Automated Anonymization of Text Documents

Francisco Dias
Instituto Superior Técnico
Universidade de Lisboa
Av. Rovisco Pais
1049-001 Lisboa, Portugal
francisco.m.c.dias@tecnico.ulisboa.pt

Nuno Mamede
INESC-ID Lisboa
L2F – Spoken Language Lab
R. Alves Redol, 9
1000-029 Lisboa, Portugal
nuno.mamede@inesc-id.pt

Jorge Baptista
Universidade do Algarve
FCHS and CECL
Campus de Gambelas
8005-139 Faro, Portugal
jbaptis@ualg.pt

***Abstract*—Sharing data in the form of text is important for a wide range of activities but it also raises a concern about privacy when sharing data that could be sensitive. Automated text anonymization is a solution for removing all the sensitive information from documents. However, this is a challenging task due to the unstructured form of textual data and the ambiguity of natural language. In this work, we present our implementation of an automated anonymization system, built in a modular structure, for documents written in Portuguese. Four different methods of anonymization are evaluated and compared. Two methods replace the sensitive information by artificial labels: suppression and tagging. The other two methods replace the information by textual expressions: random substitution and generalization. Evaluation showed that the use of the tagging and the generalization methods facilitates the reading of an anonymized text while preventing some semantic drifts caused by the remotion of the original information.**

I. Introduction

Data anonymization is a process of masking or removing sensitive data from a document while preserving its original format. This process is important for sharing data without exposing to third parties any sensitive information contained in databases or documents.

Free-form text is a special type of document where data is contained in an unstructured way, as represented in natural language. Examples of this type of document may include email messages, newspaper articles or reports. From the content of these documents, it is necessary to identify text structures that represent names or unique identifiers, known as *entities*.

The problem of sharing information with third parties without causing a privacy breach already raises a concern. In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) [1] states that clinical data cannot be published unless it is deidentified. HIPAA specifies a list of seventeen categories of identifiers that should be removed prior to exposing a document to the public. Also, the General Data Protection Regulation (GDPR) [2] proposes a unified law of privacy protection on the processing of personal data and on the free circulation of such data, and it advises to take seriously the privacy of personal data in order to avoid privacy breaches.

In this work, we present an implementation of an anonymization system for free-form text documents in Portuguese, and then we evaluate the impact of different methods of anonymization over texts. This implementation is built in a modular fashion and it is based on an Named Entity Recognition (NER) tool.

This paper is divided into six sections as follows: Section II gives an overview of related work in the area of text anonymization. Section III describes the anonymization system that has been implemented in this paper, detailing the functioning of each module. Section IV presents the anonymization methods that were implemented in this paper. Section V explains the way the system was evaluated in order to present and discuss the results. Finally, Section VI concludes the paper and lists some future work.

II. Related Work

The main use of automatic text anonymization systems is to deidentificate medical records. Anonymization of medical records is crucial for publishing and using clinical data for research purposes without the risk of unauthorized access to the patients' identification. Most of these anonymization systems have been applied to texts written in English. Although some of these systems are multilingual, there is no specific anonymization system for texts written in Portuguese.

A generic anonymization system is usually composed of up to four modules: i) a module that performs text normalization and feature extraction; ii) one or more Named Entity (NE) classifiers in parallel; iii) a poll to vote the most probable class of NE; and iv) module that applies an anonymization method over the NEs and replaces the occurences of these entities in the text. From those systems, we could list the following six.

One of the first automated anonymization systems was *Scrub*. It was introduced by Sweeney [3] in 1996 and it uses pattern-matching and dictionaries. The system runs multiple algorithms in parallel in order to detect different classes of entities. In 2006, part of the i2b2 (Informatics for Integrating Biology to the Bedside) Challenge was dedicated to de-identification of clinical data. Seven systems participated in this challenge. The MITRE system, developed by Wellner *et al*. [4], achieved the highest performance. The MITRE system uses two model-based NER tools, one based on Conditional Random Fields (CRF) and another on Hidden Markov Models. Gardner *et al*. [5] developed the Health Information DE-Identification (HIDE) framework for de-identification of private health information (PHI), which uses a NER tool based on CRF. Neamatullah *et al*. [6] developed the MIT Deid package.

