Anonymization of Unstructured Data Via Named-Entity Recognition

Fadi Hassan, Josep Domingo-Ferrer and Jordi Soria-Comas

Universitat Rovira i Virgili
Department of Computer Science and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia
UNESCO Chair in Data Privacy
Av. Països Catalans 26
43007 Tarragona, Catalonia
{fadi.hassan, josep.domingo, jordi.soria}@urv.cat

Abstract. The anonymization of structured data has been widely studied in recent years. However, anonymizing unstructured data (typically text documents) remains a highly manual task and needs more attention from researchers. The main difficulty when dealing with unstructured data is that no database schema is available that can be used to measure privacy risks. In fact, confidential data and quasi-identifier values may be spread throughout the documents to be anonymized. In this work we propose to use a named-entity recognition tagger based on machine learning. The ultimate aim is to build a system capable of detecting all attributes that have privacy implications (identifiers, quasi-identifiers and sensitive attributes). In particular, we present a proof of concept focused on the detection of confidential attributes. We consider a case study in which confidential values to be detected are disease names in medical diagnoses. Once these confidential attribute values are located, one can use standard statistical disclosure control techniques for structured data to control disclosure risk.

Keywords: Anonymization, Unstructured Data, Named-Entity Recognition, Conditional Random Fields

1 Introduction

Nowadays, large amounts of data are being collected from very diverse sources, quite often without the affected individuals being aware of it. Such a systematic data collection, coupled with new data analysis techniques, has given rise to big data. Although sometimes qualified as a buzzword, big data entail a significant change in the way data are managed. In this work, we are concerned with the privacy implications of big data, in particular unstructured big data.

In the traditional setting, data were mainly collected through surveys or from other administrative data sources. As a result, they usually had a structured nature (a table). The wide variety of data sources in the current big data context (e.g. emails sent and received, participation in social networks, etc.) forces us to

consider other types of data, such as semi-structured or unstructured data (free text). Already in 2005, it was claimed in [9] that as many as 80% of the business and medical data were stored in unstructured form. In the health-care context, a proper use of such data is critical for research and policy-making purposes, and useful for related industries such as health insurance.

The new European General Data Protection Regulation (GDPR, [8]) states that explicit consent from the affected individuals is needed to use personally identifiable information (PII) for secondary purposes (different from the primary purpose that motivated the collection, such as healthcare or service billing). Ideally, the data collector should strive to gather such consent. However, in practice this may not be feasible. It may be difficult to contact individuals to obtain their consent. Additionally, individuals with rare conditions are more likely be concerned about their privacy, which makes them less prone to grant consent for their data to be used. Due to these shortcomings, the resulting data sets will probably be biased.

To avoid the need for consent, data used for secondary purposes should no longer be personally identifiable. Anonymization, also known as statistical disclosure control (SDC), provides a way to turn PII into information that cannot be linked to a specific identified individual any more and hence is not subject to privacy regulations.

There is a substantial amount of literature on SDC for the case of structured data [10, 5, 4]. Structured data are those that can be described as a set of records each of which corresponds to an individual and contains the values of a fixed set of attributes for that individual. A common approach to anonymize structured data is to remove attributes that are identifiers and then mask quasi-identifier attributes. The latter are attributes that are not identifiers but together might allow linking the record with some external data source containing identifiers, and therefore might allow re-identifying the individual to whom a record corresponds. Alternatively, instead or in addition to masking quasi-identifiers, one can mask the confidential attributes, to introduce uncertainty about the confidential attribute values.

Once a decision has been made on which attributes are quasi-identifiers and which are confidential ones, anonymization of structured data can be fully automated. (Admittedly, in some cases the above decision may be unclear, as it depends on the background information that is assumed to be available to an intruder.) However, automation of unstructured data anonymization is much more difficult, because there is no database schema that can be followed to classify the data into identifiers, quasi-identifiers and confidential attributes. As a result, anonymizing unstructured data remains today a largely manual task.

In fact, it can be argued that unstructured textual data are the ones for which anonymization is hardest. Other types of data that might seem more difficult at first sight can be either reduced to unstructured text by using tools for automated semantics extraction (as it occurs with video and audio) or are not amenable to anonymization because their semantics is not yet sufficiently understood (as is the case for genetic data).

Contribution and plan of this paper

The purpose of this work is to automate the extraction of quasi-identifier and/or confidential attributes from unstructured textual data. That is, we want to be able to automatically identify attributes such as passport number, name, location, age, birth date, etc. For the sake of concreteness, in this work, the focus will be on medical diagnosis reports. Once this automatic identification of the relevant attributes is completed, we can apply some of the methods designed for anonymizing structured data. To identify attributes, we will take advantage of a named-entity recognition (NER) tagger [7].

