Automated Anonymization as Spelling Variant Detection

Steven Kester Yuwono **Hwee Tou Ng**
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
{kester,nght}@comp.nus.edu.sg

Kee Yuan Ngiam
Department of Surgery
National University Hospital
5 Lower Kent Ridge Road
Singapore 119074
kee_yuan_ngiam@nuhs.edu.sg

Abstract

The issue of privacy has always been a concern when clinical texts are used for research purposes. Personal health information (PHI) (such as name and identification number) needs to be removed so that patients cannot be identified. Manual anonymization is not feasible due to the large number of clinical texts to be anonymized. In this paper, we tackle the task of anonymizing clinical texts written in sentence fragments and which frequently contain symbols, abbreviations, and misspelled words. Our clinical texts therefore differ from those in the i2b2 shared tasks which are in prose form with complete sentences. Our clinical texts are also part of a structured database which contains patient name and identification number in structured fields. As such, we formulate our anonymization task as spelling variant detection, exploiting patients' personal information in the structured fields to detect their spelling variants in clinical texts. We successfully anonymized clinical texts consisting of more than 200 million words, using minimum edit distance and regular expression patterns.

1 Introduction

Clinical discharge summaries are an essential source of information to facilitate medical research. However, they contain patients' personal health information (PHI) which, if disclosed, would compromise patients' privacy. Various techniques have been applied to create de-identification systems and they have performed well (Uzuner et al., 2007). These de-identifier systems utilize either machine learning approaches such as support vector machines (Uzuner et al., 2008), conditional random fields (Wellner et al., 2007), and decision trees (Szarvas et al., 2007), or rule-based approaches with pattern matching (Douglass et al., 2004).

In this paper, we tackle the task of anonymizing clinical discharge summaries written in English from the National University Hospital in Singapore. Our work is novel in the following aspects: (1) Our clinical discharge summaries are written in sentence fragments and they frequently contain symbols, abbreviations, and misspelled words, unlike the clinical texts in the i2b2 shared tasks which are in prose form with complete sentences. (2) We treat anonymization as a spelling variant detection task, by exploiting patient health information stored in structured fields. (3) We have applied our anonymization algorithm on actual hospital discharge summaries containing more than 200 million words. Manual evaluation on a sample test set shows that our algorithm achieves very high recall.

2 Task Description

The corpus of hospital discharge summaries used in this paper is obtained from the National University Hospital, spanning a period of ten years. The patients in these discharge summaries came from a variety of countries with varied names from different races and cultures. In all, there are about 570,000 discharge summaries with a total size of more than 700MB. Each discharge summary has an average of 400 word tokens. Given a discharge summary, the anonymization task is to remove patients' PHI which includes the following items: names of patients; identification numbers; telephone, fax, and pager numbers;

Proceedings of the Clinical Natural Language Processing Workshop,
pages 99–103, Osaka, Japan, December 11-17 2016.

geographical locations; dates; and names of doctors and hospitals. It is highly improbable that a patient can be identified without the personal information listed above. Any PHI detected will be replaced by an appropriate surrogate, e.g., a patient name will be replaced by PNAME, a patient identification number will be replaced by PID, etc.

As mentioned earlier, our discharge summaries are written in sentence fragments and organized in bullet points. They frequently contain symbols, abbreviations, and misspelled words. As such, our discharge summaries are significantly different from those in the i2b2 shared tasks in 2006 (Uzuner et al., 2007) and 2014 (Stubbs et al., 2015), which are in prose form with complete sentences. Samples of discharge summaries from our corpus and from the i2b2 shared task in 2006 are given below.

```
This 68 year old female had rheumatic
    fever in the past , and has had
    chronic atrial fibrillation .
She has had progressive heart failure
    and an evaluation demonstrated
    worsening mitral stenosis with
    severe pulmonary hypertension .
Because of her deteriorating status ,
    she underwent prior cardiac
    catheterization , which confirmed
    severe mitral stenosis with
    secondary tricuspid valve
    regurgitation due to pulmonary
    hypertension .
She was referred for valve surgery .
She had undergone a previous nasal
    arterial embolization for treatment
    of recurrent epistaxis .
She had a partial gastrectomy in 1972 .
Her MEDICATIONS ON ADMISSION included
    Coumadin , digoxin , 0.125 , qD ,
    Lasix , 40 , q.i.d. , and Vanceril
    inhaler .
```

A sample discharge summary snippet from the i2b2 de-identification challenge in 2006.

```
33/Chinese/M
PMHX:
- anemia
- previously on iron supplement
- nil OGD done

Currently c/o:
epigastric pain 1500H
nil nasuea / vomiting
nil fever noted
nil dysuria / hematuria
no changes in bowel movement
no LOW/LOA
no chest pain or SOB

O/E on admission:
Pt alert, attentive
CVS: PR 78/min, Bp 120/70 S1S2 no
    murmurs, TWC 14 UC10 - nad
soft abdo, normoactivew BS. direct and
    rebound tenderness RIF. nil guarding
    . nil rebound

Imperssion: Acute appendicitis
Pt was sent for op
```

A sample discharge summary snippet from our hospital.