It is a dictionary and rule-based system. MIT Deid package was made available for free on the Internet by PhysioNet. Uzuner *et al*. [7] developed the Stat De-id in 2008. This system runs a set of classifiers in parallel. Each classifier is specialized in detecting a different category of entities. More recently (2013), the Best-of-Breed System (BoB) by Ferrández *et al*. [8], a hybrid design system, uses rules and dictionaries to score a higher recall, and it also uses model-based classifiers in order to score a higher precision.

III. Anonymization System

The anonymization system here proposed is composed of a pipeline with four modules: pre-processing, NER, Co-reference Resolution (CRR), and Anonymization (Fig. 1). This pipeline receives a text document, and returns an anonymized version of the same document and a table of solutions. The *table of solutions* contains all the replacements made in the document and their positions inside the text. This modular structure makes it possible to implement this anonymization system for other languages by simply changing the tools used in each module.

Our system utilizes the STRING chain (Mamede *et al*. [9]), a hybrid, statistical and rule-based, NLP chain for Portuguese as a NLP tool. This tool is hosted at INESC-ID L2F[1]. The STRING chain performs the basic operations of a NLP tool such as text pre-processing, POS tagging, chunking, extraction of depedencies, NER, among others.

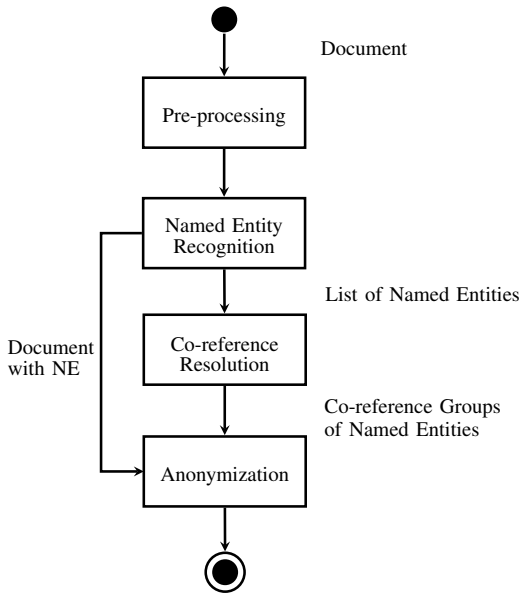

Fig. 1. Flowchart of the anonymization pipeline implemented in this work.

A. Pre-processing

The pre-processing module prepares the text to be processed by the pipeline. This operation is performed by the STRING chain. The text is normalized, splitted into sentences and then divided in tokens. After that, the STRING chain assigns some morphosyntactic features to the tokens, such as the grammatical gender, which are utilized by the next modules of our system.

[1] INESC ID's Spoken Language Systems Laboratory

B. Named Entity Recognition

The function of the NER module is to detect candidates for sensitive information in the text as is also performed by the STRING chain. It was decided, for the present system, to use a NER tool in order to detect candidates because: (i) NEs provide the most probable data to contain private information; (ii) a NER tool also provides information about the class of the entity; and (iii) previous works in anonymization were based on NER achieved above-average results (Wellner *et al*. [4], Uzuner *et al*. [7], among others).

This module receives a document and returns a list of possible entities contained in the document. Each entity is identified by its position and its class. Our implementation aims only at the three main classes of NEs from MUC-6 [10] (person, location, and organization), although STRING features a larger set of NE classes, with higher granularity.

C. Co-reference Resolution

The CRR module determines if two NEs within the same document mention the same extralinguistic object. As NEs mention the entities of interest within a text, they are important to the context where they appear in that text. During an anonymization process, the entities that mention the same object could be replaced by different expressions, making the text lose the specific information regarding the original entities. In order to preserve the information of the original document, it is important to replace all the occurrences of NEs that refer to the same extralinguistic object by the same expression. In this work, we aimed exclusively at the co-reference between NEs.

This module receives a list of NEs and returns the same list with the NEs grouped by mentions to the same extralinguistic object. A simple rule-based CRR tool was implemented for this task.

D. Anonymization

This module applies an anonymization method to the entities previously detected. Given a list of recognized entities, it removes or replaces these entities by a specific expression and returns an anonymized version of the text, along with a *table of solutions*. The table of solutions contains the anonymized entities, their positions in the document and the expressions used to replace the entities.

In order to implement some methods of anonymization, the anonymization module also needs to access external tables of entities and knowledge bases. The methods implemented in this paper are presented in the section below.

