This document is published in:

Rebeca Pérez-Laínez, Ana Iglesias, César de Pablo-Sánchez, Anonimytext: anonimization of unstructured documents, Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, 6-8 October, 2009. First International Joint Conference, IC3K 2009, Funchal, Madeira, Portugal. INSTICC, 2009, p. 284-287. ISBN 978-989-674-011-5

ANONIMYTEXT: ANONIMIZATION OF UNSTRUCTURED DOCUMENTS

Rebeca Perez-Lainez, Ana Iglesias, Cesar de Pablo-Sanchez
Universidad Carlos III de Madrid, Avenida de la Universidad 30,28911 Leganes (Madrid) España
rplainez@inf.uc3m.es,aiglesia@inf.uc3m.es,cdepablo@inf.uc3m.es

Keywords: Anonymization, Medical Records, Sensible Data, Private Data, De-identification, Clinical Notes

Abstract: The anonymization of unstructured texts is nowadays a task of great importance in several text mining applications. Medical records anonymization is needed both to preserve personal health information privacy and enable further data mining efforts. The described ANONYMITEXT system is designed to de-identify sensible data from unstructured documents. It has been applied to Spanish clinical notes to recognize sensible concepts that would need to be removed if notes are used beyond their original scope. The system combines several medical knowledge resources with semantic clinical notes induced dictionaries. An evaluation of the semi-automatic process has been carried on a subset of the clinical notes on the most frequent attributes.

1 INTRODUCTION

Nowadays the task of anonymizing texts is fundamental to preserve the security of information in certain application domains. After anonymizing texts, they should be legible but they could not disclose individual information. For example, in the health information domain, de-identification is an important task if medical records are used for judicial purpose, epidemiologic studies, research, etc.

Most countries have developed its own legislation to preserve medical records privacy. In this paper, the American Health Insurance Portability and Accountability Act (HIPPA) of 1996, and the Spanish Law for Protection of Personal Data (LOPD) (1999) have been taken into account to de-identify the clinical documents.

De-identification is defined as the process of identify, select and remove sensible data that appear in a text. Sensible data can be defined as personal data which could be used to identify a person and do not have an explicit purpose for the final application.

The de-identification task can be addresed using Natural Language Processing (NPL) and Information Extraction (IE). This challenge remains interesting because usually these medical records are often unstructured, ungrammatical and they usually present some misprints, dificulting the de-identification task.

The main objective of ANONIMITEXT is to de-identify clinical notes used in a spanish hospital. The system acquires sensible data from text by using the dictionary induction technique.

2 RELATED WORK

Several research groups have been working on developing techniques to de-identify unstructured English medical records according to HIPAA. Most of the present approaches fall into two categories: Dictionary Based Techniques (DBT) or Machine Learning Techniques (MLT).

The UMLS metathesaurus (Bodenreider, 2004) is an essential clinical resource that some authors like (Ruch, Baud, Rassinoux, Bouillon, & Rober, 2000), (Gupta, Saul, & Gilberston, 2004) and (Morrison & Li, 2008) use to recognize medical terminology. The remaining tokens should be considered as candidates

for de-identification. To avoid removing too much content which is sometime not accurately identified, other resources like dictionaries for personal names, surnames, etc. are often used.

On the other hand, MLT and IE were used by authors like (Aramaki & Miyo, 2006), (Szarvas, Farkas, & Busa-Fekete, 2007) to extract medical records information, however (Sim & Wright, 2005) use it to minimize the number of values which should be hidden.

Both techniques have advantages and disadvantages: DBT are fast, but protection is limited by the coverage of used dictionaries. Moreover, the use of DBT is hindered by the problem of ambiguous terms, that is, terms which could have more than one meaning (Ruch et al, 2000). However, MLT are useful to obtain inference rules which generalize the model beyond the training data, but a large amount of training data is needed to learn effective models. Besides, if the source of the data changes, retraining the models is needed to guarantee the performance.

A complementary idea that has been applied in Adaptive IE consists on the acquisition or induction of semantic dictionaries from a large collection of documents in the same domain of the application. This technique only requires specifying a set of interesting concepts that are prominent (seeds) in the domain. Semantic dictionary induction is often less expensive than annotating full documents as it only requires to specify related seeds.

3 ANONYMITEXT ARCHITECTURE

ANONYMITEXT system is designed to recognize sensible data in unstructured documents to enable their de-identification. The system combines general and domain knowledge resources with automatically induced dictionaries. The system input is a set of unstructured documents. The output is the de-identified input corpus.

The system architecture is composed of five steps: *Dictionary Induction, Tagger, Adviser, Expert Revision and Anonymizer* (Figure 1).

In the *Dictionary Induction* step, a domain expert selects the set of seeds examples that are frequent in the corpus like person, hospital names, etc. The dictionaries are created by extending the seeds with new terms that co-occur in similar contexts using the collection of clinical notes.

