Swiss Medical Informatics

Original article | Published 01 November 2018 | doi:10.4414/smi.34.00417
Cite this as: Swiss Med Informatics. 2018;34:00417

De-identification of French medical narratives

Gaudet-Blavignac Christophe[a], Foufi Vasiliki[a], Wehrli Eric[b], Lovis Christian[a]

[a] Division of Medical Information Sciences, Geneva University Hospitals and University of Geneva, Switzerland
[b] Laboratoire d'Analyse et de Technologie du Langage, University of Geneva, Switzerland

Summary

RESEARCH QUESTION: Maintaining data security and privacy in an era of cybersecurity is a challenge. The enormous and rapidly growing amount of health-related data available today raises numerous questions not only about data collection, storage, analysis, comparability and interoperability but also about data protection. The US Health Information Portability and Accountability Act (HIPAA) of 1996 provides a legal framework and guidance for using and disclosing health data. The approach proposed by HIPAA is the de-identification of medical documents by removing certain protected health information (PHI).

METHODS: In this work, a rule-based method for the de-identification of French free-text medical data using natural language processing (NLP) tools is presented.

RESULTS: A random sample of 1000 discharge summaries were manually evaluated. The system achieved 0.93 total recall and 0.99 precision.

DISCUSSION: The evaluation results showed very good performance of the system. However, spelling, orthographic and typographic errors that are present in discharge summaries can affect the de-identification process.

Keywords: *medical data, data protection, privacy, HIPAA, Natural Language Processing (NLP), de-identification, anonymisation*

Introduction

Medical data contain various types of personally identifiable information (PII) or otherwise sensitive personal information (SPI). In this context, legislation has been defined to ensure personal data protection. The most significant legal document produced to face the challenge of healthcare data management is the US Health Insurance Portability and Accountability Act (HIPAA) of 1996 and its revisions. In Europe, the General Data Protection Regulations (GDPR) have been approved (April 2016) and implemented (May 2018). These texts provide a legal framework and guidance for using and disclosing personal data. The approach proposed by HIPAA is the de-identification of medical documents by removing certain protected health information (PHI).

This paper deals with the de-identification of French free-text medical data for secondary usage (medical research, quality measurement and improvement, public health, epidemiology and other purposes). Since it has been proved that manual de-identification of medical records is time-consuming [1], automating the work with the use of natural language processing (NLP) tools to perform this task is mandatory. In particular, pre-processing tools, electronic dictionaries and finite state automata constructed in the Unitex corpus processing system will be applied to the medical narrative data.

De-identification is generally approached as a specific named entity recognition (NER) task targeting PHI. NER is defined as "the task of recognizing expressions denoting entities, such as diseases, drugs, people's names in free text documents" [2].

The great need for de-identification techniques is reflected by the large number of systems that have been built over the last 20 years. Some of them are rule-based systems, whereas others, such as MIST [3], use conditional random field (CRF) models trained for text processing. Systems such as BoB of the Veteran's Health Administration [4] and the Cincinnati Children's Hospital Medical Center's (CCHMC) in-house de-identification system [5] follow hybrid approaches. Other de-identification systems are the Scrub system [6], Datafly [7], the MIMIC de-identification filter [1, 8, 9], HIDE [10] and deid [11]. Most of the tools are available for the English language, but a rule-based de-identification system for Serbian medical narrative texts was also built [12]. Moreover, a system for removing identifiers in French medical records, with a success rate of about 99%, has been designed [13]. Finally, some multilingual systems have been constructed [14, 15].

The system presented in this paper has two major characteristics: on one hand, it is designed for French narrative data based on a symbolic NLP method and on the other hand it preserves the data integrity since PHI is not removed but replaced with credible structures.

Method

Following HIPAA, 18 categories of information, including names, geographic locations, elements of dates, social security numbers, telephone and fax numbers, must be removed from medical texts. In the framework of the 2014 i2b2/UTHealth Natural Language Processing (NLP) shared task3, where one of the tracks focused on identifying PHI in longitudinal clinical narratives, new categories

Correspondence:
Vasiliki Foufi, Division of Medical Information Sciences , University Hospitals of Geneva (HUG), University of Geneva (UNIGE), Campus Biotech G6, Vasiliki.Foufi[at]unige.ch

such as hospital, room, department and identities of devices, vehicles and biometrics were added. By removing only a given number of identifiers, de-identification preserves data integrity.