The objective of this module is to obfuscate the entities from the text in such a way that those entities cannot be re-identified while the text can still be understood by a human reader.

IV. Anonymization Methods

Different methods of anonymization can be applied to the text depending on the purpose of anonymization. In this system, we have implemented four methods of anonymization: i) suppression, ii) tagging, iii) random substitution, and iv) generalization.

A. Suppression

The *suppression* method is a simple way of anonymizing a text that consists of the suppression of the NE using a neutral indicator that replaces the original text, e.g. 'XXXXXX'.

B. Tagging

The *tagging* method consists of the replacement of the NE by a label that could indicate its class and a unique identifier. This method was implemented concatenating the class given by the NER tool and a unique numeric identifier, e.g. `[**Organization123**]`. All the entities that refer to the same object within the document are replaced by the same label, as provided by the output of the CRR module.

C. Random substitution

The *random substitution* method replaces a NE by another random entity of the same class and morphosyntactic features. This method was implemented using a *default list* containing random entities of each class. In this study, we have considered only the grammatical gender as feature. The grammatical gender of the NE is given by its headword. The headwords of NEs and their features are determined at the pre-processing stage, otherwise in the absence of features, a default grammatical gender is assigned to the headword .

Then, an entry with the same class and same gender is looked up in the default list of entities. If no suitable entity is found, the tagging method is used instead of a substitution.

Use: $defaults$ ▷ List with default entities by class
function RandSub($entity, class$)
  $hw \leftarrow$ headword($entity$) ▷ Gets the head of the NE
  $gender \leftarrow$ get-features(hw)
  if $gender = \emptyset$ **then**
    $gender \leftarrow DEFAULT$ ▷ Default gender
  end if
  if $\langle class, gender \rangle \in defaults$ **then**
    return lookup-list($class, gender, defaults$)
  else
    return tagging($class$) ▷ Uses tagging method
  end if
end function

Fig. 2. Implementation of random substitution method. The function receives as arguments an entity and its class.

Determining the gender of the headword is important for replacing NEs that refer persons by another NE from the same gender, e.g. replacing *John* by *Peter*, or replacing *Mary* by *Anna*. In Portuguese, as in some highly inflected languages, all nouns and proper names feature gender, and determiners or modifiers must show agreement with them. For this reason, it is important to replace entities of every class by another entity with the same gender, e.g. replacing *Aeroporto* ('airport'; masculine noun) by *Recinto* ('venue'; masculine noun), in order to make the output more natural.

The grammatical gender is then used to lookup a substitution entry from the list of defaults. Fig. 2 presents a pseudocode of the implementation.

D. Generalization

Generalization is any method of replacing an entity from the text by another entity that mentions an object of the same class but in a more general way, e.g.: *University of Lisbon* could be generalized to *University*, or even to *Institution*.

Use: KB ▷ Knowledge base
function Generalize($entity, class, lvl$)
  if $class = PERSON$ **then**
    return RandSub($entity, class$)
  else
    $ents \leftarrow$ query-KB($entity, class, KB$)
    if $ents = \emptyset$ **then** ▷ Then uses only the head word
      $h \leftarrow$ headword($ents$)
      $ents \leftarrow$ query-KB($h, class, KB$)
    end if
    if $ents \neq \emptyset$ **then**
      for each $e \in ents$ **do**
        $root \leftarrow$ root-KB($class, KB$)
        while $e \neq root$ **and** anon-meas(e)<lvl **do**
          $e \leftarrow$ get-parent-entry(e, KB)
        end while
        return e
      end for
    else
      return RandSub($entity, class$)
    end if
  end if
end function

Fig. 3. Implementation of generalization method. The function receives as arguments: an entity, its class and the desired level of anonymity. The function *query–KB* looks up for a given entity in the KB; the function *root–KB* returns the top object in the KB for a given class; the function *anon-meas* returns the a level of anonymization for an entry.

In order to implement a method of generalization, this module needs to access a Knowledge Base (KB) that can provide the superclasses of a given entity. The current implementation uses Wikidata (Vrandečić [11]) as an KB. Wikidata is a free collaborative knowledge base available online[2]. By querying Wikidata for an entity, it is possible to retrieve a list of statements of an entity that may include: superclasses, pseudonyms, acronyms, membership, among others. Names of superclasses can be used as generalizations of an entity. We initially consider that the superclass of a given entity is pointed by its statement '*subclass of*'. When this statement is not

[2] https://www.wikidata.org

available, the entries pointed by the statements '*instance of*' or '*member of*' can be used as the superclass. Entities with the class `Person` are not generalized. For this class of entities, the anonymization method calls the random substitution presented in subsection IV-C.