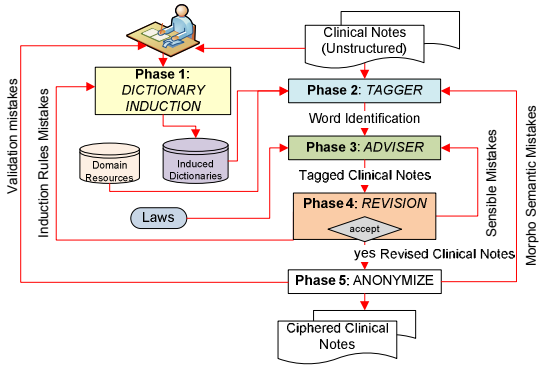

Figure 1: ANONYMITEXT De-Identification architecture.

The *Tagger* step performs a morphological and semantic analysis of the text, which is tokenized, split in sentences and enriched with part of speech information. Induced dictionaries are used to include semantic information. Moreover, other domain resources could be used to improve semantic tagging, as the UMLS metathesaurus. This phase finds the semantic information that will be used to de-identify sensible information in next phases.

In the *Adviser* step, the system detects sensible data from the documents according to the country information security laws in the domain of the system. Sensible data is marked for the expert to make their final decision on which data will be preserved in the next step.

An interface will show to the expert the documents semantically tagged and the source of these tags (induced dictionary, biomedical resource, etc). The task of the expert is to accept or not the recommendation of the system reporting the cause of reject. Among the causes for rejection we can find: 1) a *dictionary induction mistake*, 2) a *morpho-semantic mistake*, that occurs when an ambiguous word is incorrectly tagged, 3) *an advice mistake*, this is when the system advises to hide all words not tagged as personal information. Feedback data is logged and would be useful to adjust previous phases (*Dictionary Induction, Tagger, Adviser*) .The models for the different parts could be retrained or improved using a similar idea than active learning. The system could learn continuously to become more efficient reducing the time that the expert spends in the *Revision* step.

Finally, in the *Anonymizer* step sensible data are ciphered with a public-key algorithm or a hash function.

4 EXPERIMENT DESIGN

ANONYMITEXT system has been evaluated using a corpus of 60 Spanish clinical notes from a Spanish hospital. These clinical notes contain sensible data such as *patient names, patient ages, phones, cities, dates, medical facility names or doctor names*, according to the HIPAA and LOPD security laws.

Three domain experts participated in the experiment annotating manually the gold standard corpus. Moreover, these experts collaborated in the *Dictionary Induction* phase obtaining frequent seeds examples. The induced dictionaries were obtained automatically by using the whole collection (210.700 clinical notes).

Due to the low frequency of some sensible data in the corpus, this paper is focused in the evaluation of the de-identification of *doctor, medical facility and patient names*. From the corpus used for the experiment, 172 tokens belong to the *Doctor Name class*, 79 *Patient Names*, and 107 *Medical Facility Names* were identified.

In the corpus, most of clinical notes present sentences that are not tabulated and they do not fulfil grammar rules, so sentence analysis become difficult.

Therefore, the evaluation process is composed of the next phases:

1) *Corpus Annotation*: Firstly, the domain experts annotated the medical records using a common set of tags for sensible data. These tags were clearly defined taking into account the HIPPA and LOPD laws. Secondly, we checked if tags were correctly defined and if they were understand in the same way by the annotators. To ensure that the annotation process had been correctly executed, the agreement level between annotators was calculated with the Kappa measure (Sim & Wright, 2005).

2) *Dictionary Induction*: Next, the domain experts were asked to obtain seeds from the corpus. These seeds were used to induce *person name, doctor name and medical facility dictionaries*. The tool used for this induction is SPINDEL (De Pablo-Sanchez & Martínez, 2009).

3) *De-Identification*: This phase includes morpho-semantic analysis of the clinical texts and the anonymization phase in which sensible data is hidden. For the morpho-semantic analysis, ANONYMITEXT uses STILUS tool (Villena, González, & González, 2002). STILUS includes resources for classifying semantically a token as person, organization or location. To tag sensible tokens, two alternatives have been taken into account: A) search the token into an induced dictionary, if it is found then it will be tagged. B) If STILUS tags a word as organization, location or person then the word is searched into the induced dictionary. If the semantic category of the induced dictionary matches up with the semantic category of STILUS, then the word is tagged, otherwise not. STILUS includes few biomedical terms so it has been necessary to use biomedical specific resources as a Spanish health acronyms dictionary (Yetano & Alberola, 2003), an active principles dictionary (Cantalapiedra, 1989) and the SNOMED metathesaurus (Spackman, Campbell, & Cote, 1997).

Once medical records have been analyzed, and the tokens are tagged as *Patient_Name* or *Medical_Facility*, are ciphered using SHA-1 security algorithm (De Cannièr et al, 2006).

4) *Evaluation*: In this phase, we compare the annotations provided by ANONIMITEXT with the manually annotated documents. Precision, Recall and $F_{0.5}$-Measure have been calculated at the token level. (β=0.5 weights precision twice as much as recall).

5 RESULTS

Table 1 shows a summary of main results obtained for the experiment.

