Telematics and Informatics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Telematics and Informatics

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

DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text

Vincent Menger[a,*], Floor Scheepers[b], Lisette Maria van Wijk[a], Marco Spruit[a]

[a] Department of Information and Computing Sciences, Utrecht University, P.O. Box 80089, 3508 TB Utrecht, The Netherlands
[b] Department of Psychiatry, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, The Netherlands

ARTICLE INFO

Keywords:
De-identification
Dutch medical text
Pattern matching
Protected Health Information
Patient privacy

ABSTRACT

In order to use medical text for research purposes, it is necessary to de-identify the text for legal and privacy reasons. We report on a pattern matching method to automatically de-identify medical text written in Dutch, which requires a low amount of effort to be hand tailored. First, a selection of Protected Health Information (PHI) categories is determined in cooperation with medical staff. Then, we devise a method for de-identifying all information in one of these PHI categories, that relies on lookup tables, decision rules and fuzzy string matching. Our de-identification method DEDUCE is validated on a test corpus of 200 nursing notes and 200 treatment plans obtained from the University Medical Center Utrecht (UMCU) in the Netherlands, achieving a total micro-averaged precision of 0.814, a recall of 0.916 and a F_1-score of 0.862. For person names, a recall of 0.964 was achieved, while no names of patients were missed.

1. Introduction

Data from Electronic Health Records (EHR) is since long being used for medical research purposes, and with the increasing digitalization in the medical world even more so (Milovic, 2012; Murdoch and Detsky, 2013). Since many hospitals today have adopted an EHR system, the produced data can be used for clinical research with few further limitations (Koh and Tan, 2005; Lee et al., 2013). Both health care institutions and patients can directly benefit from the results of this kind of research that can improve diagnosis, treatment, hospital operations and more (Jensen et al., 2012; Menger et al., 2016). The use of patient data for research however puts a strain on patient privacy, since this requires getting the data out of their health care context (Patel et al., 2014). This entails for example copying the data to different databases, where it can be accessed by data analysts (Simon et al., 2000). Technical staff such as data managers or data analysts typically do not have a treatment relation with the patient, and therefore should not be able to identify individual patients in a research dataset. Medical staff on the other hand is allowed to see patient information under medical confidentiality, but lack the technical skills to perform advanced analysis that is needed for obtaining direct clinical value from data.

Multiple ways exist to solve this problem for structured data. One rigorous approach is to remove all variables that could identify a person, such as patient names, addresses and social security numbers from the dataset (El Emam et al., 2006). A more sophisticated solution is a k-anonymity method, where the information of a patient is indistinguishable from at least $k-1$ individuals (Toledo and Spruit, 2016). In recent years however, the more widespread availability and quality of text mining approaches have shifted attention to analyzing unstructured textual data in addition to structured data. It is becoming ever more apparent that these approaches offer

* Corresponding author.
E-mail addresses: v.j.menger@uu.nl (V. Menger), f.e.scheepers-2@umcutrecht.nl (F. Scheepers), l.m.vanwijk2@students.uu.nl (L.M. van Wijk), m.r.spruit@uu.nl (M. Spruit).

http://dx.doi.org/10.1016/j.tele.2017.08.002
Received 31 December 2016; Received in revised form 8 May 2017; Accepted 1 August 2017

substantial benefits for data driven research (e.g. Harpaz et al., 2014; Patel et al., 2014; Maenner et al., 2013), and that textual data from the medical domain holds valuable information that should not be disregarded. Therefore, if we require research datasets to be anonymous to mitigate the potential negative impact on patient privacy, we must also de-identify the medical free text variables.

From a patient perspective, protecting the private details of a disease from the public is essential in retaining the trust bond between a physician and the patient (Krishna et al., 2007). Any violation of this confidentiality can therefore have serious consequences for the relation between a healthcare institution and a patient. A patient may be adverse to their data being used for research when non medical staff has access to private details, and might even consider seeking treatment elsewhere. Moreover, a potential data breach may expose private patient information to the general public. In the USA alone, between 2010 and 2013, 29,000,000 patient records were compromised (Liu et al., 2015). Clearly, such events have serious consequences for both the hospital and the patient.