In Section 2, we briefly introduce some concepts that are important to understand this work. In Section 3, previous work on document anonymization is recalled. In Section 4, we describe our proposal. Experiments are presented in Section 5 and conclusions and future work ideas are gathered in Section 6.

2 Background

2.1 Named-Entity Recognition

Named entity recognition (NER) is the task of locating and categorizing important terms in a text [17]. Named-entity recognition is a source of information for different natural language processing applications. NER has been used to improve the performance of many applications, such as answering questions [12], automatic text translation [1], information retrieval [23], and sentiment analysis of tweets [11].

NER is also useful in the anonymization of unstructured data (e.g. free text documents). In particular, it can detect those terms that might be used to reidentify an individual and those terms that contain sensitive information. Once these terms have been located, they constitute structured information that can be anonymized as usual using SDC methods (e.g. generalization, supression, etc.) to keep the disclosure risk under control.

There are many tagging schemes for NER. In this work we use the IOB2 tagging scheme [21]. In IOB2, each word in the text is labeled using one of three possible tags: I, O, or B, which indicate if the word is inside, outside, or at the beginning of a named entity. Usually, in the IOB2 tagging scheme, the B and I letters come as prefix and are followed by the category name of the named entity to distinguish between the B and I tags of different entities, e.g. in our case B-DIS refers to beginning of named entity Disease and I-DIS means within entity Disease.

2.2 Conditional Random Fields

In natural language processing, there are two common models used to solve NER tasks: hidden Markov models (HMMs), used in works such as [16, 27], and conditional random fields (CRFs), used in works such as [3, 6, 11]. NER using

CRFs is widely used and applied, and usually gives the best results in many domains, so in this work we design our model using CRFs.

CRFs [15] are conditionally trained undirected graph models often applied in pattern recognition. These models are used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes.

3 Related Work

Several techniques to anonymize unstructured textual data have been proposed. Most of them can be classified into one the following two categories: dictionary-based techniques and machine learning techniques [19].

In the past, document anonymization was carried out by manual search and replacement of the named entities. Sweeney [24] proposed the Scrub method that relies on the definition of some templates for the named entities, like location, name and country. Once these entities are found, the related value is masked.

Neamatullah et al. [18] proposed a software for document anonymization that uses lexical look-up tables, regular expressions and simple heuristics that perform context checks to locate named entities. After that, they replace these entities by non-indexed category values (e.g. replace "New York" by "[**Location**]").

Vico and Calegari [26] proposed a software architecture for document anonymization. The key idea is to recognize the named entities with an architecture of multiple natural language processing tools. After that, they replace the sensitive entities by a generic indexed category value (e.g. replace "Fever" by "generic_term_1").

In 2016, the United Kingdom Data Archive (UKDA) released a text anonymization helper tool [22]. This tool identifies numbers and words starting with a capital letter, and replaces them with "XXX".

Kleinberg et al. [13] designed Netanos, a tool to allow researchers to anonymize large texts. They use machine learning to recognize named entities (e.g. persons, locations, times and dates). Then, they replace them by a privacy-preserving indexed category value (e.g. "Location_1", "Person_1").

4 Methodology

The aim of this work is to locate terms in an unstructured text that can have privacy implications, either because they can be used to re-identify an individual or because they contain confidential information.

4.1 General Approach

Formally, given a collection of text documents $D_1, \ldots, D_n$, we want to locate supersets of all the privacy-relevant attributes they contain. Specifically, we want to come up with a superset of identifier attributes $\mathcal{ID} = \{ID_1, \ldots, ID_p\}$, a superset of quasi-identifier attributes $\mathcal{QID} = \{QID_1, \ldots, QID_q\}$, and a superset

of confidential attributes $\mathcal{C} = \{C_1, \ldots, C_r\}$. The set $\mathcal{ID}$ should contain the identifier attributes that appear in at least one of the documents; for example, $\mathcal{ID}$ will contain "Passport no." if at least one of the documents contains a passport number (even if the other documents contain no passport number). Similarly, the set $\mathcal{QID}$ should contain the quasi-identifier attributes that appear in at least one document, and the set $\mathcal{C}$ the confidential attributes that appear in at least one document.

