Artificial Intelligence in Medicine xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

De-identification of health records using *Anonym*: Effectiveness and robustness across datasets

Guido Zuccon[a,b,*], Daniel Kotzur[a], Anthony Nguyen[a], Anton Bergheim[c]

[a] *The Australian e-Health Research Centre (Commonwealth Scientific and Industrial Research Organisation), Level 5 – UQ Health Sciences Building 901/16, Royal Brisbane and Women's Hospital, Herston, QLD 4029, Australia*
[b] *School of Information Systems, Queensland University of Technology, Y Block Level 6, Gardens Point Campus, Brisbane, QLD, Australia*
[c] *Cancer Institute NSW, Australian Technology Park, Level 9, 8 Central Avenue, Eveleigh, NSW 2015, Australia*

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form 17 March 2014
Accepted 18 March 2014

Keywords:
Conditional random fields
Pattern matching
De-identification
Health records

ABSTRACT

Objective: Evaluate the effectiveness and robustness of Anonym, a tool for de-identifying free-text health records based on conditional random fields classifiers informed by linguistic and lexical features, as well as features extracted by pattern matching techniques. De-identification of personal health information in electronic health records is essential for the sharing and secondary usage of clinical data. De-identification tools that adapt to different sources of clinical data are attractive as they would require minimal intervention to guarantee high effectiveness.
Methods and materials: The effectiveness and robustness of Anonym are evaluated across multiple datasets, including the widely adopted Integrating Biology and the Bedside (i2b2) dataset, used for evaluation in a de-identification challenge. The datasets used here vary in type of health records, source of data, and their quality, with one of the datasets containing optical character recognition errors.
Results: Anonym identifies and removes up to 96.6% of personal health identifiers (recall) with a precision of up to 98.2% on the i2b2 dataset, outperforming the best system proposed in the i2b2 challenge. The effectiveness of Anonym across datasets is found to depend on the amount of information available for training.
Conclusion: Findings show that Anonym compares to the best approach from the 2006 i2b2 shared task. It is easy to retrain Anonym with new datasets; if retrained, the system is robust to variations of training size, data type and quality in presence of sufficient training data.

1. Background

Electronic health records (EHRs) often contain personal health information (PHI) that can uniquely identify a patient. The United States's Health Information Portability and Accountability Act (HIPAA) has stipulated 17 categories of PHIs that must be de-identified, the most prevalent are outlined in Table 1.

Access to EHRs outside of the primary health provider and the sharing of such data for research purposes is fundamental for critical data mining and information retrieval tasks in the health domain; for example, the identification of adverse drug reactions or patient recruitment for clinical studies [1,2]. However, PHIs are pervasive in unstructured portions of EHRs, which undermines access and sharing of such important data [3].

De-identification is the process of removing PHIs from medical records. Manual de-identification of electronic health records is time and resource consuming. Dorr et al. [4] found that on average 87.2 ± 61 s are required to manually de-identify a narrative text of an EHR; an EHR on average contains 7.9 ± 6.1 PHI entities.

Anonym is a software tool developed at the Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation (CSIRO), that automatically de-identifies EHRs. Anonym is based on the combination of conditional random fields (CRF) classifiers, informed by a number of linguistic and lexical features and pattern matching techniques. The de-identification method used in Anonym is described in Section 3. The results of the empirical evaluation reported in Section 5 shall show that, if enough training data is provided, Anonym is capable of effectively de-identify free-text EHRs, irrespective of type, source or quality of data. In addition, results demonstrate that Anonym is comparable to the best state-of-the-art de-identification system proposed

* Corresponding author at: School of Information Systems, Queensland University of Technology, Y Block Level 6, Gardens Point Campus, Brisbane, QLD, Australia. Tel.: +61 (0)731381412.
E-mail addresses: g.zuccon@qut.edu.au, zuccon.guido@gmail.com (G. Zuccon).

http://dx.doi.org/10.1016/j.artmed.2014.03.006
0933-3657/

Table 1
Subset of the United States's Health Information Portability and Accountability Act personal health identifiers types considered for the evaluation of Anonym.

PHI type	Meaning
Patients	First, middle and last names of patients and their family members (including initials of names).
Doctors	Similar to patients category, includes names and initials of health professionals.
Dates	All numerical and literal reference to dates, including years and days of the week.
Hospitals	Names of medical facilities and practices.
IDs	Any combination of digits and letters that refer to medical records, patient numbers, accession numbers, doctors identifiers, laboratory identifiers, etc.
Locations	Names of cities, regions and states, as well as addresses, zip codes and building names.
Phone numbers	Any reference to landline, fax and mobile phone numbers or phone extension numbers.

by Uzuner et al. [5]. In addition, the results also demonstrate that retraining is necessary when changing datasets.

2. Related work

Two areas of related work are reviewed: de-identification and named entity recognition.