In our discharge summary snippet above, the words nasuea and Imperssion are misspelled words. Pt, PMHX, LOW, and LOA are abbreviations for patient, past medical history, loss of weight, and loss of appetite respectively. As such, our discharge summaries pose additional challenges to anonymization and to subsequent processing by downstream natural language processing modules like part-of-speech tagging, coreference resolution, etc.

In addition, our hospital discharge summaries are part of a structured database which contains patients' PHI such as names, identification numbers, phone numbers, etc. in structured fields. As such, we exploit the meta-data in these structured fields and formulate our anonymization task as spelling variant detection. That is, the objective of our anonymization task is to find spelling variants of patient names and other PHI items and replace them with appropriate anonymized surrogates. This is in contrast to the i2b2 shared tasks, where external structured information is not utilized. Since hospitals are required to keep track of patients' PHI in addition to their discharge summaries, admission notes, etc, one can expect structured PHI items of a patient to be available in a real-world setting when processing the discharge summary of a patient. As such, the anonymization task that we address is a more realistic one.

3 Anonymization Algorithm

Our anonymization algorithm uses regular expression matching and the minimum edit distance algorithm to identify spelling variants, assuming that patients' PHI stored in the structured database is correct.

3.1 Patient Name

Patient name is the most important personal information present in a discharge summary. Even a misspelled patient name may be used to trace and identify a patient. A patient's full name associated with a discharge summary is first taken from the structured field in the database. The full name is first split into individual name tokens. Each word in a discharge summary is compared against each name token of the patient. The minimum edit distance algorithm (Wagner and Fischer, 1974) is used to compute the minimum edit distance between a name token n from the structured field and a candidate word w in the discharge summary. We set the insertion, deletion, and replacement cost to 1. The edit distance ratio R is computed as $\frac{d}{min(|n|,|w|)}$, where d is the minimum edit distance of n and w. Since a longer name has a higher probability of being misspelled than a shorter one, we use R to take into account the length of a string. If R is less than a specified threshold, the current candidate word w will be taken as the patient's name, and will be anonymized and replaced by a surrogate. We set the threshold to be 0.33.

A person's name is often preceded by an honorific (a title prefixing a person's name). As such, we replace the word after an honorific by a surrogate. The list of honorifics used in our anonymization algorithm is as follows: `mr, mrs, miss, ms, madam, mdm, lady, sir, col, dr, doctor, a/prof, e/prof, professor, prof, general, gen, senator, sen.` By detecting the honorifics, our anonymization algorithm is able to detect names that might otherwise be missed by the minimum edit distance algorithm.

3.2 Identification Number and Contact Number

To detect a patient's identification number and contact numbers, we make use of regular expressions that capture the generic formats of patients' identification numbers and contact numbers. The format of patient identification numbers in our hospital consists of fixed numbers of letters and digits arranged in a fixed order, which can be readily detected by a regular expression. Similarly, the format of contact numbers consists of digits interspersed with space or dash ("-") characters, which again can be readily detected by a regular expression.

3.3 Date

Anonymization of dates is challenging because there are many possible date formats. Days can be written in single or double-digit. Months can be written in single-digit, double-digit, short name (e.g., Jan), or long name (e.g., January). Years can be written in double-digit or four-digit. The delimiters allowed between day, month, and year include dash (-), comma (,), slash (/), colon (:), and white space (space and tab). Therefore we have created regular expressions for all possible combinations of the date format to cover all possibilities: day/month/year, month/day/year, year/day/month, year/month/day, day/month, month/day, year/month, and month/year.

3.4 Doctor's Name, Hospital's Name, and Geographical Location

Most doctors' names are handled by patient name anonymization above due to the common occurrences of "dr" or "prof" preceding a doctor's name. In addition, we obtain a list of names of doctors, hospitals, and geographical locations in Singapore. For each entry in the list, we check if it is present in a discharge summary and replace it by a surrogate if found.

4 Evaluation

One key advantage of our anonymization algorithm that relies on regular expression matching and the minimum edit distance algorithm is that manual annotation of training data is *not* required, unlike in a machine learning approach. To evaluate the performance of our anonymization algorithm, 100 discharge summaries were randomly selected as the test set. The accuracy of our anonymization algorithm is reported in Table 1.