Once the above supersets have been determined, the collection of documents can be viewed as a *structured* data set with records $D_1, \ldots D_n$ and attributes that are the elements of $\mathcal{ID} \cup \mathcal{QID} \cup \mathcal{C}$. Obviously, this structured data set is likely to be a sparse one, as not all attributes take values in all documents. To anonymize this data set, we proceed as usual in the case of structured data sets. The values of attributes in $\mathcal{ID}$ should be suppressed from all records/documents and masking should be applied to attributes in $\mathcal{QID}$ and/or $\mathcal{C}$. Depending on the type of masking used, it may be necessary to deal first with the missing attribute values in some documents; imputing them by partial synthesis is a possibility [5, 10].

Thus, the problem of anonymizing unstructured data reduces to locating the appearances of the various privacy-relevant attributes in the collection of documents and then anonymizing the resulting structured data set. We can tackle the task of locating attribute appearances by building several machine learning models, each of them recognizing a different type of named entity. For example, a first model to recognize identifier attributes (e.g. passport number, social security number, etc.), a second model to recognize quasi-identifier attributes (e.g. location, birth date, age, postal code, etc.), and a third model to recognize confidential attributes (e.g. disease names, etc.).

4.2 Proof of concept

As a proof of concept, we focus on locating confidential data within medical diagnoses. We propose a model based on conditional random fields to extract the disease names from a given medical record. For a given text, this model predicts a sequence of corresponding IOB2 tags.

Once we have the predicted sequence of IOB2 tags for every token in the medical record, we can interpret this sequence of labels and extract the Disease entity entity. For instance, if we have the sentence "Retinopathy was assessed by ophthalmoscopy" and the corresponding IOB2 tags sequence {B-DIS, O, O, O}, we move through the IOB2 sequence tags and every word corresponding to a B-DIS label is considered as the beginning of a disease entity and every word corresponding to an I-DIS label is considered as being within a disease entity. Thus, a B-DIS word with all directly following I-DIS words forms one disease entity. In fact, B-DIS and I-DIS labels do the same job but B-DIS has the particular job of distinguishing between two consecutive disease entities.

Figure 1 shows the structure of the proposed model for disease name recognition. It consists of three steps:

– The first step is the tokenizer, which splits a sentence into tokens.
– The second step is the feature extractor; in this step, we use a window of three words (the current word, the previous word and the next word), and we extract the features of these words. Table 1 explains all the features we considered.
– The third step uses a CRF model, which takes the features from the second step and produces a sequence of tags for the whole sentence.

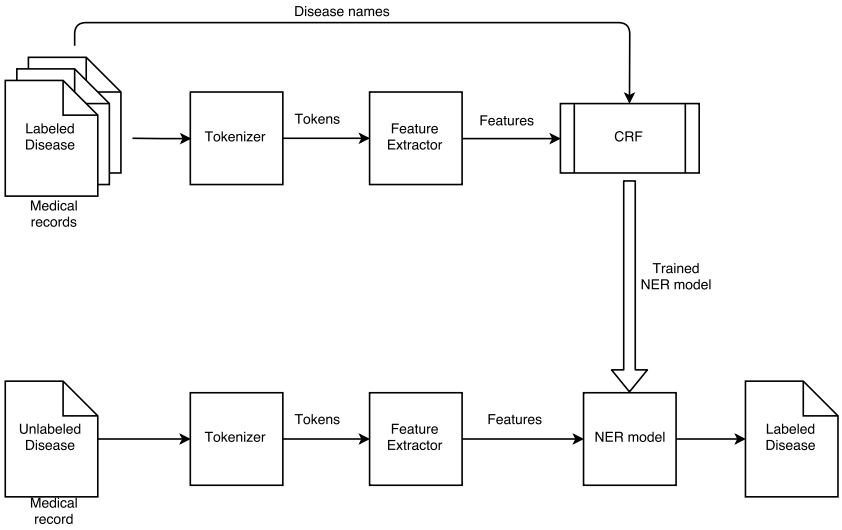

Fig. 1. Architecture of the named-entity recognition tagger

Table 1. Feature extraction

Feature	Explanation
Word stem	E.g. the stem of "illness" is "ill". We extract stems using SnowballStemmer from the nltk library [2].
Word length	The length of the word
Word shape	The shape of the word, which can be 'lowercase', 'uppercase', 'capitalized', 'mixed'
Word POS	Part of speech for the word. We use the Stanford POS tagger to extract this feature [25].

5 Experimental Results

In this section we describe the experimental results of the above-mentioned proof of concept. We programmed the experiments in Python, and we used sklearn-crfsuite for CRF [14] and SnowballStemmer for word stemming [2].