Given an entity, the system looks up for any entry in the KB whose name, pseudonym or acronym matches the entity. Then, it filters the entries by class, checking if they have a superclass in common with the current entry. In order to ensure a given level of anonymization, the implementation is also able to choose superclasses that satisfy an anonymity measure, e.g. a minimum number of child entities of the current generalization.

Wikidata provides multilingual statements for some entities, making it possible to be used also for texts in Portuguese. Fig. 3 presents a pseudocode of the implementation.

V. Evaluation and Results

A. Datasets

The anonymization system was evaluated using a dataset with 75 documents. In order to evaluate the system over different styles of text, the documents were retrieved from two different corpora: 50 documents from the golden collection of Segundo HAREM (Carvalho *et al*. [12]) with a total of 23,342 tokens, and 25 reports from the Digital Corpus of European Parliament (DCEP) (Najeh *et al* [13]) with a total of 11,497 tokens.

TABLE I
Distribution of NEs by Class in each Dataset

	Person	Location	Organization
HAREM	607	399	257
DCEP	60	43	509

These corpora have been chosen because they are rich in NEs and are structured into documents. Table I lists the number of NEs contained in each corpus, showing that the two corpora present a very different profile of NE distribution by the 3 classes here considered. The HAREM corpus contains free-styled texts from newspapers, with a wide variety of NEs. On the other hand, DCEP reports are formal texts and contain a set of NEs specific from parliamentary reports.

The documents from the DCEP were manually annotated for NEs in order to create a golden standard. The HAREM golden collection already provides annotations for NEs. The NE tagset from the HAREM golden collection was converted into the NE tagset used in the STRING chain. The same tagset was also used to annotate the set of documents from the DCEP. Only NEs from classes `Person`, `Location` and `Organization` were considered during the annotation.

All documents have been annotated for co-reference between entities. This task was performed by annotators that grouped the NEs into groups of *mentions* that are referred by two or more NEs. The documents were pre-annotated using the STRING chain before being given to the annotators.

The annotation process was perfomed by a total of three annotators. Three distinct subsets of 15 documents from the DCEP dataset and two distinct subsets of 10 documents from the HAREM dataset were randomly selected, and distributed among the annotators. A common subset of 5 documents from each dataset was given to all annotators in order to estimate the inter-annotator agreement. The resulting average pairwise percent agreement, the Cohen's kappa [14], and the Fleiss' Kappa [15] for agreement between annotators of the DCEP dataset are presented in the Table II.

TABLE II
Results for Inter-Annotator Agreement

Dataset	DCEP
Average pairwise percent agreement	97.9%
Cohen's kappa	0.913
Fleiss kappa	0.913
Observed agreement	0.980
Expected agreement	0.766

The results showed that we can consider a very high agreement between annotators and consider that the annotation is consistent among the different annotators. According to Landis & Koch [16], we could consider an almost perfect agreement between annotators based on the values of Cohen's kappa and Fleiss kappa.

The texts from this dataset were normalized, tokenized and fed into the anonymization system. We have applied the four methods of anonymization introduced in Section IV to each document. Then, the output was evaluated for its performance.

B. Metrics

The metrics of *precision* (equation 1) and *recall* (equation 2) are defined as:

$$\text{precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

where TP are true-positives, FP are false-positives and FN are false-negatives.

F1-score (equation 3) is the harmonic mean of precision and recall, and it is defined as:

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

The B^3 algorithm (Bagga & Baldwin [17]) was used to measure the performance of co-reference tasks. B^3*-precision* is defined in equation 4 and B^3*-recall* is defined in equation 5. *N* is the total number of entities. B^3*-precision* is computed as the number of correct elements in the group of an entity *i* over the number elements classified as part of the group of an entity *i*. B^3*-recall* is computed as the number of correct elements in the group of an entity *i* over the true number of elements in the group of the entity *i*. B^3-precision and B^3-recall are hereinafter called B^3*-score*.

$$\text{B}^3\text{-precision} = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{\text{\# correct elements in group } i}{\text{\# elements grouped with } i} \quad (4)$$

$$\text{B}^3\text{-recall} = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{\text{\# correct elements in group } i}{\text{\# expected elements in group } i} \quad (5)$$

The metrics of *availability* and *relevance* are used to evaluate the quality of the entities' replacements in the anonymization methods of generalization and random substitution.

