Journal of Biomedical Informatics 58 (2015) S6–S10

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Special Communication

Creation of a new longitudinal corpus of clinical narratives

Vishesh Kumar [a], Amber Stubbs [b,*], Stanley Shaw [c,d], Özlem Uzuner [e]

[a] Dartmouth-Hitchcock Medical Center, Division of Cardiology, Lebanon, NH, USA
[b] School of Library and Information Science, Simmons College, Boston, MA, USA
[c] Harvard Medical School, Boston, MA 02115, USA
[d] Center for Systems Biology, Massachusetts General Hospital, Boston, MA 02114, USA
[e] Department of Information Studies, State University of New York at Albany, Albany, NY, USA

ARTICLE INFO

Article history:
Received 5 September 2015
Revised 22 September 2015
Accepted 23 September 2015
Available online 1 October 2015

Keywords:
Corpus
NLP
Medical records
Machine learning

ABSTRACT

The 2014 i2b2/UTHealth Natural Language Processing (NLP) shared task featured a new longitudinal corpus of 1304 records representing 296 diabetic patients. The corpus contains three cohorts: patients who have a diagnosis of coronary artery disease (CAD) in their first record, and continue to have it in subsequent records; patients who do not have a diagnosis of CAD in the first record, but develop it by the last record; patients who do not have a diagnosis of CAD in any record. This paper details the process used to select records for this corpus and provides an overview of novel research uses for this corpus. This corpus is the only annotated corpus of longitudinal clinical narratives currently available for research to the general research community.

1. Introduction

The 2014 i2b2[1]/UTHealth[2] Natural Language Processing (NLP) shared task featured four tracks: (1) de-identification of medical records, (2) identifying risk factors for coronary artery disease (CAD) in diabetic patients over time, (3) assessment of software usability, and (4) novel uses of the i2b2/UTHealth data set. Of these, tracks 1, 2, and 4 relied on a new corpus and annotations created for that year's task.

The goal of Track 2, the Risk Factor ("RF") track was to look at CAD risk factors in patients over time. Accordingly, the corpus for this task included longitudinal data: multiple records for each patient, separated in time, that when put together would show changes in the patient's health status. In order to provide comparison points for patients with different medical histories, we chose for this corpus to represent three different diabetic patient cohorts. The first cohort contains patients who have a diagnosis of CAD in their first record, and continue to have it in subsequent records. The second cohort contains patients who do not have a diagnosis of CAD in the first record, but get a diagnosis of CAD by the last record. The third cohort contains patients who do not have a diagnosis of CAD in the first record and do not get it over the course of their records.

The 2014 i2b2/UTHealth NLP shared task tracks and annotations are all very different, but they all have the corpus in common. In this paper, we compare the 2014 corpus to other biomedical corpora and shared task data sets (Section 2), explain the source of the corpus and the process we used to select the records (Section 3), and examine the corpus in terms of data relating to represented groups (Section 4). This paper also discusses novel uses of this corpus outside of the tracks defined by the i2b2/UTHealth shared task organizers (Section 5), and closes with a discussion of our observations on the corpus-building process.

2. Related work

Shared tasks in the biomedical field have existed for over two decades: the OHSUMED corpus was used for an interactive task in 1994 [1], the KDD Challenge Cup was held in 2002 [2] and the TREC Genomics tracks started in 2003. However, the data for these shared tasks relied on abstracts from MEDLINE and similar sources, not clinical notes from medical facilities [3]. Laws pertaining to patient privacy make releasing clinical notes difficult, and as a result there are relatively few datasets of these notes available to researchers who are not affiliated with medical facilities [4].

One of the most widely-used collections of clinical notes is the Mimic II Clinical Database [5], a collection of medical records, including nursing notes and discharge summaries, gathered over

* Corresponding author at: School of Library and Information Science, Simmons College, 300 The Fenway, Boston, MA 02115, USA. Tel.: +1 617 521 2807.
E-mail address: stubbs@simmons.edu (A. Stubbs).

[1] Informatics for Integrating Biology and the Bedside.
[2] University of Texas Health Science Center at Houston.

http://dx.doi.org/10.1016/j.jbi.2015.09.018
1532-0464/

7 years from ICUs in the Beth Israel Deaconess Medical Center. These de-identified records have been used as a source for a variety of NLP shared tasks, such as the ShARe/CLEF eHealth Evaluation Labs 2013 [6], and 2014 [7].

Other notable collections of medical records include the THYME corpus, a collection of over 1200 de-identified notes from the Mayo Clinic, representing patients from the oncology department, specifically those with brain or colon cancer [8]; a recently created corpus of 3503 de-identified medical records of 22 different types, including discharge summaries, progress notes, and referrals [9]; and TREC Medical Records corpora [10].

