Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text

Q1 Brett R. South [a,b,d,*], Danielle Mowery [e,f,g], Ying Suo [b,d], Jianwei Leng [b,d], Óscar Ferrández [c], Stephane M. Meystre [a,b], Wendy W. Chapman [a,b,e,f,g]

[a] *Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*
[b] *VA Salt Lake City Health Care System, Salt Lake City, UT, USA*
[c] *Nuance Communications Inc., Burlington, MA, USA*
[d] *Department of Internal Medicine, University of Utah, Salt Lake City, UT, USA*
[e] *Department of Biomedical Informatics, University of Pittsburgh, PA, USA*
[f] *VA Health Care System, San Diego, CA, USA*
[g] *Division of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA*

ARTICLE INFO

Article history:
Received 8 September 2013
Accepted 6 May 2014
Available online xxxx

Keywords:
Q3 Natural language processing
Medical informatics
Confidentiality
Patient data privacy
De-identification
Anonymization
Electronic health records
Clinical corpora

ABSTRACT

The Health Insurance Portability and Accountability Act (HIPAA) *Safe Harbor* method requires removal of 18 types of protected health information (PHI) from clinical documents to be considered "de-identified" prior to use for research purposes. Human review of PHI elements from a large corpus of clinical documents can be tedious and error-prone. Indeed, multiple annotators may be required to consistently redact information that represents each PHI class. Automated de-identification has the potential to improve annotation quality and reduce annotation time. For instance, using machine-assisted annotation by combining de-identification system outputs used as pre-annotations and an interactive annotation interface to provide annotators with PHI annotations for "curation" rather than manual annotation from "scratch" on raw clinical documents. In order to assess whether machine-assisted annotation improves the reliability and accuracy of the reference standard quality and reduces annotation effort, we conducted an annotation experiment. In this annotation study, we assessed the generalizability of the VA Consortium for Healthcare Informatics Research (CHIR) annotation schema and guidelines applied to a corpus of publicly available clinical documents called MTSamples. Specifically, our goals were to (1) characterize a heterogeneous corpus of clinical documents manually annotated for risk-ranked PHI and other annotation types (clinical eponyms and person relations), (2) evaluate how well annotators apply the CHIR schema to the heterogeneous corpus, (3) compare whether machine-assisted annotation (experiment) improves annotation quality and reduces annotation time compared to manual annotation (control), and (4) assess the change in quality of reference standard coverage with each added annotator's annotations.

1. Introduction

In most electronic medical record (EMR) systems, a great deal of relevant clinical information is stored in clinical documents. Clinical documents and other medical records data are rich in protected health information (PHI). Preserving a patient's privacy and confidentiality of PHI is fundamental to the physician-patient relationship. In order to use patient medical records for purposes other than providing health care (e.g. clinical research), informed consent from the patient is required. Indeed, use of patient medical record data is subject to the ethical and legal considerations defined by the Health Insurance Portability and Accountability Act (HIPAA) codified as 45 CFR §160 and 164 and the Common Rule [1]. However, obtaining the informed consent of a large population of patients, especially for retrospective research is difficult, time-consuming, and costly. This requirement can be waived if clinical documents are de-identified (i.e., all information identifying the patient has been redacted). Although de-identification of clinical documents remains a significant challenge, fulfilling these ethical and legal requirements is often a necessary step prior to using them for clinical research. However, manually de-identifying clinical documents represents a considerable expense in terms of time and human workload.

Q2 * Corresponding author at: University of Utah, Department of Biomedical Informatics, 421 Wakara Way, Suite 140, Salt Lake City, UT 84112, USA.
E-mail address: brett.south@hsc.utah.edu (B.R. South).

http://dx.doi.org/10.1016/j.jbi.2014.05.002
1532-0464/© 2014 Published by Elsevier Inc.

Please cite this article in press as: South BR et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.05.002

Automated methods that apply natural language processing (NLP) techniques may reduce the time and effort required to manually de-identify clinical documents, especially for large-scale projects applied to tens of thousands of patient records in which manual redaction of PHI is impractical. An NLP de-identification system must accurately remove the 18 types of PHI identifiers specified under the HIPAA *Safe Harbor* method for clinical documents to be considered "de-identified". NLP systems that de-identify clinical documents are readily available [2–17], but are often developed and evaluated using specific document types. The approaches used by these systems may not be generalizable to all document types due to document specific formatting, clinical sublanguages, and prevalence of PHI [2]. Indeed, there is always the possibility that even with "de-identified" documents a PHI identifier may slip by and not be removed by all review methods [18]. A combined approach may reduce the likelihood of missing PHI identifiers and achieve acceptable coverage for certain PHI types by combining the efforts of many human reviewers with the outputs of an NLP system used as pre-annotations [19–21]. By leveraging NLP system outputs, this approach could offer a lower cost solution by pre-annotating potential PHI identifiers that human annotators review i.e., modifying existing, adding missing, or deleting spurious machine annotations. However, with any human review task relying on understanding of guidelines and tools, the cost of manual effort is high and may produce marginal returns of improved coverage as additional reviewers are added. The number of judges required to achieve acceptable coverage may also correlate with the risk of re-identification for different PHI types.

