Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 243-246 recibido 26-06-10 revisado 03-07-10 aceptado 06-07-10

Complementary Methods for De-identifying Sensitive Data with a focus on Clinical Discourse

Métodos complementarios para la anonimización de datos en el dominio clínico

Dimitrios Kokkinakis
Språkbanken, Department of Swedish Language,
University of Gothenburg, Box 200, SE-405 30, Sweden
dimitrios.kokkinakis@svenska.gu.se

Resumen: En la era de las notas clínicas en formato electrónico, las disponibilidad de datos personales para investigación, planificación y estadísticas sobre salud y seguimiento de enfermedades son algunas de las áreas en las cuales la protección de la información de los pacientes se ha convertido en un importante asunto. El objetivo de este estudio es adaptar y aplicar métodos para la anonimización de documentos, en particular en el dominio clínico. El principal reto y objetivo de esta investigación es mantener importantes conceptos en los documentos en una manera estandar y neutral que significa la encriptación sin violar la integridad de los datos personales y sin sacrificar la calidad y el significado previsto por los autores.

Palabras clave: Anonimización, Confidencialidad, Historia Médica

Abstract: In the era of the Electronic Health Record (EHR) the release of individual data for research, public health planning, health care statistics, monitoring of diagnostic tests, automated data collection for health care registries and tracking disease outbreaks are some of the areas in which the protection of Personal Health Information (PHI) has become an important concern. The purpose of this study is to adapt and apply synergetic methods to document de-identification, particularly clinical, or other sources of sensitive data. The main challenge and goal of this research is to retain important concepts and PHI in the documents in a standardized and neutral manner as means of encryption without violating the integrity of the PHI and without sacrificing the quality and intended meaning of the authors.

Keywords: Information dissemination, confidentiality, personal health information, clinical/medical data, terminology recognition, data scrubbing.

1 *Introduction*

De-identified data can be used as a source of information and knowledge to broad spectrum of services related to the growing demands for better forms of dissemination of confidential information about individuals in EHRs and other clinical free text (i.e. phenotype information, human biological specimens). On a daily basis, hospitals produce, manage and store vast amounts of patient-related data. Due to confidentiality requirements these data – mostly in textual form – remain inaccessible for research and knowledge mining. Thus, although there is a growing need for accessing EHR data for clinical and language technology research, these are underutilized or not utilized at (Pestian et al., 2006). The need of disclosure of PHI in EHRs for secondary purposes, i.e. retrospective PHI uses outside of direct health care delivery[1] (Safran et al., 2007) is expected to increase dramatically the coming years and research in the area of de-identification has attracted the attention of many research groups worldwide working actively for sustainable solutions. For instance, in the *Challenges in NLP for Clinical Data* workshop (Uzuner et al., 2006) there are details of systems participated in a shared task of automatic de-identification of medical summaries.

[1] This implies analysis, research, quality and safety measurement, public health, provider certification, marketing, and other applications.

ISSN 1135-5948

2 Background

Access of clinical text for research purposes (e.g. EHR) which can ensure protection of PHI, can be either granted by the patients themselves, by obtaining permission from institutional review boards, or by data use agreements under which e.g. researchers must obtain approvals for use of the data by regional ethic committees. In any case de-identification of various explicit identifiers (such as names of relatives or doctors' names) is often required. De-identification is defined as the process of recognizing and deliberately changing, replacing or concealing the names and/or other identifying information of relevance about entities (PHI) from clinical or other sensitive to disclosure data. *Data scrubbing* is another term used for the same purpose (cf. Sweeney, 1996) sometimes with a bit lower understandability ambitions (Berman, 2003). In our context we do not make a differentiation between these two terms. There are many de-identification techniques described in the literature. One of the earliest systems is the "Scrub" system (Sweeney, 1996) which was based on a set of detection algorithms utilizing word lists and templates that each detected a small number of name types in pediatric records. The *k-anonymisation* approach, described in Sweeney (2002), de-associates attributes from the corresponding identifiers, each value of an attribute, such as date of birth, is suppressed or generalized. For a description of several methods for making data anonymous, see Hsinchun et al. (2005); El Emam & Fineberg (2009) also provide a description of four techniques, namely randomization/masking, pseudonymization, heuristics and analytics. Finally, a thorough review on de-identification in the clinical domain is given in Meystre et al. (2008); while one of the first publicly available de-identification software and relevant test data are described in Neamatullah et al. (2008). For Swedish, which is our application language, the works by Kokkinakis & Thurin (2007) and Velupillai et al. (2009) are relevant on the topic.

2.1 HIPAA and PHI

In different parts of the world confidentiality is regulated and protected by various mechanisms such as the *US Health Insurance Portability and Accountability Act* (HIPAA) (2003) or the *European Commission's Directive on Data Protection* (95/46/EC). Such policies state that for a text to be rendered as safely de-identified, information such as e.g. *names*, *geographic subdivisions*, *dates,* etc. must be removed. In many cases such information has been modified to suit different needs. HIPAA defines 18 different data elements that should be replaced from any type of sensitive data in order for them to be considered de-identified. Researchers such as Hrynaszkiewicz et al. (2010) discuss a 28 item list of patient identifiers in datasets some of which are complementary to the previous (e.g. *biometric data*). Velupillai et al. (2009) discuss that e.g. *ethnicity* might also be another identifier that can reveal crucial to re-identification identifiable information. To these identifier lists *rare disease names*, certain forms of *ethnic clothing*, *offending words* or *personal attributes* such as exceptional qualities (e.g. "Olympic medallist") can be added.