The i2b2 NLP shared tasks have existed since 2006, with the first challenges in de-identification [11] and smoking status classification [12]. Other i2b2 challenges included identifying obesity and its comorbidities [13], extracting medications and associated information [14], identifying concepts, assertions, and relations [15], coreference [16], and temporal relations [17]. Each of these shared tasks included a corpus of clinical narratives, each annotated in accordance with the task description and split into training and test sets; some of these corpora re-use documents from previous years. These corpora are available from http://i2b2.org/NLP/DataSets with a data use agreement.

The 2014 i2b2/UTHealth corpus is unique among the existing corpora in that it consists entirely of longitudinal clinical records (as opposed to scientific papers and abstracts) that represent a particular medical population. The longitudinal nature of the data provides an interesting challenge for de-identification, and allows us to glean information about changes to patients over time. As part of the de-identification process, we generated realistic surrogates that maintained the narrative nature of the clinical records, the temporal relationships between dates in the patients' timelines, and co-references between locations [18]. Other corpora use generic placeholders for identifiable information, which can change the narrative structure, or swap the information between patients, breaking the continuity of the records. The i2b2/UTHealth corpus avoids these problems and is therefore suited for both de-identification work and medical research.

Institutional review boards of MIT, Partners Healthcare, and SUNY Albany approved the collection, annotation, and distribution of this corpus. The 2014 i2b2/UTHealth corpus will be available in November 2015 at https://i2b2.org/NLP/DataSets with a data use agreement.

3. i2b2/UTHealth 2014 corpus selection

The medical records for this corpus came from the Partners HealthCare Electronic Medical Records (EMR). EMR at Partners HealthCare comprises a platform shared by two large academic tertiary hospitals – Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH). First, we created a datamart of 314,000 possible Type II Diabetes (T2D) patients using highly sensitive and lenient criteria based on codified data such as the T2D ICD-9-CM billing codes (250.xx), T2D medications, abnormal glycated hemoglobin or plasma glucose. This datamart is thus enriched for patients in whom diabetes is suspected, but the majority of patients in fact do not have T2D.

To identify the actual cases of T2D in this population, an expert provided terms related to T2D. The cTAKES NLP platform [19] analyzed the narrative notes for the presence of these terms, including diabetes medications, complications (e.g. retinopathy, dialysis) and mentions of the term "diabetes mellitus". This narrative data was used, along with the codified data, by a logistic regression with LASSO penalty to develop a T2D classification algorithm. Algorithm development and validation used a training set of 400 random patients (180 T2D cases by physician gold standard, manual chart review), and a test set of 200 patients (170 T2D cases). This method successfully identified a cohort of 65,099 T2D patients at 0.97 specificity and positive predictive value (PPV) 0.96.

The smaller and validated set of patients ($n = 65{,}099$) greatly narrowed the search space, but identifying patients who fit our cohorts and finding records that met our criteria was still a challenge. In order to further reduce the number of eligible records, we searched for all of the following criteria to identify patients who have or develop CAD, applied to the patient's entire medical history:

- at least 3 CAD codes or 1 procedure code for a coronary revascularization
- at least 4 codified mentions of beta-adrenergic inhibitor medications
- at least 4 codified mentions of anti-platelet agents (such as aspirin)
- at least 4 codified mentions of statins (cholesterol lowering drugs)

These criteria returned 6382 patients, and helped us to ensure that the selected patients would have information in their records relevant to the RF task. We used similar restrictions for the patients without CAD:

- no CAD procedure
- no CAD codes at all
- at least 4 beta blockers
- at least 4 codes of anti-platelet (aspirin)
- at least 4 statins

Final section of patients and notes for the NLP challenge required manual review of the records, and selection of 2–5 records per patient based on the following:

1. Each record had to be longer than 300 words.
2. The notes needed to contain information about different aspects of the patient's health related to CAD risk factors and diabetes.
3. By the last record, the patient's medical status should be different than in the first (i.e., different medications, more risk factors present), and these changes should occur over the course of the intervening records.

One of the authors examined the records (VK) for suitability for the NLP challenge, and estimates that for every five records he selected for the corpus, he had to review and reject 80–100 records. In total, it took approximately 120 h to select the records of 301 patients for the 2014 corpus. Of those, we later removed five patients from the corpus as their file formats were incompatible with our de-identification systems.

4. 2014 i2b2/UTHealth shared task corpus

In total, the 2014 i2b2/UTHealth shared task corpus contains 1304 medical records describing 296 patients. The records are a mixture of discharge summaries and correspondences between medical professionals. For training and testing purposes, we used a 60/40 split: 790 records in the training set and 514 records in the test set. The entire corpus contains 805,124 whitespace-separated tokens, an average of 617.4 tokens per record. Each cohort (those with CAD, those who develop CAD, or those without CAD) is equally represented in the training and test sets.

The records in this corpus reflect a variety of different tones and styles. One of the represented styles in this corpus are letters between medical professionals, which take a familiar tone to