Journal of Biomedical Informatics 58 (2015) S20–S29

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus

Amber Stubbs [a,*], Özlem Uzuner [b]

[a] School of Library and Information Science, Simmons College, Boston, MA, USA
[b] Department of Information Studies, State University of New York at Albany, Albany, NY, USA

ARTICLE INFO

Article history:
Received 10 March 2015
Revised 24 July 2015
Accepted 26 July 2015
Available online 28 August 2015

Keywords:
Natural language processing
HIPAA
De-identification
Annotation

ABSTRACT

The 2014 i2b2/UTHealth natural language processing shared task featured a track focused on the de-identification of longitudinal medical records. For this track, we de-identified a set of 1304 longitudinal medical records describing 296 patients. This corpus was de-identified under a broad interpretation of the HIPAA guidelines using double-annotation followed by arbitration, rounds of sanity checking, and proof reading. The average token-based F1 measure for the annotators compared to the gold standard was 0.927. The resulting annotations were used both to de-identify the data and to set the gold standard for the de-identification track of the 2014 i2b2/UTHealth shared task. All annotated private health information were replaced with realistic surrogates automatically and then read over and corrected manually. The resulting corpus is the first of its kind made available for de-identification research. This corpus was first used for the 2014 i2b2/UTHealth shared task, during which the systems achieved a mean *F*-measure of 0.872 and a maximum *F*-measure of 0.964 using entity-based micro-averaged evaluations.

1. Introduction

Clinical narratives (i.e., free text records of patients' health and medical history) provide information to researchers that cannot be found in structured medical records, such as family history, reasoning behind prescribed treatments, and details of the patient's health. These clinical narratives are therefore an important resource for medical applications such as decision support [1,2] and cohort selection [3,4]. However, clinical narratives also contain information that identifies patients, such as their names, home addresses, and phone numbers. The Health Insurance Portability Accountability Act (HIPAA) requires that all information that identifies a patient be removed from these records before sharing the records outside of the clinical setting in which they were produced. The process of determining and removing patient-identifying information from medical records is called *de-identification*, also called *anonymization*. Often, removal of the patient-identifying information requires replacements with realistic placeholders, which we refer to as *surrogates*, also called *pseudonyms*. The replacement process is called *surrogate generation*.

HIPAA refers to patient-identifying information as Protected Health Information (PHI), and defines 18 categories of PHI as they relate to "the [patients] or of relatives, employers, or household members of the [patients]" (45 CFR 164.514). These categories are shown in Table 1.

The 2014 Informatics for Integrating Biology and the Bedside (i2b2) and the University of Texas Health Science Center at Houston (UTHealth) natural language processing (NLP) shared task featured a track focused on the de-identification of longitudinal medical records [5]. Longitudinal medical records represent multiple time points in the care of a patient, making references to past records as appropriate; their de-identification needs to pay attention to indirect identifiers that can collectively reveal the identities of the patients, even when none of those indirect identifiers would be sufficient to reveal the identity of the patient on their own. For example, the description of a patient's injuries as "resulting from Superstorm Sandy" would not be covered under the HIPAA guidelines, but they indirectly provide both a location and a year for that medical record. This information, paired with other hints about the patient's identity, such as profession and number of children, could lead to the patient's identity.

However, there are some rewards to mitigate the increased risks. Automated systems can take advantage of the repeated information: a name identified in one record as PHI can be searched for in other records in order to boost accuracy. Additionally, longitudinal records contain significantly more medical information about a patient, and they allow researchers to study a patient's health over time. We selected the 2014 de-identification corpus in order to

* 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).

http://dx.doi.org/10.1016/j.jbi.2015.07.020
1532-0464/

Table 1
18 HIPAA PHI categories (45 CFR 164.514).

1. Names;
2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly-available data from the Bureau of the Census:
   (a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
   (b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code.

support research into the progression of Coronary Artery Disease (CAD) in diabetic patients, a different track for the 2014 i2b2/UTHealth shared task [6].

In addition to paying attention to the longitudinal aspects of the corpus, the preparation of the corpus for the shared task was guided by the following goals:

1. Given the intended widespread distribution of the corpus, we needed to apply a risk-averse interpretation of the HIPAA guidelines.
2. Given the intended use of the corpus for automatic system development, we needed to maintain the semantics and integrity of the data so that systems developed on these data could be useful on authentic data.
3. We needed to have sufficient representation of PHI categories, both in type and in quantity, so that machine learning based systems could learn automatically from available samples.
4. We needed to have granular PHI categories to maximize the usability of the data for research on any subsets of the PHI, and
5. We needed to replace the authentic PHI with realistic surrogates to maintain readability.

Given our goals, we developed annotation guidelines which we applied to the 2014 i2b2/UTHealth shared task corpus for manual de-identification of 1304 longitudinal clinical narratives, generating gold standard annotations that researchers can use for automatic de-identification system development. We replaced the authentic PHI with realistic surrogates using a combination of automated systems and hand curation.

This paper describes the manual de-identification and the automatic surrogate generation processes applied to the 2014 de-identification shared task data, as well as the annotation guidelines generated for this shared task. Institutional review boards of MIT, Partners HealthCare, and SUNY Albany approved this study, and Partners HealthCare approved the de-identification methods described in Section 5.

2. Related work

Due to the strict regulations surrounding the release of medical records, very few clinical narrative data sets are currently available for de-identification research. The 2006 i2b2 NLP shared task had a de-identification track, and the corpus consisted of 889 hospital discharge summaries, which in total contained 19,498 PHI [7]. This corpus is available on i2b2.org/NLP for researchers who sign a data use agreement (DUA). PhysioNet [8] includes a de-identification dataset created by Neamatullah et al. [9], which is available at http://www.physionet.org/physiotools/deid/ with appropriate log ins and a DUA. The PhysioNet dataset contains 2434 nursing notes and 1779 instances of PHI.

Deleger et al. [10] recently created a corpus of 3503 de-identified medical records of 22 different types, including discharge summaries, progress notes, and referrals. In all, their corpus contains 30,815 instances of PHI and is available upon request.

All three of the above corpora, and the 2014 i2b2/UTHealth corpus described here, have the PHI replaced with realistic surrogates, making them suitable for NLP research into automated de-identification. All of the corpora follow HIPAA guidelines as a base for the PHI annotations, the annotations generally only have minor differences. For example, the corpus from Deleger et al. [2] conflates patient and doctor names into a single "name" category, while the other corpora maintain a distinction between patients and doctors. The 2014 i2b2/UTHealth de-identification corpus described in this paper is the only one that provides longitudinal data for patients, and it includes additional PHI categories, which we describe in Section 4.

Research into the annotation process for PHI has led to some interesting findings. South et al. [11] performed an experiment to determine if pre-annotating a corpus using automated de-identification software had a substantial effect on the quality of the PHI annotations or the time it took human annotators to check the PHI when compared to their performance on un-annotated documents. They found that the pre-annotations did not, in fact, improve inter-annotator agreement or significantly decrease the amount of time that it took the annotators to complete the task.

Additionally, in a preliminary study to the de-identification process described in this paper, we performed an experiment to determine whether PHI annotation is more accurate when done in parallel (i.e., two annotators working separately on each document) or in series (one annotator reviews the document, then the second reviews the first one's work and checks for un-annotated PHI). We found that the annotation process used had no effect on the quality of the annotations [12].