5.1 Data Set

In our experiments, we took advantage of medical texts that were labeled to study the relation between diseases and treatments. These files were obtained from MEDLINE 2001 using the first 100 titles and the first 40 abstracts from the 59 files medline01n*.xml, that are available in [20].

These data contain 3,654 labeled sentences. The labels are: "DISONLY", "TREATONLY", "TREAT PREV", "DIS PREV", "TREAT SIDE EFF", "DIS SIDE EFF", "DIS VAG", "TREAT VAG", "TREAT NO" and "DIS NO". As we were only interested in diseases, we only kept the 629 sentences with the "DISONLY" labels.

5.2 Evaluation Metrics

We used three metrics to evaluate the performance of the proposed model for the recognition of diseases:

- *Precision.* Number of diseases correctly identified by the classifier divided by the total number of identified diseases:

$$\text{Precision} = \frac{|S \cap T|}{|S|},$$

  where S is the set of all diseases identified by the classifier and T is the set of correct diseases according to the original dataset.
- *Recall.* Number of diseases correctly identified by the classifier divided by the number of correct diseases in the original dataset:

$$\text{Recall} = \frac{|S \cap T|}{|T|}.$$

- *F1.* Harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

5.3 Results and Discussion

We did the experimental evaluation in two phases: model training and model testing. Out of the 629 samples of labeled sentences, 503 were devoted to model training (80% of the samples), and 126 to model testing (20% of the samples).

Table 2. Evaluation of the model on the test dataset at word level

	Precision	Recall	F1-score
B-DIS	0.766	0.677	0.719
I-DIS	0.789	0.709	0.747
avg / total	0.778	0.693	0.733

The training phase was performed via 10-fold cross-validation, as follows. We partitioned the training data set into 10 equal-size subsamples. Out of the 10 subsamples, a single subsample was retained as validation data for testing the model while in the training phase, and the remaining 9 subsamples were used in training.

While most words in the data set were labeled as O (outside disease), we were interested in words labeled as B-DIS (beginning of disease) and I-DIS (in disease). Thus, we computed the precision, the recall and the F1 score only for B-DIS and I-DIS. For example, if we have the sentence "Diagnostic evaluation of the patient with high blood pressure", its word tokens are {"Diagnostic", "evaluation", "of", "the", "patient", "with", "high", "blood", "pressure"} and the corresponding labels are {O, O, O, O, O, O, B-DIS, I-DIS, I-DIS}. The named entity here contains three words "high blood pressure". Table 2 shows the evaluation of the predicted tags against the correct tags at the word level (separately for each word). In contrast, Table 3 reports the same evaluation metrics for whole entities. That is, in the previous example, Table 2 would separately refer to the three words "high", "blood" and "pressure", while Table 3 would refer to the entity "high blood pressure"; in the latter case, unless *all three* words of the entity were correctly labeled, the whole entity would be considered as misclassified.

Table 3. Evaluation the model on the test dataset at entity level

	Precision	Recall	F1-score
Disease Entity	0.742	0.660	0.698

According to Table 3, our model performed significantly better regarding the precision than regarding the recall. It is very likely that the recall can be increased by using more training samples. Nonetheless, we consider the above results to be promising, as a recall similar to manual labeling is achieved. Indeed, the authors of [18] asked 14 clinicians to detect and anonymize named entities in approximately 130 patient notes: the result of this manual procedure varied from clinician to clinician, with recall ranging between 0.63 and 0.94 on the data they used.

6 Conclusions and Future Work

In this work, we have dealt with the anonymization of unstructured textual data. As a proof of concept, we have focused on locating disease names (i.e sensitive attributes) in medical records. Once located, these sensitive attributes can be protected using common SDC techniques for structured data.

The main contribution of this work relates to the architecture of the recognizer for named entities. The proposed model is based on machine learning and outperforms dictionary-based NER approaches. Specifically, it avoids the out-of-dictionary problem that arises when the entities to be located are not in the dictionary being used.

As future work, we plan to extend the presented proof of concept to the detection of identifiers and quasi-identifiers. This will require investing substantial effort to generate annotated datasets for attributes such as name, location, age, etc. These annotated data sets will subsequently be used to train the identifier and the quasi-identifier detection models.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: European Commission (projects H2020-644024 "CLARUS" and H2020-700540 "CANVAS"), Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer) and Spanish Government (projects TIN2014-57364-C2-1-R "SmartGlacis" and TIN2015-70054-REDC). The views in this paper are the authors' own and do not necessarily reflect the views of UNESCO or any of the funders.