De-identification is viewed as a NER task targeting PHI (person names, dates, geographical locations, contact information). To perform this task, pre-processing tools (tokenisation, sentence splitting, part-of-speech tagging), lexicons of simple and compound words, and rules with orthographic (capitalisation, punctuation), pattern, negation, lexical and context features, symbols and special characters are applied to medical texts. The grammars that have been constructed recall data from the electronic dictionaries of simple and compound words incorporated in Unitex and produce some output based on the notion of transduction. Furthermore, the use of right and left context – either positive or negative – contributes to the identification of PHI. For instance, the presence of the determinant *de, de la* in certain proper names should be predicted in the finite state automata (positive left context). All identified information is replaced by credible surrogate structures and not by generic strings. For instance, dates contained in the documents are replaced by surrogate ones consistent with the various types of dates found in the text. Some representative examples of dates are cited below:

– *le 06 janvier 2018* (on 6th January 2018)
– *en novembre 2018* (in January 2018)

After being identified, days and months are replaced but years are kept in their initial form, following the HIPAA rules. More precisely, the patterns *le 06 janvier 2018* (on 6th January 2018) and *en novembre 2018* (in November 2018) are transformed to *le 30 février 2018* (on 30th February 2018) and *en février 2018* (in February 2018), respectively.

For the detection of names, trigger words have been used. In particular, titles such as *Monsieur* (Mr), *Madame* (Mrs), *Professeur* (Professor), *Docteur* (Doctor) and others are considered as triggers for person named entities (NEs). Like dates, patients' names also present various structures:

– *Title (Mr or Mrs) + first name + last name* (small or capital letters)
– *Title (Mr or Mrs) + last name* (formed by two or more constituents with or without dash in small or capital letters)
– *Title (Mr or Mrs) X'X* (apostrophe between the constituents of the name).

Likewise, in doctors' names, the title (*doctor, Dr, professor,* etc.) could precede the name, followed or not by the specialisation (*general practitioner, oncologist, cardiologist,* etc.).

On the other hand, "anatomic locations, devices, diseases and procedures could be erroneously recognised as PHI and removed" [12]. During the processing of the discharge summaries, similar remarks have been made. In the terms *classification de Los Angeles* (Los Angeles classification system), *score de Lille* (Lille model), and *maladie de Parkinson* (Parkinson's disease), the proper noun should not be de-identified. Diseases, syndromes, classifications and scores containing a proper name are detected by the system and excluded by the de-identification process.

Results

The finite state automata have been applied to a corpus of 11,000 discharge summaries in French. An example of a de-identified sentence is given below:

Initial sentence: ***Monsieur Gaudet-Blavignac*** *a été transféré* ***aux Hôpitaux Universitaires de Genève le 5 novembre 2018.*** (**Mr *Gaudet-Blavignac*** was transferred to the **University Hospitals of Geneva on 5th November 2018.**)

De-identified sentence: ***Monsieur Foufi*** *a été transféré* ***à l'Hôpital le 30 février 2018***. (**Mr Foufi** was transferred to the **Hospital on 30th February 2018**.)

Next, a random sample of 1000 discharge summaries were manually evaluated. The system achieved 0.93 total recall and 0.99 total precision. Table 1 displays the results per category and the total performance of the system:

Further evaluation taking into account all categories showed that 71.8% of the corpus of 1000 was totally de-identified. However, as shown in table 1, the categories where the system performs less well are dates and locations. If these two categories are not taken into account, it leads to a higher performance (95.2% of the corpus totally de-identified).

Discussion

This approach allows a fine-grained development of high precision and high recall text recognition patterns which in turn guarantees the preservation of data quality. Finite state automata have some important characteristics. They can be easily and effectively explained and understood by future users and by the general public without the need of a specific background. In addition, rules can be deleted, changed and enriched after testing. This means that once applied to medical documents, an error can be located and corrected with the aim to improve the de-identification results.