2.1. De-identification

Research on de-identification of EHRs has flourished as a result of the introduction of the 2006 Integrating Biology and the Bedside (i2b2) dataset shared task [5]. This shared task provided an evaluation framework for de-identification, consisting of a dataset of manually annotated medical discharge summaries populated with ambiguous PHIs and metrics to measure the performance of de-identification systems. Uzuner et al. [5] provide an overview of systems that participated in i2b2. The techniques used by participants included conditional random fields, rule-based approaches, hidden Markov models (HMM), and support vector machines. The best system in i2b2 was developed by Wellner et al. [6]. Their system is similar to the approach considered in this paper: both use CRFs to label tokens and regular expressions to form one of the feature classes. However, our approach differs in that we do not use lexicons of locations and English words and we consider additional features such as part of speech.

Uzuner et al. [7] have studied the role of local context (i.e. the words that are immediate neighbours of the target PHI or that have immediate syntactic relation with it) for de-identification when using support vector machine classifiers. They observed that features that thoroughly capture local context are beneficial to the PHI de-identification task. While not relying on local context features as thoroughly as Uzuner et al. [7], Anonym does use features that implicitly capture local context information, such as token n-grams and part-of-speech.

An overview of approaches to PHI de-identification is provided by Meystre et al. [8]. From their analysis, they concluded that methods based on linguistic resources, such as dictionaries, tend to perform better with rarely mentioned PHIs. Vice versa, they found that machine learning techniques better generalise to PHIs that are not mentioned in dictionaries, although machine learning tends to have problems identifying PHI types that rarely occur in the training corpus. Rule-based techniques and machine learning algorithms have been recently integrated in the stepwise hybrid approach proposed by Ferrandez et al. [9]. Anonym uses rule-based techniques in its pattern-matching component for feature generation.

Recent work has focused on semi-supervised or iterative approaches that improve the human-supervised de-identification workflow process as a whole, rather than producing a fully automatic de-identification system. Hanauer et al. [10] constructed statistical de-identification models by iteratively performing (i) annotation of a small EHRs sample; (ii) training of a CRF model; (iii) automatic identification of PHIs on a small sample of unseen data; (iv) manual correction of the errors on the unseen data; and (v) retraining of the model. Boström and Dalianis [11] used active learning to train a random forest classifier to detect PHIs from Swedish EHRs. They also investigated different strategies to select the most discriminative samples for online manual annotation.

In a previous paper [12], we presented the approach underlying Anonym and initial results that showed our tool is comparable to state-of-the-art approaches on the 2006 i2b2 shared task. In that work, we have also briefly investigated the effectiveness on a small set of pathology reports supplied by an Australian cancer registry. This article extends that work by considering (1) additional datasets, including a larger set of cytology and pathology reports from a statewide Australian cancer registry and 1885 clinical notes from the MTSamples dataset [13]; (2) further investigation of the adaptability of Anonym across the different datasets.

2.2. Named entity recognition and conditional random fields

De-identification is a specialisation of named entity recognition (NER), i.e., the task of recognising references in text to information units like names (e.g., persons, organisations, locations) and numeric expressions (e.g., dates, money). While early NER systems were based on highly engineered rules, the most recent and successful approaches adopt supervised machine learning to automatically induce recognition rules from a corpus of training examples. Popular supervised algorithms for NER include HMMs, decision trees, maximum entropy, support vector machines and CRF. A survey of NER models, common features, and evaluation techniques is given by Nadeau and Sekine [14].

Anonym is based on the conditional random fields approach to learn PHIs and then identify new occurrences of PHIs from unseen data. A CRF is a discriminative undirected probabilistic graphical model that, given an observed sequence, defines a log-linear distribution over labelled sequences [15]. Mathematically, given an observed sequence x, a CRF predicts a label y from the set of possible labels Y if y maximises the conditional probability $p(y|x)$, i.e., if $p(y|x)$ is greater than any $p(y^*|x)$, for all y^* in $Y \setminus \{y\}$. This conditional nature of CRF is the key characteristic distinguishing CRF from HMM; it also means that the independence assumption necessary to ensure tractable inference in HMM is relaxed in the CRF approach.

The CRF approach underneath Anonym uses, among others, features generated by a set of pattern matching rules (regular expressions). This feature generation approach is similar to that of Collins [16], who introduced pattern features that map tokens onto a set of patterns.

3. Anonym: de-identifying EHRs with CRF and pattern matching

Anonym consists of three main modules: (i) the automatic feature generation component, (ii) the model training component that uses the features generated by the first module, and (iii) the classification component which applies the learnt model to unseen data. A fourth module is responsible for the generation of PHI surrogates consistent with those identified and the replacement of the identified PHI with its surrogate. This component has not been used to post-process the PHIs identified by Anonym in this work. Instead, we used this component to pre-process the data of two of the three datasets considered here as they could not be distributed with the

Please cite this article in press as: Zuccon G, et al. De-identification of health records using *Anonym*: Effectiveness and robustness across datasets. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.03.006