Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

De-identification of clinical notes in French: towards a protocol for reference corpus development

Cyril Grouin *, Aurélie Névéol

LIMSI–CNRS, UPR 3251, Orsay, France

ARTICLE INFO

Article history:
Received 2 August 2013
Accepted 22 December 2013
Available online xxxx

Keywords:
Confidentiality
Electronic Health Records
France
Information Dissemination
Natural Language Processing

ABSTRACT

Background: To facilitate research applying Natural Language Processing to clinical documents, tools and resources are needed for the automatic de-identification of Electronic Health Records.
Objective: This study investigates methods for developing a high-quality reference corpus for the de-identification of clinical documents in French.
Methods: A corpus comprising a variety of clinical document types covering several medical specialties was pre-processed with two automatic de-identification systems from the MEDINA suite of tools: a rule-based system and a system using Conditional Random Fields (CRF). The pre-annotated documents were revised by two human annotators trained to mark ten categories of Protected Health Information (PHI). The human annotators worked independently and were blind to the system that produced the pre-annotations they were revising. The best pre-annotation system was applied to another random selection of 100 documents. After revision by one annotator, this set was used to train a statistical de-identification system.
Results: Two gold standard sets of 100 documents were created based on the consensus of two human revisions of the automatic pre-annotations. The annotation experiment showed that (i) automatic pre-annotation obtained with the rule-based system performed better ($F = 0.813$) than the CRF system ($F = 0.519$), (ii) the human annotators spent more time revising the pre-annotations obtained with the rule-based system (from 102 to 160 minutes for 50 documents), compared to the CRF system (from 93 to 142 minutes for 50 documents), (iii) the quality of human annotation is higher when pre-annotations are obtained with the rule-based system (*F*-measure ranging from 0.970 to 0.987), compared to the CRF system (*F*-measure ranging from 0.914 to 0.981). Finally, only 20 documents from the training set were needed for the statistical system to outperform the pre-annotation systems that were trained on corpora from a medical speciality and hospital different from those in the reference corpus developed herein.
Conclusion: We find that better pre-annotations increase the quality of the reference corpus but require more revision time. A statistical de-identification method outperforms our rule-based system when as little as 20 custom training documents are available.

1. Introduction

Medical knowledge is routinely advanced through clinical studies involving patient volunteers who provide informed consent to participate in a carefully designed study, planned before any medical information is collected or any health care procedure is performed. Medical knowledge can also be greatly advanced through retrospective studies exploiting the wealth of information contained in Electronic Health Records (EHRs). This type of study also requires the patients involved to provide informed consent. However, because the study design is crafted after the patients have received the care described in the EHRs, it can be difficult to obtain consent from each patient (e.g., logistics issues arise for contacting the patients or their family).

De-identification is the process of hiding or removing content that explicitly identifies persons involved in patient care, including patients themselves and health care providers [1]. The use of de-identified clinical data provides researchers with the means to carry out studies that can advance the state of medical knowledge while protecting patients' privacy and confidentiality. Specifically, in the absence of informed consent, the Personally Identifiable Information (PII) and Protected Health Information (PHI) contained in clinical data must be processed according to privacy rules and regulations.

A significant body of research has addressed the issue of de-identification in the past decades, covering different types of data,

* Corresponding author.
E-mail addresses: cyril.grouin@limsi.fr (C. Grouin), aurelie.neveol@limsi.fr (A. Névéol).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.014

such as text, images, biological samples and DNA sequences [2]. In this paper, we focus on the de-identification of clinical free-text, as a preliminary step to prepare clinical text for further Natural Language Processing (NLP) and analysis of clinical documents. In order to ensure the quality and robustness of NLP tools, real clinical data must be used for development and testing.

An increasing number of efforts recently targeted the de-identification of clinical text in English [3]. Other efforts also addressed the de-identification of clinical documents in languages other than English such as French [4,5] and Swedish [6,7]. The lack of a freely available de-identification reference corpus similar to the i2b2 corpus available for English [8] has prevented any rigorous comparison between the two approaches developed for French.

Our goal is to support research in Natural Language Processing for biomedical texts in French through the development of a de-identified corpus that can be distributed to the scientific community for research purposes [9].