Table 1: Results for ANONYMITEXT

	Precision	Recall	F-Measure
Person_Name	89.5	67.85	84.15
Medical_Facility	26.21	23.68	25.66
Overall	67.22	53.6	63.97

Due to STILUS classify the tokens in the same way as induced dictionaries; precision, recall, and F-Measure obtained good values for *Person_Name* class. However, precision is not 100% because STILUS does not allow splitting certain tokens like a surname followed by a punctuation sign. It is one of the STILUS limitations.

On the other hand, the system did not achieve good results for the de-identification of *Medical_ Facility* names. Analysing the results, two main causes were found: 1) *semantic ambiguity of terms*, that is, polysemous words which depending on the context, could refer to a sensible data or not. Moreover, a great majority of Spanish *medical facility* names contain a *person name*, which generates a semantic ambiguity between tokens belonging to *Person_Name* class and *Medical_Facility* class. 2) *Acronyms of Medical Facility Names* make the de-identification process

more difficult. Later analysis of the human tagged results has shown that some *medical facilities* are written with acronyms by clinicians. However the Spanish health acronyms dictionary used during the experiments only contained acronyms related with diseases and medical concepts. Unfortunately, the use of acronyms in medical records are usual, so it would be necessary to upgrade SPINDEL in order to include acronyms into the induced dictionary.

Overall results are calculated by micro-averaging for all semantic concepts. They indicate that the system is not ready yet to work automatically, although, the current system configuration, the system proposes a fast solution to identify sensible data that would make the task easier and more effective.

6 CONCLUSIONS AND FUTURE WORK

The main difference between ANONYMITEXT and previous approaches is the combination of medical resources with the use of the dictionary induction technique. The main advantage of this approach lies in the minimal effort required for a human annotator, which only needs some seeds from a subset of the corpus.

Both stages *Dictionary induction* and *Revision*, allows including tagging tules, new induced dictionaries and new system steps. Therefore, those stages make possible system scalability.

Moreover, ANONYMITEXT preserves the integrity and confidentiality of documents, because it replaces sensible data by ciphered information.

Currently we are working towards a more comprehensive evaluation of the tool including a larger number of documents and representative categories of sensible data. Besides, we are working on improving dictionary acquisition techniques. Finally, we are developing the framework that allows taking profit of the expert feedback to improve final results.

ACKNOWLEDGEMENTS

This work has been partially supported by MAVIR (S-0505/TIC-0267) and by the TIN2007-67407-C03-01 project BRAVO.

REFERENCES

Aramaki, E., & Miyo, K. (2006). Automatic De-identification by Using Sentence Features and Label Consistency. *I2b2 Workshop on Challenges in Natural Language* Processing for Clinical Data.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res. 3*, 267-270.

Boletín Oficial del Estado. (1999, December 14). http://www.boe.es/boe/dias/1999/12/14/index.php

Cantalapiedra, J. (1989). Diccionario de excipientes de las especialidades farmacéuticas españolas. Madrid: Ministerio de Sanidad y Consumo.

De Cannière, C., & Rechberger, C. (2006). Finding SHA-1 Characteristics: General Results and Applications. *In Advances in Cryptology – ASIACRYPT 2006* (pp. 1-20). Springer Berlin / Heidelberg.

De Pablo-Sanchez, C., & Martínez, P. (2009). Building a Graph of Names and Contextual Patterns for Named Entity Classification. *In Advances in Information Retrieval* (pp. 530-537). Springer Berlin / Heidelberg.

Gupta, D., Saul, M., & Gilberston, J. (2004). Evaluation of a de-identification (De-Id) software engine to share pathology reports and clinical documents for research. *American Journal of Clinical Pathology*, 176-186.

Im, S., & Raś, Z. W. (2005). Ensuring Data Security Against Knowledge Discovery in Distributed Information Systems. *In Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* (pp. 548-557). Springer Berlin / Heidelberg.

Morrison, F. P., & Li, L. (2008). Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes? *J Am Med Inform Assoc* , 37-39.

Ruch, P., Baud, R. H., Rassinoux, A.-M., Bouillon, P., & Rober, G. (2000). Medical Document Anonymization with a Semantic Lexicon. *AMIA Annu Symp Proc* , 729-733.

Spackman, K. A., Campbell, K. E., & Cote, R. A. (1997). SNOMED RT: a reference terminology for health care. Proceedings of the AMIA Fall Symposium, (pp. 640-644).

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *In Physical Therapy* (pp. 257-268).

Szarvas, G., Farkas, R., & Busa-Fekete, R. (2007). State-of-the-art in anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association* (pp. 574-580).

Thelen, M. (2002) Simultaneous Generation of Domain-Specific Lexicons for Multiple Semantic Categories.

U.S. Department of Health & Human Services. (2006). http://www.hhs.gov/ocr/privacy/index.html

Villena, J., González, J., & González, B. (n.d.). STILUS: Sistema de revisión lingüística de textos en castellano. *Procesamiento del Lenguaje Natural núm 29* , 305-306.

Yetano, J., & Alberola, V. (2003). Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias.