Journal of Biomedical Informatics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Learning to identify Protected Health Information by integrating knowledge- and data-driven algorithms: A case study on psychiatric evaluation notes

Azad Dehghan [a,b], Aleksandar Kovacevic [c], George Karystianis [d], John A Keane [a,f], Goran Nenadic [a,e,f,g,*]

[a] School of Computer Science, University of Manchester, Manchester, UK
[b] The Christie NHS Foundation Trust, Manchester, UK
[c] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[d] Macquarie University, Australian Institute of Health Innovation, Australia
[e] Health eResearch Centre, The Farr Institute of Health Informatics Research, UK
[f] Manchester Institute of Biotechnology, Manchester, UK
[g] Mathematical Institute, SANU, Serbia

ARTICLE INFO

Article history:
Received 2 February 2017
Revised 1 June 2017
Accepted 5 June 2017
Available online xxxx

Keywords:
De-identification
Named entity recognition
Information extraction
Clinical text mining
Electronic health record

ABSTRACT

De-identification of clinical narratives is one of the main obstacles to making healthcare free text available for research. In this paper we describe our experience in expanding and tailoring two existing tools as part of the 2016 CEGS N-GRID Shared Tasks Track 1, which evaluated de-identification methods on a set of psychiatric evaluation notes for up to 25 different types of Protected Health Information (PHI). The methods we used rely on machine learning on either a large or small feature space, with additional strategies, including two-pass tagging and multi-class models, which both proved to be beneficial. The results show that the integration of the proposed methods can identify Health Information Portability and Accountability Act (HIPAA) defined PHIs with overall F_1-scores of ~90% and above. Yet, some classes (*Profession*, *Organization*) proved again to be challenging given the variability of expressions used to reference given information.

1. Introduction

Clinical free text data (including, for example, consultation notes, discharge letters, imaging reports etc.) contain a number of variables that are key for understanding patients' health conditions and their responses to treatments. Extracting such information is challenging due to inherent ambiguity and variability of clinical text, but one of the main obstacles to accessing such data in the first place is the presence of Protected Health Information (PHI). While de-identification and pseudo-anonymization of well-structured health data has been used routinely, it is still not clear what acceptable levels of masking PHI mentions in clinical narrative are [1–3].

The task of finding PHI instances in text is by and large a text mining task, where the aim is to identify mentions of specific PHI data types (e.g. patient names, age, address). This is a challenging task even for human annotators [4–6], and there have been several community challenges such as the 2006 i2b2 de-identification challenge [7], the 2014 i2b2/UTHealth Shared Task in de-identification of longitudinal clinical narratives [8]; with an increasing number of systems and papers addressing this issue [9]. The task is typically approached as named entity recognition (NER) of PHI data types. Two main approaches have been followed and quite often combined: knowledge-driven methods that rely on dictionaries and rules for regularized PHI types [10–13] and machine-learning and hybrid approaches that aim at learning from data [14–19]. The results of the community challenges have suggested that machine-learning approaches, in principle, provide better and more consistent performance [7,20].

A recent challenge in this area (the 2016 CEGS N-GRID Shared Tasks Track 1b [21]) further focused on NER of up to 25 PHI types (see Table 1). The organizers provided a high-quality training and a held-out test data set of initial psychiatric evaluation notes. In this paper we describe two methods developed and evaluated as part of that task, as well as the outcome of their integration. Our methods rely on previous work [22]. mDEID is a knowledge-driven approach

* Corresponding author at: School of Computer Science, University of Manchester, Manchester, UK.

E-mail addresses: azad.dehghan@manchester.ac.uk (A. Dehghan), kocha78@uns.ac.rs (A. Kovacevic), george.karystianis@mq.edu.au (G. Karystianis), john.keane@manchester.ac.uk (J.A Keane), goran.nenadic@manchester.ac.uk (G. Nenadic).

http://dx.doi.org/10.1016/j.jbi.2017.06.005
1532-0464/© 2017 Published by Elsevier Inc.

Table 1
Composition of the submissions. CRF(mDEID) denotes the CRF-expanded version of mDEID; all references to CliDEID refer to the new version introduced here. Count is the number of instances in the held-out data; Union represents merging of the results as explained below.

