureas, thiazolidinediones, GLP-1 agonists, Meglitinides, DPP-4 inhibitors, Amylin, anti-diabetes medications, combinations
	CAD	Aspirin, Thienopyridines, beta blockers, ACE inhibitors, nitrates, calcium-channel blockers, combinations
	Hyperlipidemia	Statins, fibrates, niacins, ezetimibes, combinations
	Hypertension	Beta-blockers, ACE inhibitors ARBs, Thiazide diuretics, calcium-channel blockers, combinations
	Obesity	Orlistat (xenical) or Lorqess (Lorcaserin)
Family history	Present if the patient has a first-degree relative (parents, siblings, or children) who was diagnosed prematurely (<55 for male relatives, <65 for female relatives) with CAD	
Smoker	Status: CURRENT, PAST (quit > 1 year ago), EVER (smoked at some point but it is unclear), NEVER (never smoked), or UNKNOWN (not mentioned)	

Please cite this article in press as: J. Urbain, Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.08.009