How to de-identify a large clinical corpus in 10 days

C. Grouin, PhD[1], L. Deléger, PhD[1], J.B. Escudié, MD[2], G. Groisy, MD[2], A.S. Jannot, MD, PhD[2,3], B. Rance, PhD[2,3], X. Tannier, PhD[1,4], A. Névéol, PhD[1]

[1]CNRS UPR 3251 LIMSI, Orsay, France; [2]AP-HP, University Hospital HEGP, Biomedical Informatics and Public Health Department, Paris, France; [3]INSERM U1138, Université Paris Descartes, Sorbonne Paris Cité, Faculté de médecine, Paris; [4]Université Paris Sud, Orsay, France

Abstract

To de-identify a large corpus of clinical documents in French supplied by two different health care institutions, we apply a protocol built from previous work. We show that the protocol can be installed and executed by outside collaborators with little guidance from the authors. The automatic de-identification method used reaches 0.94 F-measure and human validation requires about 1 minute per document.

Introduction

Clinical corpora are useful for scientists to develop natural language processing methods for the clinical narrative. In order to ensure the robustness of those methods, access to real data is a critical point. As clinical records contain personal health information, de-identification tools have been designed to help protecting privacy.[1]

Methods

Herein, we apply the protocol we designed to rapidly de-identify a new set of clinical records:[2] new documents to be de-identified are automatically pre-annotated with a statistical model built on similar resources. For new documents from the same hospital as our previous study[2] (corpus 1), we directly apply a model built with 100 training documents. For documents coming from a distinct hospital (corpus 2) we iteratively build a model as follows:

1. 20 documents were de-identified automatically, using a rule-based method[3] for the first iteration and a CRF (Conditional Random Fields) model for subsequent iterations, and manually validated by scientists and physicians

2. a statistical (CRF) model was built using all the manually validated de-identified documents available from all iterations. The latest model is then used for the next iteration of step 1. The features used to build the CRF model include surface features *(case, punctuation, digit, token length, etc.)*, deep features *(part-of-speech, lexical look-up, etc.)*, and external features *(position of the token in the record, cluster id).*

For the initial seed corpus of 20 documents, six different types of clinical notes were chosen to ensure robustness: 3 medical certificates, 3 consultation reports, 3 exam certificates, 3 hospitalization reports, 5 follow-up care letters, and 3 staff consultation reports.

Results and Conclusion

Corpus 1: The automatic de-identification method reached 0.94 F-measure and human validation required less than 1 minute per document for 800 documents. Corpus 2: About 1 day was needed to install the tools to use in the de-identification protocol. The automatic de-identification method reached 0.92 F-measure and human validation required as little as 1 minute per document for 100 documents. Validation time decreased with annotator experience and improvement in the performance of the de-identification method (i.e. after more training documents become available as we progress through more iterations). The protocol can be executed by outside collaborators with no prior experience of de-identification. It provides adequate support for the de-identification of a large corpus requiring little time and guidance.

References

[1]Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10.

[2]Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform. 2014. In press.

[3]Grouin C, Zweigenbaum P. Automatic de-identification of French clinical records: comparison of rule-based and machine-learning approaches. Stud Health Technol Inform. 2013;192(Part 1):476–80.