Ao Li. 798657

aol3@student.unimelb.edu.au

PROJECT PROPOSAL

Clinical Text De-identification and Colander

# 1 Introduction

With the development of Electronic Health Records (EHR) system, a large amount of patient data nowadays can be readily accessed and provided through clinical text for several various background people, for example, clinicians for diagnosis, professors and authorities for statistics analysis (Meystre et al, 2010). Also, it is worth noting that these generous clinical texts contain a great deal of patients’ confidential information, which are the information like Name, Age, Address, Phone number …. etc. that do not want the other people know.

In this project, it will focus on eliminating the problem of leaking confidential information from the clinical text problem and there are two aspects of method will be proceeded in the project: 1) De-identification approach and 2) Colander approach.

# 

# 2 Methodology

Undeniably, there are many issues around the original clinical text. Two major problem are narrative clinical texts are generally unstructured and different EHR systems provide distinctive framework of clinical text (Ferrandez et al. 2012, p.88). So, for the sake of avoiding those problem, this project will only use i2b2 (Informatics for Integrating Biology and the Bedside, 2016) dataset for the whole experiment. The reasons for choosing i2b2 dataset are: 1) i2b2 dataset has provided sets of clinical notes from the Research Patient Data Repository at Partners HealthCare for general research purposes and it has become increasingly use to explore the value locked in unstructured text that is essential to the success of research. 2) It has already tagged the Protected Hospital Identifier (PHI), so it can help for the measurement.

## 2.1 De-identification Approach

De-identification approach is generally a process to identify and remove the patients’ confidentiality information from the clinical texts so as to protect their personal details. In this project, this general idea has been improved by drawing the concept of external specific dictionary into and the overview processes are as following:

1. Establishing the external specific dictionaries:

For each concerned PHI category, part of the training data (e.g. 80% of the random training data) and methods like cross-validation can be used for generating a specific dictionary. For example, if a training clinical text is like *“Essentially, Mr. Cornea is a 60 year old male who noted the onset of dark urine during early January.”*. “Cornea” can be added into the “Name” dictionary. So, when later in the test clinical text, it is much easier to recognize “Cornea” as a name and remove it from the text. Multiple various dictionaries are established from the training data and each dictionary is only for one particular PHI category.

1. Using tools for de-identification:

There are four de-identification tools chosen to use in this project, which are HMS Scrubber (Beckwith, 2006), MIT de-id software package (Neamatullah, 2008), PHI reducer (PHI-Reducer, 2017) and MIST (Aberdeen, 2010).

1. Measure the performance for de-identification:

The comparison among tools and among external dictionaries will be introduced in this part. For instance, compare the tools’ performance when with the external dictionaries to the without situation or compare the impact of different dictionaries on one specific tools… etc. Also, 1) Recall: the number of correct de-identified terms over all PHI terms.

and 2) Precision: the number of correct de-identified terms over all de-identified terms. will be considered as the capability of systems.

## 2.2 Colander Approach

Colander Approach is an opposite way of processing to the de-identification approach. The key idea is to pre-process and only keep the terms that we want in the clinical texts, which means that it considers all other terms as the PHI and then try to get rid of them. The overview processes in this project are as below:

1. Establish protected word dictionaries:

For each concerned problem scenario, create a protected word dictionaries. For example, consider the scenario about ‘the correlation between smoking and cancer’, manually or import online related open source tools (e.g. cTakes, MetaMap etc.) to create the protected word dictionary that contains several words related to ‘smoking’ and ‘cancer’.

1. Using program for colander:

Writing a program to remove all the words that are not in the protected word dictionary from the original clinical text.

1. Analysis the result for each scenario

Because colander method cannot be measured exactly, some guess can be introduced from the results. For instance, the total number of patients are only related to smoking, the total number of patients are only related to cancer and the total number of patients are related both smoking and cancer can be summed up. Then compare the number of each category and analyse if there is a relationship between them.

# 3 Time Schedule and Preliminary Results

To ensure the quality of each part in the project, the following plan will be considered:

|  |  |
| --- | --- |
| Time Schedule | Preliminary Results |
| 01/03/2018 - 01/04/2018 | Complete all of the experiment on de-identification approach part. |
| 01/04/2018 - 01/05/2018 | Complete all of the experiment on colander approach part. |
| 01/05/2018 - 01/06/2018 | Do the extension and complete the whole report. |

# 4 Reference

Aberdeen, J, Bayer, S, Yeniterzi, R, Wellner, B, Clark, C, Hanauer, D, Malin, B & Hirschman, L, 2010, ‘The MITRE Identification Scrubber Toolkit: design, training, and assessment’, International Journal of Medical Informatics, vol.79, no.12, pp.849-859.

Beckwith, BA, Mahaadevan, R, Balis, UJ, & Kuo, F, 2006, ‘Development and evaluation of an open source software tool for de-identification of pathology reports’, *BMC Medical Informatics and Decision Making*, vol.6, no.1, pp.12.

Ferrandez, O, South, BR, Shen, S, Friedlin, FJ, Samore, MH & Meystre, SM, 2012, ‘BoB, a best-of-breed automated text de-identification system for VHA clinical documents’, *American Medical Informatics Association,* vol.20, no.1, pp.88-83.

Informtics for Integrating Biology and the bedside, 2017, viewed 15 March 2018, <http;//www.i2b2.org>.

Meystre, SM, Friedlin, FJ, South, BR, Shen, S, & Samore, MH, 2010, ‘Automatic de-identification of textual documents in the electronic health record: a review of recent research’, *Bmc Medical Research Methodology,* vol.10, pp.70.

Neamatullah, I, Douglass, MM, Liwei, HL, Resiner, A, Villarroel, M, Long, WJ, Szolovits, P, Moody, GB, Mark, RG, and Clifford, GD, 2008, ‘Automated de-identification of free-text medical records’, *BMC Medical Informatics and Decision Making*, vol.8, no.1, pp.32.

PHI-Reducer, 2017, viewed 15 March 2018, <https://pypi.python.org/pypi/phi-reducer>