INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 83 (2014) 303–312

journal homepage: www.ijmijournal.com

Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records

Emmanuel Chazard[a,*], *Capucine Mouret*[b], *Grégoire Ficheur*[a], *Aurélien Schaffar*[a], *Jean-Baptiste Beuscart*[c], *Régis Beuscart*[a]

[a] *Department of Public Health, CHU Lille, UDSL EA 2694, Univ Lille Nord de France, F-59000 Lille, France*
[b] *Department of Occupational Medicine, CHU Lille, F-59000 Lille, France*
[c] *Department of Geriatrics, CHU Lille, UDSL EA 2694, Univ Lille Nord de France, F-59000 Lille, France*

ARTICLE INFO

Article history:
Received 26 March 2012
Received in revised form
28 November 2013
Accepted 28 November 2013

Keywords:
Anonymization
De-identification
Confidentiality
Free text
Natural language processing

ABSTRACT

Purpose: Medical free-text records enable to get rich information about the patients, but often need to be de-identified by removing the Protected Health Information (PHI), each time the identification of the patient is not mandatory. Pattern matching techniques require pre-defined dictionaries, and machine learning techniques require an extensive training set. Methods exist in French, but either bring weak results or are not freely available. The objective is to define and evaluate FASDIM, a Fast And Simple De-Identification Method for French medical free-text records.

Methods: FASDIM consists in removing all the words that are not present in the authorized word list, and in removing all the numbers except those that match a list of protection patterns. The corresponding lists are incremented in the course of the iterations of the method.

For the evaluation, the workload is estimated in the course of records de-identification. The efficiency of the de-identification is assessed by independent medical experts on 508 discharge letters that are randomly selected and de-identified by FASDIM. Finally, the letters are encoded after and before de-identification according to 3 terminologies (ATC, ICD10, CCAM) and the codes are compared.

Results: The construction of the list of authorized words is progressive: 12 h for the first 7000 letters, 16 additional hours for 20,000 additional letters. The Recall (proportion of removed Protected Health Information, PHI) is 98.1%, the Precision (proportion of PHI within the removed token) is 79.6% and the *F*-measure (harmonic mean) is 87.9%. In average 30.6 terminology codes are encoded per letter, and 99.02% of those codes are preserved despite the de-identification.

Conclusion: FASDIM gets good results in French and is freely available. It is easy to implement and does not require any predefined dictionary.

* *Corresponding author at*: Pôle de santé publique, 154 rue Yersin, CHRU de Lille, 59037 Lille Cedex, France. Tel.: +33 3 20 44 60 38.
E-mail addresses: emmanuel.chazard@univ-lille2.fr, emmanuelchazard@yahoo.fr (E. Chazard).
1386-5056/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijmedinf.2013.11.005

1. Introduction

1.1. A need for de-identifying discharge letters

Computerized free-text medical records are important information sources for research. In most countries, each time a patient is discharged from a healthcare facility, a discharge letter has to be written: it summarizes all the pertinent information from the reason for admission to the discharge drug treatment. Those letters are routinely produced and provide the researchers with a big amount of medical information. On the other hand, the confidentiality must imperatively be respected: as soon as a discharge letter is not used with direct benefit to the patient and if the patient does not need to be identified, the letter must be de-identified. The anonymization consists in removing the patients' names from the records: unfortunately, other pieces of information enable to identify the patients. The de-identification is a more exhaustive removal of the entire Protected Health Information (PHI), so that the patients cannot be identified, directly nor indirectly. In the US, privacy rules have been enacted by the Department of Health and Human Services further to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [1]. In order to de-identify a high number of records, it is necessary to use automated methods, as manual methods require too high workload [2].

1.2. State of the art

Several methods exist for automated de-identification of free-text records [3], including procedures reports and discharge letters.

Pattern matching methods [4–16] consist in applying rules that enable to keep or remove some words that belong to dictionaries that have been predefined by experts or institutions. For instance, it is possible to remove all the words that belong to a list of town names, or to preserve all the words that belong to a list of medical terms (such as the Unified Medical Language System [17]). Additional rules may be used to take into account words declension and verbs conjugation. This approach requires that such lists are available. When they exist, those lists are language-dependent, and are suitable for a specific context only (e.g. town names or current family names are useless in another country).

Machine learning methods [14,18–26] are derived from artificial intelligence. A learning phase requires that a corpus of records is previously de-identified manually by experts. Those methods are often very efficient, depending on the quality and the completeness of the learning corpus.

Whatever the method used, the de-identification is evaluated by computing three rates:

- The recall (or sensitivity or completeness, Eq. (1)), which is the proportion of removed token within the PHI. A high recall enables to preserve the confidentiality.
- The precision (or positive predictive value or correctness, Eq. (2)), which is the proportion of PHI within the removed token. A high precision enables to preserve the readability of the text.
- The *F*-Measure, which is the harmonic mean of the recall and the precision (Eq. (3)).

$$\text{recall} = R = \frac{TP}{\#\text{identifiers}} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{precision} = P = \frac{TP}{\#\text{removed}} = \frac{TP}{TP + FP} \tag{2}$$

$$F\text{-measure} = F = \left(\frac{R^{-1} + P^{-1}}{2}\right)^{-1} \tag{3}$$

Table 1 presents the main results obtained in the literature by the authors for medical free-text de-identification. Most of methods are developed for English language and can hardly be used for other languages. Some methods have been developed in French, but either their results are disappointing, or they are not freely available.

1.3. Unsolved situations

Despite the good results obtained by many methods, text de-identification is still not obvious and some situations may not be addressed with current tools. We shall illustrate it through 4 use cases.

Case 1: a team has to de-identify English free-text records using pattern-matching. Some tools are freely available. However, it cannot be guaranteed that those tools could be applied in a different context without any adaptation. Indeed, pattern matching techniques rely on lists of words that are context-dependent: for instance "lime tree" should be removed in most reports as it is often part of a street name, but should not be removed in an allergy-related report. Lists of town names or family names also depend on the country. Finally, misspellings are most often not taken into account by existing methods.

Case 2: a team has to de-identify English free-text records using machine learning. Here again, some tools are freely available but, in a like manner, machine learning techniques require a pre-existing corpus of de-identified records. Such corpuses are available in English [11,36,37], but they may be used only if the type of document to de-identify is the same as the documents of the training corpus.

Case 3: a team has to de-identify French free-text records (the problem is the same with most of non-English languages): no free and efficient method, no list of words, and no training corpus are available. Everything has to be built.

Case 4: a team has only little time (e.g. 1 man-week) to de-identify a few records (e.g. 25,000 records). Whatever the language, the context and the technique, it will probably take more time to understand, adapt, implement and execute an existing tool.

The conception of FASDIM relies on the idea that a simple de-identification technique could enable to de-identify French discharge letters with an acceptable workload, particularly when the number of records is low. The main idea is to supply the workload in the course of the method, and not before the first document can be de-identified.