3 Materials and Methods

Ethical issues might be a barrier to directly accessing patient data, without approval from ethical committees. There is however other means that can circumvent this barrier, for instance by looking at similar data with fewer restrictions on their content. In order be able to rapidly test technologies in realistic scenarios and avoid ethical and administrative problems we explore texts given to medical students in the form of written examination papers. These texts mirror the reality the students are suppose to meet as they start their professional career (*fake* discharge summaries and case reports, where the students have to read and comprehend in order to evaluate the clinical problems and solutions proposed therein; these reports are considered equivalent to *real* reports). In our study we have assembled such a corpus of 52 EHR-like reports (150,000 tokens) from medical faculties across Sweden, e.g. <http://courses.ki.se/ utbildningsprogram/Lakare>.

The following methodology has been applied to the Swedish corpus, originally inspired by the work of Berman (2003) and which is based on a number of complementary, techniques described below:

1. (Generic) *tokenization* and generic *multi word expression identification* (e.g. prepositions, adverbs as well as idioms taken from monolingual general lexica).
2. *Terminology recognition* using the Swedish and English Medical Subjects

Headings (MeSH) and parts of the Swedish Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT). Both resources have been extended with numerous variant forms (not part of the original listings) and linked to the original (Kokkinakis, 2009; Kokkinakis & Gerdin, 2009). This way during processing we can safely replace variant terms with their recommended term or code found in the reference terminologies, minimizing risks for confidentiality attacks – e.g. no matching to the original report and the output can be made using idiosyncratic spelling.

3. Applying generic *named entity recognition* (NER) to the texts and filtering out the content in a way that only the labels of the entities are retained. This way sensitive text content is filtered out during processing. The NER system recognises eight main classes of entities: PERSON, ORGANIZATION, LOCATION, TIME, MEASURE, EVENT, ARTIFACT and WORK; cf. Kokkinakis (2004).
4. Optionally, it is possible to *introduce bias* by changing person names to the most frequent Swedish male/female first/last names and also health care professionals to frequent names with suitable attributes such as ‘dr’ (i.e. doctor) or ‘ssk’ (i.e. nurse).
5. *Data scrubbing*: using a large general text corpus (>50 million tokens) we extracted two lists of the 5000 and 10000 most frequent tokens. During processing if a token in the test texts is among the most frequent tokens then it is kept in tact. Also, punctuation markers, numbers of length <3 and tokens consisting of 1 character are also kept intact in the texts. All other tokens, not part of the previous steps, are scrubbed according to their length and orthographic characteristics. This implies that each orthographic character is changed to an asterisk ‘*’, and each number (not part of the NER) to ‘N’.

The following example illustrates some of the steps in the approach: ‘*Du arbetar på akuten i Lund och får in en 55-årig man med Hodgkinlymfom som heter Karjalainen och kommer från Karelen i Finland*’ i.e. ‘You are working in the emergency room in Lund and receive a 55-year-old man with Hodgkin Disease named Karjalainen who comes from Karelen in Finland’. After processing, the text (with the 5000 general language threshold) becomes: ’*Du arbetar på ****** i LOCATION/CITY och får in PERSON/MALE med C04.557.386.355; C15.604. 515.569.355; C20.683.515.761.355 som heter PERSON/ MALE och kommer från LOCATION/CITY i LOCATION/COUNTRY*’.

4 Evaluation Issues and Conclusions

The main goals of this work were to retain important concepts and PHI in the documents in a standardized and neutral manner without violating the integrity of the PHI and without sacrificing the quality and intended meaning of the authors. In an evaluation in a small scale conducted (by the author) the results showed that roughly all sensitive PHI in these texts have been either replaced by neutral labels by the NER process or scrubbed rendering the textual data harmless (the evaluation material can be found here: <http://demo.spraakdata.gu.se/svedk/pbl/scrubbCorpusText.txt>). The results seem adequate for fulfilling the first goal and we have implemented an interface that can be used for testing the validity of these results.

With respect to our second goal, that of information preservation (‘intended meaning’) the scrubbing approach is sensitive to the amount of the general language that can be retained in the original texts since in many cases there is a risk that the meaning of a sentence is changed or lost. The 10000 word limit seems an acceptable threshold for this purpose (the top-5000 tokens are in many cases limited), however a higher threshold might be more suitable for information preservation. Unsafe terms in the threshold lists, e.g. *suicide* and *rape* can be pruned and forbidden words can be listed and excluded during the scrubbing process (not implemented yet). The methodology previously outlined revealed also a number of characteristics that can be considered as potential drawbacks (depending on the application). For instance, a number of domain-specific acronyms have been scrubbed which implies that some valuable information is lost, data cleansing might be necessary in order not to lose valuable information; e.g. by expanding acronyms. There were also a few medical terms that we could identify as scrubbed, this depends on either the limitations of the standardised taxonomies with respect to their coverage or because of misspellings or ad hoc variant term forms such as *brsm-buksm* (lit. ‘bröstsmärta-buksmärta’) i.e. chest pain-abdominal pain. A demonstration interface has been implemented that illustrates the functionality of the scrubbing process; at the same time the user has the possibility to test the text understandability by choosing appropriate values that can be used for refining the evaluation. More tests are planned during the near future.

