Chapter 48
Patient Data De-Identification: A Conditional Random-Field-Based Supervised Approach

Shweta Yadav
Indian Institute of Technology Patna, India

Sriparna Saha
Indian Institute of Technology Patna, India

Asif Ekbal
Indian Institute of Technology Patna, India

Parth S Pathak
ezDI, LLC, India

Pushpak Bhattacharyya
Indian Institute of Technology Patna, India

ABSTRACT

With the rapid increment in the clinical text, de-identification of patient Protected Health Information (PHI) has drawn significant attention in recent past. This aims for automatic identification and removal of the patient Protected Health Information from medical records. This paper proposes a supervised machine learning technique for solving the problem of patient data de- identification. In the current paper, we provide an insight into the de-identification task, its major challenges, techniques to address challenges, detailed analysis of the results and direction of future improvement. We extract several features by studying the properties of the datasets and the domain. We build our model based on the 2014 i2b2 (Informatics for Integrating Biology to the Bedside) de-identification challenge. Experiments show that the proposed system is highly accurate in de-identification of the medical records. The system achieves the final recall, precision and F-score of 95.69%, 99.31%, and 97.46%, respectively.

INTRODUCTION

With the start of the golden era in the medical interpretation, the vast amount of information in the clinical domain is increasing at a rapid rate. In the past decade, with the development of the health information technology and health data documentation, there has been progress in how heath care is performed (Berner et al., 2005).

DOI: 10.4018/978-1-7998-0951-7.ch048

With the widespread use of health information technology, there has been huge pace in the increment of clinical data in addition to the fast adoption of the Electronic Clinical Records and with the conversion of narrative data to the electronic form. The amount of information can be improved further with the minimization of the medical error. This requires the development of some sophisticated tools for Medical Language Processing (MLP). Most medical records are in the narrative forms which are formed as the result of transcription of dictations, direct entry by providers, or use of speech recognition applications. However, their use in this form is restricted to any organization or research, as medical records have a sufficient number of personal health information or protected health information (PHI). According to Health Insurance Portability and Accountability Act (HIPAA), 1996, the PHI terms need to be enclosed and protected. This has lead to de- identification problem. Paragraph 164.514 of the Administrative Simplification Regulations promulgated under the Health Insurance Portability and Accountability Act (HIPAA) states that for data to be treated as de-identified, it must clear one of two hurdles (HIPPA ACT 1996).

1. An expert must determine and document "that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."
2. Or, the data must be purged from a specified list of seventeen categories of possible identifiers relating to the patient or relatives, household members and employers, and any other information that may make it possible to identify the individual.

Studies showed that there was a significant drop in the patient consent request reducing the participation rate and also, this is quite infeasible for the huge population. Even, in the case when a patient provides the permission, documents must be tracked to stop any unauthorized disclosure. This emerging problem of consent, waiver, and tracking can be effectively handled if the patient personal health information is properly de-identified facilitating the clinical NLP research (Wolf & Bennett, 2006).

De-identification task is more specifically defined as the step where the private information is removed or replaced while keeping the record as it is (Stubbs et al., 2015). De-identification is a type of traditional named entity recognition (NER) problem, with the property of defining a term to be PHI type or not. The main aim of de-identification challenge as pointed out earlier is to remove the PHI terms maintaining data integrity as much as possible. Every record is enclosed in the RECORD_ tags and is provided a unique ID which is randomly generated. Figure 1 shows Sample Discharge Summary Excerpt; a sample discharge summary from the training dataset where the goal is to identify the PHI (private health information) terms. In this summary, some the PHI terms are doctors' name ("Dr. Do Little"), patient name ("John Doe") and hospital name ("ABHG", "SBHG"). A TEXT_ tag encloses the text of different records. Each PHI instance is enclosed within PHI_ tags and the PHI TYPE represents the category of the PHI term as shown in Figure 1.

As shown in Figure 1, the task is to enclose the PHI terms such as ("Dr. Do Little", "John Doe", "ABHG", "SBHG"). This task can be seen as the typical sequence labeling task where for e.g. "Dr. Do Little" should be labeled as whole with Doctor. Here, we present an example where input forms the sequence of words and output is the label sequence and its corresponding de-identified sentence. "BIO" notation was followed to label the NE where "B" represents the beginning of label sequence, "I" denotes intermediate of label sequence and "O" represents others.