Availability is defined in equation 6 as the ratio between number of entities with a parent entity in a given KB, *P*, and the total number of entities, *N* :

$$\text{availability} = \frac{\text{\# } (N \cap P)}{\text{\# } N} \quad (6)$$

Relevance is defined in equation 7 as a ratio of the sum of the relevance of all entries, R(e), over the total number of available entries, A, made in a document:

$$\text{relevance} = \frac{\sum_{e=1}^{A} \text{R}(e)}{\text{\# } A} \quad (7)$$

The relevance of a entry e is a binary function (equation 8):

$$\text{R}(e) = \begin{cases} 1 & \text{, if generalization } e \text{ is suitable in context} \\ 0 & \text{, otherwise} \end{cases} \quad (8)$$

C. Named Entity Recognition

The output of the NER classifier was evaluated against the golden standards of HAREM and DCEP, and the overall performance was measured by precision, recall and f1-score as presented on Table III and Table IV. The performance presented in the row '*All Classes*' is the performance of the NER classifier without taking into account the different classes of NEs, as if there was only one class of NE.

TABLE III
NER PERFORMANCE IN HAREM DATASET

Class	Recall	Precision	F1-score
Person	0.752	0.733	0.778
Location	0.771	0.559	0.648
Organization	0.474	0.387	0.426
All Classes	0.786	0.751	0.768

TABLE IV
NER PERFORMANCE IN DCEP DATASET

Class	Recall	Precision	F1-score
Person	0.506	0.756	0.606
Location	0.989	0.654	0.788
Organization	0.296	0.900	0.445
All Classes	0.401	0.892	0.553

The overall performance of NER was lower in the DCEP dataset as compared to the HAREM dataset mainly because of its different domain of text. The presence of foreign proper nouns and organizational entities with long forms (e.g. '*Comissão do Meio Ambiente, da Saúde Pública e da Política Urbana*') that were not detected by NER could be the main contribution to a lower recall in the DCEP dataset.

Recall of organizations was also considerably lower in the HAREM dataset when compared with other NE classes mostly because this dataset features a wider range of types of NE that refer to organizations.

D. Co-reference Resolution

The performance of the CRR module was evaluated using B^3-scores. The performance of co-reference was also compared to the case when entities are grouped into *singletons*. Singletons are groups of entities that contain only one NE per mention.

The CRR module presented a very acceptable performance, as shown in Table V. The values of B^3-score are shown for both cases when: i) CRR module is used to group NEs into mentions; ii) all NEs are considered *singletons*, i.e. each NE mentions a different extralinguistic object.

TABLE V
COMPARISON OF B^3-SCORES AFTER CO-REFERENCE RESOLUTION AND ASSUMING SINGLETON GROUPS

	Co-reference resolution		Singleton groups	
Corpus	B^3-precision	B^3-recall	B^3-precision	B^3-recall
HAREM	0.88	0.65	1.00	0.61
DCEP	0.91	0.43	1.00	0.36

The B^3-precision of the CRR module was not significantly affected by the use of different corpora. Acronyms, ambiguous entity boundaries and inconsistent classification of the entities were the main cause of resolution errors. The ratio of singleton groups in the total of entities is higher in HAREM dataset than in DCEP dataset, being a major cause to the higher B^3-recall in the HAREM dataset.

E. Anonymization

The anonymization module was evaluated for the methods of suppression, random substitution and generalization. This evaluation was based on: i) the ability of human readers to find co-references between anonymized entities only by the context surrounding the NEs; ii) on the availability of generalizations for an NE; iii) and on the relevance of the provided generalizations and substitutions.

First, three human evaluators were asked to find co-reference relations between anonymized entities, in order to understand how a human is able to retrieve some of the original information from an anonymized text. The human evaluator received the texts anonymized using the suppression method. The result is presented in table VI in terms of B^3-score. Although the human evaluator failed to find the majority of the co-reference relations between anonymized entities, as most of

the entities are contained in singleton groups, they have not contributed to a lower B^3-precision. The slightly higher B^3-recall value when compared to the Table V showed that human readers were able to group correctly some of the suppressed entities. The surrounding words acted like a major clue to detect co-reference relations, especially in the DCEP dataset. However, the lower B^3-precision value shows that readers tend to append different mentions into the same group.

