Journal of Biomedical Informatics 58 (2015) S47–S52

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

# Automatic de-identification of electronic medical records using token-level and character-level conditional random fields

Zengjian Liu [a,1], Yangxin Chen [b,1], Buzhou Tang [a,*], Xiaolong Wang [a], Qingcai Chen [a], Haodi Li [a], Jingfeng Wang [b], Qiwen Deng [c], Suisong Zhu [c]

[a] Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China
[b] Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou 510120, China
[c] The Sixth People's Hospital of Shenzhen, Shenzhen 518052, China

ARTICLE INFO

*Article history:*
Received 30 January 2015
Revised 2 June 2015
Accepted 9 June 2015
Available online 26 June 2015

*Keywords:*
De-identification
Protected health information
Electronic medical records
i2b2
Natural language processing
Hybrid method

ABSTRACT

De-identification, identifying and removing all protected health information (PHI) present in clinical data including electronic medical records (EMRs), is a critical step in making clinical data publicly available. The 2014 i2b2 (Center of Informatics for Integrating Biology and Bedside) clinical natural language processing (NLP) challenge sets up a track for de-identification (track 1). In this study, we propose a hybrid system based on both machine learning and rule approaches for the de-identification track. In our system, PHI instances are first identified by two (token-level and character-level) conditional random fields (CRFs) and a rule-based classifier, and then are merged by some rules. Experiments conducted on the i2b2 corpus show that our system submitted for the challenge achieves the highest micro *F*-scores of 94.64%, 91.24% and 91.63% under the "token", "strict" and "relaxed" criteria respectively, which is among top-ranked systems of the 2014 i2b2 challenge. After integrating some refined localization dictionaries, our system is further improved with *F*-scores of 94.83%, 91.57% and 91.95% under the "token", "strict" and "relaxed" criteria respectively.



## 1. Introduction

With the development of electronic medical records (EMRs), more and more clinical data are generated. However, they cannot be freely used by companies, organizations and researchers because of a large amount of personally identifiable health information, known as protected health information (PHI), embedded in them. Using clinical data containing PHI is usually prohibited. De-identification, identifying and removing PHI, is a critical step in making clinical data accessible to more people. Since the Health Insurance Portability and Accountability Act (HIPAA) was passed in 1996 completely defined all kinds of PHI [1], de-identification has attracted considerable attention. De-identification resembles traditional named entity recognition (NER) tasks, but has its own property such that a word/phrase can be either a PHI instance or not. During the last decade, a large amount of effort has been devoted to de-identification including a challenge, i.e., the i2b2 (Center of Informatics for Integrating Biology and Bedside) clinical natural language processing (NLP) challenge in 2006, and various kinds of systems have been developed for de-identification [2–5]. However, no unified platform to evaluate systems on any PHI type defined in HIPAA.

In order to comprehensively investigate the performance of de-identification systems on every HIPAA-defined PHI type, the 2014 i2b2 clinical natural language processing (NLP) challenge sets up a new track to identify PHI instances in electronic medical records (EMRs) (track 1). In this track, seven main categories with twenty-five subcategories are defined, which cover all eighteen PHI types defined in HIPAA. In this paper, we describe our de-identification system for the 2014 i2b2 challenge. It is a hybrid system based on both machine learning and rule approaches. Evaluation on the independent test set provided by the challenge shows that our system achieves the highest micro *F*-scores of 94.64%, 91.24% and 91.63% under the "token", "strict" and "relaxed" criteria respectively, which is among top-ranked systems of the 2014 i2b2 challenge. We subsequently introduce refined localization dictionaries into our system, and marginally improve

* Corresponding author.
*E-mail addresses:* liuzengjian.hit@gmail.com (Z. Liu), tjcyx1995@163.com (Y. Chen), tangbuzhou@gmail.com (B. Tang), wangxl@insun.hit.edu.cn (X. Wang), qingcai.chen@gmail.com (Q. Chen), haodili.hit@gmail.com (H. Li), dr_wjf@hotmail.com (J. Wang), qiwendeng@hotmail.com (Q. Deng), 13809883596@163.com (S. Zhu).
[1] Contributed equally to this work.

http://dx.doi.org/10.1016/j.jbi.2015.06.009
1532-0464/

performance with micro *F*-scores of 94.83%, 91.57% and 91.95% under the "token", "strict" and "relaxed" criteria respectively.

## 2. Background

In the medical domain, many NLP approaches have been proposed for de-identification. The earliest de-identification system was proposed by Sweeney et al. in 1996 [6]. This system employed rules to identify twenty-five categories of personally-identifying information in pediatric EMRs. In the same year, the HIPAA was passed, and defined eighteen types of PHI. Subsequently, a large number of pattern matching-based systems were introduced for de-identification based on HIPAA. These systems used complex rules [7–12] and specialized semantic dictionaries [7,9,10,12] to perform de-identification. Most of them de-identified PHI in their own particular types of EMRs. For example, three systems were designed only for pathology reports [8–10]. Two systems were designed for multiple types of EMRs: Friedlin et al.'s [11] system for clinical notes including discharge summaries, laboratory reports and pathology reports, and Neamatullah et al.'s [12] system for nursing progress notes, discharge summaries and X-ray reports. Some pattern matching-based systems have been able to find around 99% PHI instances on their own datasets as reported [7,8,10,11]. However, we could not find which one is better due to no unified evaluation on publicly available datasets.