References

1. B. Babych and A. Hartley. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools: Resources and Tools for Building MT (EAMT '03)*, pp. 1–8. Association for Computational Linguistics, 2003.
2. S. Bird, E. Klein and E. Loper. *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit.* O'Reilly, 2009. The Natural Language Tooking software (nltk) is available at: `https://www.nltk.org`
3. A. Culotta, R. Bekkerman and A. McCallum. *Extracting Social Networks and Contact Information from Email and the Web. Computer Science Department Faculty Publication Series, no. 33. University of Massachusetts-Amherst, 2004.*
4. *J. Domingo-Ferrer, D. Sánchez and J. Soria-Comas.* Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections. *Morgan & Claypool, 2016.*
5. *J. Drechsler.* Synthetic Datasets for Statistical Disclosure Control. *LNS 201. Springer, 2011.*
6. *A. Ekbal, R. Haque and S. Bandyopadhyay. Bengali part of speech tagging using conditional random field. In* Proceedings of the Seventh International Symposium on Natural Language Processing (SNLP-2007)*, 2007.*

7. *J. R. Finkel, T. Grenager and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In* Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pp. 363–370. Association for Computational Linguistics, 2005.*
8. *EU General Data Protection Regulation, 2016/679.* https://gdpr-info.eu
9. *S. Grimes. Structure, models and meaning.* Intelligent Enterprise*, Mar. 2005.*
10. *A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer and P.P. de Wolf.* Statistical Disclosure Control. *Wiley, 2012.*
11. *M. Jabreel, F. Hassan and A. Moreno. Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks. In* Advances in Hybridization of Intelligent Methods*, pp. 39–55. Springer, 2018.*
12. *M. A. Khalid, V. Jijkoun, and M. De Rijke. The impact of named entity normalization on information retrieval for question answering. In* Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR'08)*, pp. 705–710. LNCS 4956, Springer, 2008.*
13. *B. Kleinberg, M. Mozes, Y. van der Toolen, and B. Verschuere.* Netanos - Named Entity-based Text Anonymization for Open Science*. Open Science Framework, Jan. 31, 2018.* https://osf.io/w9nhb
14. *M. Korobov. sklearn-crfsuite, 2015.* https://sklearn-crfsuite.readthedocs.io/en/latest/
15. *J. Lafferty, A. McCallum and F. C. N. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In* Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282-289. ACM, 2001.*
16. *S. Morwal, N. Jahan and D. Chopra. Named entity recognition using hidden Markov model (HMM).* International Journal on Natural Language Computing*, 1(4):15–23, 2012.*
17. *D. Nadeau and S. Sekine. A survey of named entity recognition and classification.* Linguisticae Investigationes*, 30(1):3–26, 2007.*
18. *I. Neamatullah, M. M. Douglass, L. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark and G. D. Clifford. Automated de-identification of free-text medical records.* BMC Medical Informatics and Decision Making*, 8(1):32, 2008.*
19. *R. Pérez-Laínez, A. Iglesias and C. de Pablo-Sánchez. Anonimytext: anonymization of unstructured documents. Universidad Carlos III de Madrid, 2009.* https://e-archivo.uc3m.es/handle/10016/19829
20. *B. Rosario and M. A. Hearst. Classifying semantic relations in bioscience texts. In* Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics *(ACL 2004).* Association for Computational Linguistics 2004. Data available from: `http://biotext.berkeley.edu/dis_treat_data.html`
21. E. F. Sang and J. Veenstra. Representing text chunks. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 173–179. Association for Computational Linguistics, 1999.
22. United Kingdom Data Service. *Text Anonymization Helper Tool.* Last Accessed: Mar. 24, 2018. `https://bitbucket.org/ukda/ukds.tools.textanonhelper/wiki/Home`
23. B. M. Sundheim. Overview of results of the MUC-6 evaluation. In *Proceedings of the TIPSTER Text Program: Phase II*, pp. 423–442. Association for Computational Linguistics, 1996.

24. L. Sweeney. Replacing personally-identifying information in medical records, the Scrub system. In *Proceedings of the AMIA Annual Fall Symposium*, p. 333. American Medical Informatics Association, 1996.
25. K. Toutanova, D. Klein, C. Manning and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252-259. 2003.
26. H. Vico and D. Calegari. Software architecture for document anonymization. *Electron. Notes Theor. Comput. Sci.*, 314(C):83–100, 2015.
27. G. Zhou and J. Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 473–480. Association for Computational Linguistics, 2002.