TABLE VI
EVALUATION OF CO-REFERENCE PERFORMED BY HUMANS IN AN ANONYMIZED TEXT

Corpus	B^3-precision	B^3-recall
HAREM	0.88	0.67
DCEP	0.90	0.43

When using other methods of anonymization than suppression, the B^3-precision is the same as presented in table V as all entities provided by the CRR module that refer to the same extralinguistic object are replaced by the same entry.

Secondly, entities assigned by the random substitution and generalization anonymization methods were rated for the relevance of the replacement of NEs in the context of the text. This rating was made by three human readers, who assigned a binary score (equation 8) to each replacement entry. This score is based on the proper fit of the entry into the context. The rating of the relevance was performed by three human readers and showed a strong inter-rater agreement as presented in Table VII.

TABLE VII
INTER-RATER AGREEMENT OF THE RELEVANCE RATING

Dataset	DCEP	HAREM
Average pairwise percent agreement	87.5%	89.1%
Cohen's Kappa	0.747	0.769
Fleiss Kappa	0.743	0.769
Observed agreement	0.875	0.891
Expected agreement	0.514	0.528

According to the table proposed by Landis & Koch [16], we could consider an almost perfect agreement between raters as Cohen's Kappa and Fleiss kappa present sufficiently high values. The most common cause of not agreement was the presence of generalizations entities in English. Some raters would accept these generalizations within the context, but as no translations to Portuguese were available at Wikidata, these generalizations were rated as not relevant.

The results of the relevance rating are presented in Table VIII and Table IX.

TABLE VIII
RELEVANCE OF THE REPLACEMENTS OBTAINED BY THE RANDOM SUBSTITUTION METHOD

Corpus	NE Class	Relevance
HAREM	Location	0.26
	Organization	0.20
DCEP	Location	0.17
	Organization	0.39

TABLE IX
AVAILABILITY AND RELEVANCE OF THE REPLACEMENTS OBTAINED BY THE GENERALIZATION METHOD

Corpus	NE Class	Availability	Relevance
HAREM	Location	0.47	0.54
	Organization	0.89	0.23
DCEP	Location	0.66	0.90
	Organization	0.66	0.60

F. Discussion

The performance of automatic anonymization using the suppression method only depends on the performance of the NER module and does not take into account the categorisation of the NEs into different classes. This NER performance was different for both datasets. Suppression does not maintain the information, and the context around the NEs may be insufficient for a clear understanding of the text's content, as it was confusing for a human reader to resolve references between anonymized NEs (Table VI). The B^3-recall of human-made resolution is higher than assuming singleton in both corpora (Table VI), indicating that human readers were able to successfully recognize some links between entities. The B^3-precision displays a low value for both corpora, indicating that human readers also tend to group distinct mentions into the same group. In order to ensure better anonymization, the NER module must achieve an higher recall.

The tagging method enables the reader to resolve references between NEs. Although the replacement tag may allow a reader to identify the class of the anonymized entity, the result does not resemble a natural language text. The performance of this module is dependent on the performance of the NER and CRR modules. The CRR module has been shown to run with an acceptable performance, however some resolution errors occur due to inconsistent classification of entities that mention the same object. For that reason, it is important to increase both precision and recall of NER classification module in order to improve the performance of this anonymization method.

The random substitution method provides a simple solution for anonymization with a natural language output. However, the relevance of the replacements was low because entities were chosen randomly (see Table VIII), so they were often replaced by an entity of a different type (e.g. replacing a street name by a country name), which may cause drifts in the meaning of the text. One possible way of improving the relevance of this method is to use a carefully selected list of replacement entities that are vague enough, in order to avoid semantic drifts.

The generalization method provides an output similar to a natural language text. Entities are replaced by a superclass, resulting in a considerably higher relevance than that of the random substitution method. Entities that designate locations achieved a good relevance score, as they are easier to query in the KB. Entities that designate organizations usually raise NE linking issues because an entity may have several entries in the KB with the same name (e.g. *Conselho* 'Council' was generalized into a city due to an incorrect entity linking).