In this paper, we focus on three specific aims that address both fundamental research questions and practical considerations:

1. De-identification research methods for clinical texts in French: what are the best methods for automatic text de-identification in French? What are the best methods for producing a reliable, high-quality reference corpus for de-identification? Specifically, we assess the usability of two automatic pre-annotation methods.
2. De-identification resources: development of a de-identification reference corpus freely available to the scientific community.
3. De-identification evaluation: assess the time and effort required to produce de-identified corpora and adapt existing de-identification tools to new, unseen data.

2. Related work

2.1. De-identification of clinical free-text

De-identification of clinical data, including de-identification of clinical free-text in English has been well-studied in the past decade. De-identification is generally approached as a specific named entity recognition task targetting PHIs. Named entity recognition is defined by Meystre et al. [1] as "the task of recognizing expressions denoting entities (i.e., named entities), such as diseases, drugs, or people's names, in free text documents". A review of available tools shows that de-identification can be reasonably achieved using a rule-based approach, statistical machine learning, or a combination of both [3]. The rule-based tool developed by Neamatullah et al. [10] was notably used to de-identify clinical documents in the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database [11,12] and adapted to data outside of the United States [13]. It uses the principle of surrogate PHI re-introduction, which consists of substituting PHI in the original records by similar made-up data in order to preserve language coherence while enforcing privacy. This process was shown to have minimal impact on information extraction in clinical documents [14].

Recent work used the de-identification reference corpus developed for the i2b2 2006 challenge [8] to perform a systematic evaluation of five de-identification systems available for English [15], which prompted the development of a new tool customized for VA documents [16]. Contrary to what was reported by Wellner et al. [17] in their work for the i2b2 challenge, these studies showed that a fair amount of adaptation is required for any de-identification tool to obtain acceptable results on new, unseen corpus. The study also provided valuable insight for de-identification tool adaptation by pointing out that the strength and weaknesses of rule-based and statistical systems seem to be different for the types of PHI targeted.

While de-identification as a task seems to be almost resolved, efficient adaptation of de-identification tools to new corpora (including in languages other than English) currently remains a major challenge. Recent work provided estimations of the annotator time [18,19] required to prepare training data for a statistical de-identification tool achieving 0.96 *F*-measure on clinical notes in English. The same group also assessed the effect of training corpus size [19], training document type [18] and re-identification status [20].

2.2. Development of annotated reference corpora

Many international NLP challenges require annotated reference corpora for participant evaluation. For instance, the Message Understanding Conferences (MUC) [21] notably produced reference corpora for named entities such as organization names, person names, locations, dates and times in newswire text. In the biomedical domain, challenges yielded reference corpus for named entities including bacteria [22], genes [23], medications [24], diseases [25] or PHI [8]. Throughout the tasks, a variety of document genres were covered, including scientific articles or abstracts [22,23,25], clinical text [8,24], PubMed queries [26]. Several methods have been used and assessed for producing reference annotations for these tasks, relying on human annotations for all [8,24,25] or part [23] of the final reference. Automatic pre-annotations have often been used to process corpora in the biomedical domain in order to create quality reference corpora [8,26,27]. This process was shown to be relevant and useful as it saves annotation time, contributes to the production of consistent, high quality annotations and is overall preferred by annotators [26]. Another commonly used method for producing high-quality reference corpora is the use of several annotators working either independently [25,27] or in sequence [8] and discussing disagreements to produce consensus annotations.

3. Material and methods

3.1. Corpus

The corpus we used was approved by the French administrative authority on data privacy[1] for research on Information Retrieval (IR) in large Electronic Health Records [28]. To address the IR task in the context of severe diseases (i.e., records containing a large number of documents on a given patient) 1000 patient records were selected randomly from patients with at least 50 hospital stays in a group of hospitals within a French geographic area. The entire corpus comprises about 170,000 documents. As a result, a large variety of medical specialties (e.g., Pneumology, Obstetrics, Infectious Diseases), clinical document types (e.g., radiography reports, discharge summaries, consult correspondance) and hospitals (5 locations) are covered in the corpus. In the random subsets of documents used in this study, we did not attempt to control the distribution of either specialties, document types, or original health care provider. The sheer number of document types and specialties represented in the overall corpus would make this a difficult task. While it has been shown that tools perform better if trained on documents very similar to those they are tested on [18], we are interested in assessing the portability of de-identification tools with minimal adaptation work.

The corpus was de-identified by the original health care providers based on patient information as it appeared in the local hospital information system: patients' first and last names were replaced by the string "XX" while the day and month in their dates of birth

[1] Commission Nationale de l'Informatique et des Libertés (CNIL).

Please cite this article in press as: Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2013.12.014