Different methods for de-identification of sensitive data such as clinical and financial data must be sought since it is a well know fact that manually removing PHI is a time consuming, tedious and costly enterprise. The difficulty of the task is illustrated in experiments described in Dorr et al., (2006) where it is pointed out that even simple PHI is difficult to automatically identify with the exactitude required by HIPAA. Also, a major problem that has been recently recognized is the lack of metrics that can quantify the risk of re-identification and information preservation using different de-identification techniques Hirschman & Aberdeen (2010). The first goal of the presented work is easier to evaluate the second, information preservation, harder. Therefore we will let human subjects read the results and grade in a scale how well they understand the resulted text content or if the results are sufficient for making available text databases for medical/clinical research. In the future we intend to integrate and combine more standardized resources in order to achieve a higher lever of understanding and experiment with other thresholds.

References

Berman, J.J. 2003. Concept-Match Medical Data Scrubbing. *Path Lab Med*, 127:680-86.

Dorr, DA, Phillips, WF, Phansalkar, S., Sims, SA, Hurdle, JF. 2006. Assessing the difficulty and time cost of de-identification in clinical narratives. *Inf Med*. 45(3):246-52.

El Emam, K., Fineberg, A. 2009. An Overview of Techniques for De-Identifying Personal Health Info. <www.ehealthinformation.ca/documents/>

Health Insurance Portability and Accountability (HIPAA), 2003. Privacy Rule and Public Health Guidance. CDC & U.S. Dep. of Health and Human Services.

Hirschman, L., Aberdeen, J. 2010. Measuring Risk and Information Preservation: Toward New Metrics for De-identification of Clinical Texts. Text and Data Mining of Health Documents, a NAACL workshop. LA, CA.

Hrynaszkiewicz, I., Norton, ML., Vickers, AJ., Altman, DG. 2010. Preparing Raw Clinical Data for Publication: guidance for journal editors, authors&peer reviewers. *Trials* 11:9.

Hsinchun, C., Fuller, S.S., Friedman, C. and Hersh, W. 2005. *Medical Informatics – Knowledge Management and Data Mining in Biomedicine*. Pp. 109-121. Springer.

Kokkinakis, D., Thurin, A. 2007. Anonymisation of Swedish Clinical Data. The 11th Conf. on Artificial Intelligence in Medicine (AIME). Pp. 237-241. Netherlands.

Kokkinakis, D. 2009. Lexical granularity for automatic indexing and means to achieve it - the case of Swedish MeSH®. *IR in Biomedicine: NLP for Knowledge Integratio*n. Pp. 11-37. IGI Global.

Kokkinakis, D., Gerdin, U. 2009. Issues on Quality Assessment of SNOMED CT® Subsets - Term Validation and Term Extraction. Proceedinsg of RANLP-2009 Worksop: Biomedical IE. Bulgaria.

Kokkinakis, D. 2004. Reducing the Effect of Name Explosion. LREC Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP tasks. Portugal

Meystre, SM., Savova, GK., Kipper-Schuler, KC., Hurdle, JF. 2008. Extracting Information from Textual Documents in the EHR. *Yearb Med Inf*. 128-44.

Neamatullah, I., et al. 2008. Automated De-Identification of Free-Text Medical Records. *BMC Med Info and Decision Making*, 8:32.

Pestian, JP., Itert, L., Andersen, C., Duch W. 2006. Preparing Clinical Text for Use in Bio. Research. *J Db Manag*, 17(2), 1-11.

Safran, C., et al. 2007. Toward a National Framework for the Secondary Use of Health Data: An Am Med. Inf Assoc. White Paper. JAMIA 14:1-9 doi:10.1197/jamia.M2273

Sweeney, L. 1996. Replacing Personally-Identifying Information in Medical Records, the Scrub System. *J. of the Am Med Informatics Assoc*. Pp. 333-337.

Sweeney, L. 2002. k-anonymity: a Model for Protecting Privacy. *J. Uncertain. Fuzziness Knowl.-Based Syst.* 10(5): 557-570.

Uzuner, O., Kohane, I., Szolovits, P. 2006. Challenges in Natural Language Processing for Clinical Data Workshop.

Velupillai, S., Dalianis, H., Hassel, M., Nilsson G. 2009. Developing a Standard for de-identifying Electronic Patient Records Written in Swedish: Precision, recall and F-measure in a Manual and Comput. Annot. Trial. *J. of Med. Inf.* vol. 78:12, pp. e19-e26.