In this study, we evaluate the effects of a combined machine pre-annotation plus interactive annotation interface used to de-identify clinical documents from a publicly accessible document corpus called MTSamples. This heterogeneous clinical document corpus was selected for this study because it is a publicly available data source that could be easily obtained without a rigorous institutional data release process and contains replaced PHI mentions in context ("Dr. Sample Doctor...") that are useful for de-identification research. We first describe the MTSamples corpus. We then describe an annotation experiment including the annotation scheme used and training process. Finally, we further detail our annotation training, experiment, and evaluation approaches and assess the effects of combining machine pre-annotation plus an interactive annotation interface used to de-identify clinical documents.

2. Background

Creating a reference standard that adequately identifies all HIPAA PHI identifier types and provides accurate training and evaluation data is imperative for developing rule-based or machine-learning-based de-identification systems. A few NLP researchers have championed efforts to facilitate the creation of state-of-the-art de-identifications for clinical documents and evaluate such systems against a standard corpus [16]. In 2006, NLP researchers from the University of Albany and MIT CSAIL sponsored the 2006 i2b2 Challenge task for automatic de-identification of clinical documents. A corpus of 889 discharge summaries from Partners Healthcare was annotated in two phases. In phase 1, PHI of eight types – patient names, doctor names, hospital names, IDs, dates, locations, phone numbers, and ages – were pre-annotated using an automated de-identification system that applied machine learning approaches [17]. In phase 2, three annotators sequentially annotated each report using a serial review method and achieved consensus after each review round. The inter-annotator agreement (IAA) between annotators and the performance of the NLP de-identification system was not reported as part of the 2006 i2b2 Challenge [16].

In contrast to the 2006 i2b2 Challenge, the goal for our manual de-identification task was to estimate the effects of machine pre-annotations and an interactive annotation interface on human annotator performance and quality of generated data for a heterogeneous clinical document corpus. We compare and contrast between annotators and the generated reference standard using IAA and standard performance metrics (i.e. recall, precision, and F$_1$-measure) to assess annotator task consistency and accuracy. The effects of pre-annotation on the quality of annotated data has been investigated in many studies that include annotation of medical literature [20], POS tagging [19], named entity recognition (NER) [22] and clinical named entities [23,24], as well as common PHI types [25]. Other studies have employed semi-automated annotation methods that produce machine-generated candidate spans presented in such a way that the human reviewer must either modify incorrect annotations, delete spurious annotations, or add missed annotations [26–28]. It was our goal to produce a corpus of clinical documents annotated for PHI that maximized annotation quality while minimizing annotation effort.

3. Methods

We begin by describing the annotated MTSamples corpus. Next, we describe an annotation experiment including the annotation schema and training process. We further detail our annotation training, experiment, and evaluation approaches.

3.1. MTSamples corpus

A document sample consisting of 2,330 unique clinical documents was obtained from a publicly available resource of clinical documents called MTSamples (Medical Transcription Samples at www.mtsamples.com). These clinical documents were originally created to train medical coders and transcriptionists. The sample corpus contains document samples from 40 different medical specialties – consults, discharge summaries, and specialized medical services – including some uncommon formats. Although the MTSamples corpus does include data representing most of the 18 types of PHI identifiers specified under the HIPAA regulation, names and dates that remain have been changed (or removed) to preserve confidentiality of the users providing the data. Q4

3.2. Annotation schema

We build upon previous efforts [29] by expanding PHI types defined as part of the 2006 i2b2 challenge [16] and definitions for the Veteran Affair's (VA) setting using an annotation schema and guidelines originally developed as part of the VA Consortium for Healthcare Informatics Research (CHIR) De-identification project [8,11]. These annotation guidelines go beyond the PHI types annotated from the 2006 i2b2 Challenge. We include annotation types representing clinical eponyms, organization names, military deployments, health care units, and co-referring-paired relationships between annotations for person names (Table 1). For example, "Patient **Joe Smith**... and Mr. **Smith**...", "**Joe Smith**" and "**Smith**" might refer to the same person, in which case they would be linked in a paired relationship.

Our motivation to include annotation of clinical eponyms was twofold. First, we wished to measure human performance identifying clinical information that machine systems may misclassify as PHI. Second, we wished to enrich available data sources for training classifiers and methods to identify these information types. Human reviewers more easily identify this type of information

Please cite this article in press as: South BR et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.05.002