From a legal perspective, on a European level the Directive 95/46/EG of the European Parliament on the protection of individuals with regard to the processing of personal data and on the free movement of such data was introduced in 1995 (European Data Protection Directive, 1995). All members of the European Union are obliged to take this directive into account, but how they implement it varies for each member. For example, in Sweden regional Ethics Committees give permission for the reuse of electronic patient records if the information that can identify a patient is removed (Velupillai et al., 2009). Similar measures are implemented in France in the French Data Protection Authority (Grouin et al., 2009). In the USA, the stricter Health Insurance Portability and Accountability Act (HIPAA) protects the privacy of healthcare data, requiring 18 different categories of identifying information ranging from person names to biometric identifiers such as fingerprints to be removed from medical data (HIPAA, 1996).

In the Netherlands no specific laws on the reuse of medical data exist, but there are general rules for dealing with personal data, that can be applied to medical data as well. Since EHR data is used for retrospective research, is not specifically collected for research purposes, and human subjects are only indirectly involved, the Medical Research Involving Human Subjects Act (WMO) does not need to be taken into account. Only the Agreement on Medical Treatment Act (WBGO) and the Personal Data Protection Act (WbP), which is the implementation of the European Directive 95/46/EG mentioned above, play a role in this situation. Retrospective research with medical records needs to be proposed to the medical ethics committee (METC), which verifies that the proposed research is in line with privacy legislation. An exception to this is when only anonymized data is used, which is the case if the de-identification process is executed perfectly.

For the two reasons above it is therefore important to de-identify as much of research data as possible, both to retain patient privacy and to be able to comply with legal requirements. For de-identification of personal data, a distinction can be made between directly identifying information and indirectly identifying information. Directly identifying information, such as names, phone numbers and citizen service numbers allow identification of a person with just that information. Indirectly identifying information, such as postal codes and birth dates, is not directly relatable to a person, but if pieces of indirectly identifying information are combined it is easy to identify someone (Borking and Raab, 2001). In medical text data both directly and indirectly identifying information can be present, and both types of information need to be removed to successfully de-identify medical text. Although manual de-identification is possible, it is time consuming and generally prone to error, while automatic de-identification is feasible and easily scalable to large numbers of records (Deleger et al., 2013). For this reason, we choose to develop an automatic de-identification method. Our goal is to remove as much de-identifying information as possible, while ensuring the de-identified text is still human readable, so that research can still be carried out. Even strict de-identification methods still retain good readability of the remaining text (Meystre et al., 2014). We therefore strive to balance towards developing a method with a high recall while also maintaining a good precision.

Since there are differences in legislation that exists on a national level, and because of language-specific problems that occur in the different the types of identifying information, it is clear that a separate de-identification method must be developed for each language (Grouin et al., 2009). Although many research into the de-identification problem has been performed in English, a reliable method for de-identifying Dutch medical text has yet to be developed. For the English language, most notably Neamatullah et al. (2008) obtained a recall of 0.967 using their pattern matching method that was developed on a test corpus of 1836 nursing notes. Uzuner et al. (2008) managed to achieve a 0.97 F_1-score on medical discharge summaries, based on a machine learning approach. A hybrid approach was developed by Ferrández et al. (2013), achieving a 0.922 recall by combining both pattern matching and machine learning techniques. Many more approaches in English exist (e.g. Friedlin et al., 2008; Douglass et al., 2005; Fenz et al., 2014).

Apart from text-processing the English language, Velupillai et al. (2009). attempted to port the Neamatullah et al. (2008) algorithm to Swedish, but in their own words with “poor results”. The average score over all de-identified categories was a 0.65 F_1-measure. Over a decade ago already, Ruch et al. (2000), successfully managed to de-identify discharge letters written in French with a recall of 0.98 using their MEDTAG framework for semantic tagging. For notes that are written in Korean and English, Shin et al. (2015) developed a method based on regular expressions that achieved a 0.963 recall on several categories combined. In Dutch, Scheurwegs et al. (2013) managed a recall of 0.89 on a previously unseen dataset with a machine learning approach, achieving reasonable success using limited training data.

As can be seen, the two most common methods to de-identify medical text are pattern matching based or machine learning based (Meystre et al., 2014). In the former, lookup tables and decision rules are used to determine what parts of the text contain identifying information. In the latter approach, machine learning techniques are used to classify each piece of text. A third option is a hybrid approach, combining the two. In literature, it is not immediately clear whether pattern matching or machine learning based methods perform better, although hybrid approaches generally outperform other approaches. A clear downside of the machine learning approach (and therefore also the hybrid approach), however, is the need for a large annotated training corpus, which requires