To accelerate de-identification research in the medical domain, the 2006 i2b2 clinical natural language processing (NLP) challenge issued a track to identify PHI in EMRs, which provided a unified platform to evaluate different systems. In this challenge, eight PHI categories were defined to annotate the challenge data from Partner Healthcare, only six HIPAA-defined categories. Seven teams participated in the challenge and developed de-identification systems using rule-based [13], machine learning-based [14–16] and hybrid methods [17,18]. Results showed that machine learning-based systems using rules as features performed best [2]. The machine learning algorithms used in these systems included conditional random fields (CRFs) [19], support vector machines (SVM) [20], decision trees (DTs) [21], and so on. Considering that all the documents used in this challenge were discharge summaries not annotated with all HIPAA-defined categories of PHI instances, Deleger et al. (2013) [5] evaluated a machine learning-based system using rules as features on various types of notes (over 22 types) annotated with all HIPAA-defined categories, although some of HIPAA-defined categories were collapsed into one category.

To further advance de-identification research in the medical domain, the 2014 i2b2 clinical NLP challenge organizers set up a track (track 1) to identify PHI in EMRs again. Different from the previous de-identification challenge, more refined PHI categories were annotated in the data provided by the organizers of this challenge, which makes it possible to evaluate all participating systems on every HIPAA-defined PHI type.

## 3. Material and methods

Fig. 1 shows an overview of our de-identification system for the 2014 i2b2 NLP challenge. It is a hybrid system based on both machine learning and rule approaches. The system contains two machine learning-based classifiers and a rule-based classifier. Similar to traditional NER tasks, the de-identification task is recognized as a sequence labeling problem in both two machine learning-based classifiers. In our system, PHI instances are first identified by two (token-level and character-level) conditional random fields (CRFs) and a rule-based classifier, and then are merged by some rules. The detailed description of the system is presented below.

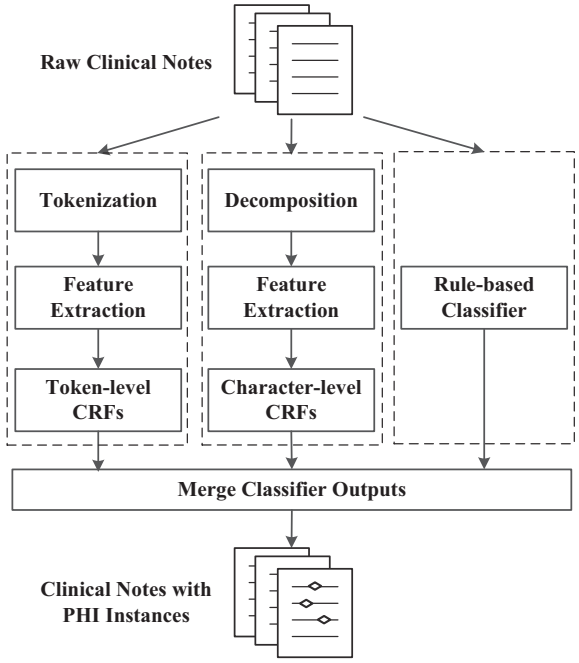


**Fig. 1.** Overview of our de-identification system for the 2014 i2b2 NLP challenge.

### 3.1. Dataset

In the 2014 i2b2 challenge, organizers manually annotated 1304 medical records of 297 patients according to the annotation guideline, and divided them into two parts: (1) 790 records of 188 patients used as a training set; and (2) the remaining 514 records of 109 patients used as a test set. 17,045 PHI instances in the training set and 11,462 PHI instances in the test sets are annotated using seven main categories with twenty-five subcategories that cover all HIPAA-defined PHI categories. The numbers of PHI instances of main categories in both two sets are listed in Table 1, where NA denotes no subcategory, numbers in parentheses in the first row are the numbers of categories and PHI instances, and asterisks indicate the HIPAA-defined categories. To get more detailed information of the dataset, please refer to the overview paper [22,23].

### 3.2. Machine learning-based classifiers

There are two (token-level and character-level) machine learning classifiers in our de-identification system, and both trained by conditional random fields (CRFs) algorithm. The main difference between those two classifiers is the representation of features. We use CRFsuite (http://www.chokkan.org/software/crfsuite/) as the implementation of CRFs, and optimize parameters of the two machine learning classifiers by 10-fold cross-validation on the training set.

#### 3.2.1. PHI instance representation

How to represent PHI instances is the chief problem we should solve in machine learning-based de-identification systems. In our system, two typical NER representation schemas are used to represent PHI instances: "BIO" and "BIOES", where 'B', 'I', 'O' and 'E' denote that a token/character is at the beginning, middle, outside and end of an instance, and 'S' denotes that a token/character itself is an instance. Fig. 2 shows examples of PHI instances represented by "BIO" and "BIOES" at token-level. The PHI instances are represented in the similar way at character-level. Our evaluation shows