On the other hand, this method is language dependent, which means that it depends on the complexity and particularities of each language. Consequently, finite state automata describing the various linguistic features occurring in text should be constructed for each target language. The rules may also differ between various types of documents as the language used may also be different. Additionally, manual validation of the results is required to identify possible errors and ellipsis in the rules and optimise them.

Table 1: Results of the evaluation of the system on 1000 discharge summaries.*

	Dates	Patient names	Physician names	Locations	Overall performance
Precision	0.9889	0.9970	1	0.9628	**0.9907**
Recall	0.9228	0.9916	0.9876	0.7872	**0.9342**

* Not all of the categories appeared in the evaluation corpus.

Finally, spelling errors can affect the de-identification process. The fact that discharge summaries are often written in a hurry and contain as a consequence spelling, orthographic and typographic errors has already been pointed out (among others [12, 16]). Dates such as *en 20**0**11* (in 20**0**11) and *du 27.07.au 01.08.2014* (from 27.07.to 01.08.2014) are difficult to detect automatically. Moreover, spelling mistakes in the trigger words (e.g., *Monseur, Monsier, Monsiuer,* instead of *Monsieur*) can prevent the system from recognising the named entities.

Conclusion

In this paper, a rule-based method for the automatic de-identification of French clinical narrative data has been presented. Finite state automata constructed via the Unitex corpus processing system have been applied to a corpus of 11,000 discharge summaries. The evaluation results show that the system is capable of detecting and de-identifying PII in texts. The corpus de-identified using this method could then be used for further research.

Acknowledgements
This research has been financed by the Swiss Personalised Health Network (SPHN). We would also like to thank Dr Christophe Fehlmann for providing us the corpus of discharge summaries.

Disclosure statement
No financial support and no other potential conflict of interest relevant to this article was reported.

References

1 Douglass M, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-Assisted Deidentification of Free Text in the MIMIC II Database. Comput Cardiol. 2004;31:341–4. doi: http://dx.doi.org/10.1109/CIC.2004.1442942.
2 Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128–44.. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.5779&rep=rep1&type=pdf PubMed.
3 Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform. 2015;58(Suppl):S20–9. doi: http://dx.doi.org/10.1016/j.jbi.2015.07.020. PubMed.
4 Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc. 2013;20(1):77–83. doi: http://dx.doi.org/10.1136/amiajnl-2012-001020. PubMed.
5 Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc. 2013;20(1):84–94. doi: http://dx.doi.org/10.1136/amiajnl-2012-001012. PubMed.
6 Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp. 1996:333–7. . Available at: https://dataprivacylab.org/projects/scrub/paper1.pdf PubMed.
7 Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. Proc AMIA Annu Fall Symp. 1997:51–5.. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233452/pdf/procamiaafs00001-0090.pdf PubMed.
8 Levine JM. De-identification of ICU Patient Records. Massachusetts Institute of Technology; 2003.
9 Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8(1):32. doi: http://dx.doi.org/10.1186/1472-6947-8-32. PubMed.
10 Gardner J, Xiong L, Kanwei L, Lu JJ. HIDE: heterogeneous information Deidentification. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09); 2009 Mar 24-26; Saint Petersburg, Russia. 2009.
11 Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):E215–20. doi: http://dx.doi.org/10.1161/01.CIR.101.23.e215. PubMed.
12 Jaćimović J, Krstev C, Jelovac D. A Rule-Based System for Automatic De-identification of Medical Narrative Texts. Informatica. 2015;39:45–53.
13 Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp. 2000:729–33.. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244050/pdf/procamiasymp00003-0764.pdf PubMed.
14 Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A De-identification method for bilingual clinical texts of various note types. J Korean Med Sci. 2015;30(1):7–15. doi: http://dx.doi.org/10.3346/jkms.2015.30.1.7. PubMed.
15 Dias FMC. Multilingual Automated Text Anonymization [thesis]. Lisbon: Instituto Superior Técnico of Lisboa; 2016.
16 Thompson P, McNaught J, Ananiadou S. Customised OCR Correction for Historical Medical Text. Digital Heritage. 2015:35–41. doi: http://dx.doi.org/10.1109/DigitalHeritage.2015.7413829.