[1]Patients' identification numbers and contact numbers

[2]Names of doctors, hospitals, and geographical locations

	Patient name	ID num[1]	Date	Other names[2]	Overall
Recall	100	100	100	93.14	97.35
Precision	85.94	100	66.03	76.71	72.67
F1-score	92.44	100	79.50	84.13	83.22
PHI count	110	29	418	350	907

Table 1: Token-level evaluation of our anonymization algorithm (in %).

Our anonymization algorithm has achieved good performance. In particular, it achieves 100% recall on anonymizing patients' names, identification numbers, and contact numbers. We favor recall over precision, since it is highly critical that personal information of patients be completely anonymized, at the cost of some false positives. The anonymization algorithm fails to detect some other names, such as doctors' names which are not present in the given list of doctors' names. Most of the false positives are contributed by some common names of doctors, and how time duration is written in the discharge summaries. To illustrate, consider the following sentence fragment: `vomiting 2/7, LOW 1/12.` `2/7` is falsely detected as a date (meaning 2 days). `1/12` is falsely detected as a date (meaning 1 month). `LOW` is falsely detected as a doctor's name, because `LOW` is a common family name in Singapore.

Our anonymization algorithm runs efficiently. It anonymizes 7 discharge summaries per second, and takes 21.7 hours to anonymize the whole corpus of discharge summaries consisting of more than 200 million words on a PC with 3.4 GHz processor in a single thread.

We have also attempted to use a machine learning approach, in particular a maximum entropy classifier, to carry out anonymization. The classifier uses the edit distance ratio as the main feature, and other additional features such as part-of-speech tags, named entity tags, binary features about the presence of a preceding honorific and whether the current word is an English word. However, preliminary experiments indicate that the maximum entropy classifier does not outperform our current anonymization algorithm of regular expression matching and the minimum edit distance algorithm. As such, we adopt our current algorithm which is simpler and requires no annotated training data.

There were several prior systems which focused on the detection or removal of certain types of PHI such as patient names (Taira et al., 2002), or both patient and doctor names (Thomas et al., 2002). However, they did not exploit knowledge of external structured information like patient names or other PHI to be removed. There were also several studies that used patients' structured fields to perform de-identification using regular expressions and lexical look-up tables (Neamatullah et al., 2008), string similarity algorithm to detect typographical errors (Friedlin and McDonald, 2008), and a combination of rule-based and machine learning approaches for de-identification (Ferrández et al., 2013). However, the performance of these systems cannot be directly compared to ours because of different test data.

5 Conclusion

In this paper, we tackle the task of anonymizing discharge summaries written in sentence fragments and which frequently contain symbols, abbreviations, and misspelled words. Our discharge summaries are therefore substantially different from the discharge summaries dealt with in the i2b2 shared tasks. We also exploit PHI of patients present in structured database fields and present a novel approach that treats anonymization as spelling variant detection. Our anonymization algorithm effectively and efficiently anonymizes more than 200 million words of actual hospital discharge summaries, achieving a very high recall.

Acknowledgments

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 1 grant T1-251RES1513.

References

M Douglass, GD Clifford, A Reisner, GB Moody, and RG Mark. 2004. Computer-assisted de-identification of free text in the MIMIC II database. In *Computers in Cardiology 2004*, pages 341–344.

Oscar Ferrández, Brett R South, Shuying Shen, F Jeffrey Friedlin, Matthew H Samore, and Stéphane M Meystre. 2013. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *Journal of the American Medical Informatics Association*, 20(1):77–83.

F Jeff Friedlin and Clement J McDonald. 2008. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15(5):601–610.

Ishna Neamatullah, Margaret M Douglass, Li-wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8:32.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19.

György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. 2007. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association*, 14(5):574–580.

Ricky K Taira, Alex AT Bui, and Hooshang Kangarloo. 2002. Identification of patient name references within medical documents using semantic selectional restrictions. In *Proceedings of the AMIA Symposium*, page 757.

Sean M Thomas, Burke Mamlin, Gunther Schadow, and Clement McDonald. 2002. A successful technique for removing names in pathology reports using an augmented search and replace method. In *Proceedings of the AMIA Symposium*, page 777.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Özlem Uzuner, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. 2008. A de-identifier for medical discharge summaries. *Artificial intelligence in Medicine*, 42(1):13–35.

R. A. Wagner and M. J. Fischer. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21:168–173.

Ben Wellner, Matt Huyck, Scott Mardis, John Aberdeen, Alex Morgan, Leonid Peshkin, Alex Yeh, Janet Hitzeman, and Lynette Hirschman. 2007. Rapidly retargetable approaches to de-identification in medical records. *Journal of the American Medical Informatics Association*, 14(5):564–573.