Entity type	COUNT	Submission1	Submission 2	Submission 3
Date	3822	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Age	2354	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Doctor	1567	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Hospital	1328	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Profession	1010	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Patient	837	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
City	820	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Organization	697	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Country	376	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
State	481	mDEID	mDEID	mDEID
Phone	113	mDEID	mDEID	mDEID
Street	34	mDEID	mDEID	mDEID
License	21	mDEID	mDEID	mDEID
Zip	17	mDEID	mDEID	mDEID
Idnum	8	mDEID	mDEID	mDEID
Email	5	mDEID	mDEID	mDEID
Fax	5	mDEID	mDEID	mDEID
Url	3	mDEID	mDEID	mDEID

that relies on dictionaries to identify relatively closed PHI types (e.g. *Country*, *State*) and a generic set of lexico-syntactic rules that model common orthographic and contextual characteristic of specific PHI types (e.g. Addresses, Phone numbers). On the other hand, CliDEID is a CRF-based tagger that uses 279 features grouped into lexical, orthographic, semantic and positional attributes. In this paper we build on top of these two approaches by adding a learning Conditional Random Fields (CRF) layer on top of mDEID and introducing multi-class labeling into CliDEID. One of our key aims was to explore how re-usable existing de-identification methods are when migrated to new settings (e.g. a move from cancer discharge notes to psychiatric evaluation notes). The results (with an overall HIPAA strict F_1 score of ~90%, ranking our system within top 3) show the potential and challenges introduced by both data-driven methods with rich (large) and focused (small) feature sets, as well as the benefits of additional processing, including two-pass tagging, multi-class models, and label priority sorting.

The following section explains the details of the proposed methodology. Section 3 presents the results and discussion, which are followed by the conclusion.

2. Method

The approaches we designed are built using two previously published methods [22], which include a knowledge-driven open source algorithm (mDEID) and a data-driven method (CliDEID) built using linear chain CRF. We used default (CRF++) parameters: L2-regularization with C = 1.00, ETA = 0.001. For some PHI types, mDEID was expanded by providing an additional CRF layer that mainly relies on rules and dictionaries as features. CliDEID on the other hand was expanded by training models for multi-class labeling for a selected set of PHI types. We have submitted three versions for official evaluation: Submission 1 combined the outputs of Submission 2 (based on mDEID) and Submission 3 (mainly based on CliDEID). Table 1 provides the details, which are further explained below.

Submission 2 is built on top of mDEID, which was initially modeled on the i2b2/UTHealth 2014 Track I [22,23]. The rules already available in mDEID were updated based on the new training data. Further, six additional NER components were developed for *Date, Hospital, Profession, City, Organization* and *License*. In addition, CRF models were trained for nine categories using a small and focused set of features generated by the mDEID pipeline. The **B**eginning-**I**nside-**O**ut (BIO) token representation was used. The core set of features used include (see Supplement, Appendix B for per category feature set):

- *Lexical features*, such as the word/token, its stem (derived from Porter's stemmer), part-of-speech and shallow parsing information.
- *Orthographic features*, including token characteristics such as word casing (upper initial, all capital, lower case, and mixed capitalization) and type (word, number, punctuation, and symbol).
- *Semantic features*, which are binary attributes indicating if a given token was tagged by mDEID knowledge-driven components.
- *Contextual features*, including a context window of two tokens before and two tokens after each current token.

We generated a minimum of 26 (*Age*) to a maximum of 44 (*Doctor*) features using forward and backward feature selection strategies. In addition, the two-pass recognition (see below) is adopted for a subset of entity types (*City*, *Country*, *Doctor*, *Hospital*, *Organization*, *Patient*, and *Profession*).

Submission 3 is a data-driven method developed on top of CliDEID, a machine learning component of our system developed for the 2014 de-identification challenge [22]. It relies on the same feature set (*lexical*, *orthographic*, *semantic*, *positional*) and the models were trained using the **I**nside-**O**utside (I-O) schema. Building on top of the 2014 system, CliDEID has the following newly introduced characteristics:

- Models with multiple class labels. In contrast to the previous version where each CRF model was aimed at a specific category and trained only with the class labels of that particular category, a subset of the CliDEID models was trained with multiple category class labels. This was done with the goals of (a) reducing confusions between lexically similar categories (e.g. 'George' can be either a *Patient* or a *Doctor*; 'Harvard' can be either a *City* or *Hospital* or an *Organization*) and (b) exploiting the fact that some of the categories frequently occur in a sequence in the same sentence (e.g. *Patient* and *Age* – 'Valentina is a 43-year old' or *Profession* and *Organization* – 'Works as medical assistant at MEDIQUIK'). We created five multi-label machine learning (ML) models: (1) *Age* and *Patient*, (2) *City*, *Doctor*, *Hospital*, *Patient* and *Organization*, (3) *Patient* and *Doctor*, and (4–5) two models for *Organization* and *Profession*, one optimized for each of the two classes. Each of the models generates separate labels

Please cite this article in press as: A. Dehghan et al., Learning to identify Protected Health Information by integrating knowledge- and data-driven algorithms: A case study on psychiatric evaluation notes, J Biomed Inform (2017), http://dx.doi.org/10.1016/j.jbi.2017.06.005