Journal of Biomedical Informatics 58 (2015) S30–S38

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Automatic detection of protected health information from clinic narratives

Hui Yang *, Jonathan M. Garibaldi

School of Computer Science, University of Nottingham, Nottingham, UK
Advanced Data Analysis Centre, University of Nottingham, Nottingham, UK

ARTICLE INFO

Article history:
Received 2 February 2015
Revised 22 June 2015
Accepted 23 June 2015
Available online 29 July 2015

Keywords:
Protected Health Information (PHI)
De-identification
Hybrid model
Natural language processing
Clinical text mining

ABSTRACT

This paper presents a natural language processing (NLP) system that was designed to participate in the 2014 i2b2 de-identification challenge. The challenge task aims to identify and classify seven main Protected Health Information (PHI) categories and 25 associated sub-categories. A hybrid model was proposed which combines machine learning techniques with keyword-based and rule-based approaches to deal with the complexity inherent in PHI categories. Our proposed approaches exploit a rich set of linguistic features, both syntactic and word surface-oriented, which are further enriched by task-specific features and regular expression template patterns to characterize the semantics of various PHI categories. Our system achieved promising accuracy on the challenge test data with an overall micro-averaged *F*-measure of 93.6%, which was the winner of this de-identification challenge.

1. Introduction

Narrative clinical texts of patient medical records that contain rich clinical information (e.g., disease treatment and medication information) are gaining increasing recognition as an important component of clinical studies and many medical applications such as disease treatment and decision-making. To protect patient privacy and facilitate the dissemination of patient-specific data, it is required that Protected Health Information (PHI) should be removed from medical records before they are publicly available for non-hospital researchers. De-identification is a step that removes or replaces all the sensitive information while keeping the records otherwise intact.

The 2014 i2b2 de-identification Challenge Task[1] [14] is to identify and extract various types of PHI data from clinical free-texts like patient discharge summaries, clinical notes and letters. The data released for this task consists of 1304 medical records with respect to 296 patients, of which 790 records (178 patients) are used for training, and the remaining 514 records (118 patients) for testing. The medical records are a fully annotated gold standard set of clinical narratives as shown in Fig. 1. The PHI categories are grouped into seven main categories with 25 associated sub-categories. The distributions of PHI categories in the training and test sets are shown in Table 1.

It is noted that in this dataset, each patient has 3–5 documents with different Document Creation Time (DCT), which allow a general timeline present in the patient's medical history. The sets of longitudinal patient records are named with the combination of patient ID and document order ID, e.g., the files, '*100-01.xml*' and '*100-02.xml*' denote the first and second timeline record for the patient with ID '*100*'.

2. Related research issues in de-identification

Here we discuss a number of research issues that arise from the analysis of the i2b2 de-identification training data, and need to be dealt with during the system development.

First, due to terminological variations and irregularities in PHI terms, PHI term identification that is resolved on the basis of token level remains a challenging task. For example, the tokens '*T-Th-Sa*' and '*TThSa*' in fact consist of three different DATE mentions, '*T*' [Tuesday], '*Th*' [Thursday] and '*Sa*' [Saturday]. The token '*3041023MARY*' contains two different PHI category mentions, i.e. '*3041023*' for the MEDICALRECORD, and '*MARY*' for the HOSPITAL.

Second, in some well-formed categories like DATE, AGE, USERNAME, PHONE, ZIP, and MEDICALRECORD, a number of regular expression template patterns can be generated to capture the characteristics of such categories. However, due to lexical variations and the non-standard 'free' forms used by the doctors, e.g.,

* Corresponding author at: School of Computer Science, University of Nottingham, Jubilee Campus, Nottingham NB8 1BB, UK. Tel.: +44 (0) 115 95 14212; fax: +44 (0) 115 95 14254.

E-mail address: Hui.Yang@nottingham.ac.uk (H. Yang).

[1] https://www.i2b2.org/NLP/HeartDisease/.

http://dx.doi.org/10.1016/j.jbi.2015.06.015
1532-0464/

Fig. 1. Example of clinical record with annotated PHI categories.

Table 1
Distributions of PHI categories in the training and test corpora.

PHI category	Sub-category	Training data	Test data
DATE	DATE	7495	4980
NAME	DOCTOR	2877	1912
	PATIENT	1315	879
	USERNAME	264	92
AGE	AGE	1233	764
CONTACT	PHONE	309	215
	FAX	8	2
	EMAIL	4	1
	URL	2	0
ID	MEDICALRECORD	611	422
	IDNUM	261	195
	DEVICE	7	8
	BIOID	1	0
	HEALTHPLAN	1	0
LOCATION	HOSPITAL	1437	875
	CITY	394	260
	STATE	314	190
	STREET	216	136
	ZIP	212	140
	ORGANIZATION	124	82
	COUNTRY	66	117
	LOCATION-OTHER	4	13
PROFESSION	PROFESSION	234	179
Total		17,389	11,462

‘*37 yoM*’, ‘*37 yo Male*’, ‘*37 yo M*’, ‘*37yoM*’, ‘*37 y.o.m*’, an additional set of morphological rules are required to cope with orthographic variants in PHI mentions.

Third, the seven main categories of PHI entities are quite different, each exhibiting distinct characteristics in lexicon, syntax, semantics, and discourse descriptions. Due to the wide variety and complexity of features inherent in different categories, a hybrid model coupled with several NLP techniques such as machine-learning approaches, keyword-based and rule/pattern-based methods, is more appropriate in this challenge task than a single language model.

Fourth, resolving ambiguity is another challenging task for the detection of PHI entities, which includes the ambiguity of PHI terms with non-PHI terms. For example, ‘*9/12*’ can be regarded as either a DATE instance or a medical test value, or the ambiguity between different PHI categories (i.e. inter-PHI ambiguity) such as whether the term ‘*40’s*’ should be considered as an AGE entity or a DATE entity (depending on context).

Fifth, we observed that quite a number of PHI mentions explicitly or implicitly correlate to each other in the challenge corpus. Several entities co-occur in a coordination-structured expression, such as ‘*GQ/NV/whalen*’ for different DOCTOR names and ‘*EDVISIT^84091519^Thomas-yosef, Julia^09/21/68^KEMPER, SYLVAN*’ for the mentions in different PHI categories. Moreover, coreference relations among different mentions in the HOSPITAL, PATIENT, and DOCTOR categories are also worth investigating for the purpose of improving the accuracy of PHI recognition. For example, the terms, ‘*Homestead Hospital*’, ‘*Homestead*’, and ‘*HH*’ all refer to the same HOSPITAL.

Sixth, it is noticed that some PHI terms frequently appear in different timeline documents regarding the same patient, because the patient is likely to visit the same HOSPITAL or DOCTOR throughout his/her medical history. To uncover the relations among PHI terms across different timeline documents is another interesting issue to explore.

In the following sections, we will discuss how we address these research issues during system development and how the de-identification task benefits from making use of various types of relations between PHI terms discovered in the challenge corpus.

3. Methods

We developed an automated system to detect, at the token level, PHI instances from full-text medical records. The system