G. Outputs

The following tables show the outputs of the anonymization system for some sentences. These outputs feature some of the issues that may occur on automated anonymization. Table X presents the output of each anonymization method for the same sentence. Sentence X.a is the original sentence selected from the HAREM corpus, having the original entities marked in bold typeface. Two entities were not detected by the NER module, resulting in a leakage of information. In sentence X.b, it is possible to notice that all information related to NEs is lost due to the entity suppression. In sentence X.c, all entities are replaced by unique labels. Although there is no specificity on the type of entities, it is possible to understand the overall meaning of the sentence and identify that all anonymized entities mention a different object. One entity was wrongly classified as a person causing a drift in the meaning. In sentence X.d, entities were successfully replaced by random entities of the same class and grammatical gender. However, the choice of some entities changed the meaning of the sentence, e.g. replacing *Força Aérea* ('Air Force') by *Escola* ('school'). As the current system only takes into account the concordance of grammatical gender, this method also failed to replace successfully an NE, by its headword in the plural, *Controladores* ('controllers').

In sentence X.e, generalizations appear to be adequated to the context. The last entity appears written in English because the KB did not provide a translation into Portuguese to that given entry.

TABLE X
OUTPUT SAMPLES FOR EACH ANONYMIZATION METHOD

(a) Original

Nuno Severiano Teixeira visitará o contingente português estacionado em **Campo Warehouse**, composto pela **22.ª Companhia de Atiradores Pára-Quedistas** da **Brigada de Reacção Rápida** e uma equipa de **Controladores Aéreos Avançados** da **Força Aérea** .

(b) Suppression

XXXXX visitará o contingente português estacionado em XXXXX, composto pela 22.ª Companhia de Atiradores Pára-Quedistas da Brigada de Reacção Rápida e uma equipa de XXXXX da XXXXX .

(c) Tagging

[**PESSOA5**] visitará o contingente português estacionado em [**LOCAL2**], composto pela 22.ª Companhia de Atiradores Pára-Quedistas da Brigada de Reacção Rápida e uma equipa de [**PESSOA2**] da [**ORGANIZACAO2**] .

(d) Random Substitution

[**Miguel**] visitará o contingente português estacionado em [**Viseu**], composto pela 22.ª Companhia de Atiradores Pára-Quedistas da Brigada de Reacção Rápida e uma equipa de [**Filipe**] da [**Escola**] .

(e) Generalization

[**Miguel**] visitará o contingente português estacionado em [**Campo**], composto pela 22.ª Companhia de Atiradores Pára-Quedistas da Brigada de Reacção Rápida e uma equipa de [**Rodrigo**] da [**Military Organization**] .

TABLE XI
OUTPUT SAMPLE WITH CO-REFERENCE BETWEEN ENTITIES

(a) Original

Henrique solicita ao **Papa Clemente VII** a anulação do casamento. Perante a recusa do **Papado**, **Henrique** faz-se proclamar, em 1531, protector da **Igreja inglesa**.

(b) Anonymized

[**PESSOA20**] solicita ao [**PESSOA7**] a anulação do casamento. Perante a recusa do Papado, [**PESSOA20**] faz-se proclamar, em 1531, protector da Igreja inglesa.

Person names were replaced by another name of the same gender using the random substitution method. Again, the misclassification of one NE designating a human collective as an individual person resulted in the semantic drift already mentioned.

Table XI presents a text sample with the result from co-reference resolution. The original entities were marked in bold typeface. The two NEs '*Henrique*' were correctly linked to the same mentioned object. In this text, two NEs were not detected by the NER module.

VI. CONCLUSION

In this paper, we have presented an implementation of an automated anonymization system of text documents. This system was tested over different styles of text, using four different methods of anonymization. The modular structure of the system makes it possible in the future to easily replace its modules in order to support new methods of anonymization, other languages, or to improve the modules' performance.

The use of the suppression method seems to be efficient as a simple anonymization method, yet it removes relevant semantic information from the text. The tagging method is able to keep some of the information and the co-referential integrity of the mentions to the entities throughout the text. The method of random substitution makes the text have a more natural appearance to a human reader, but most of the times it results in semantic drifts because the entities are chosen randomly from a list. The generalization method presents a more acceptable solution to text anonymization while keeping the appearance natural to a human reader. However, this method is limited by the recall of the KB. Even so, generalization performs much better than random substitution. In the cases where the result is not required to be a natural text, the tagging method has shown to be one of the more acceptable solutions for anonymizing a text.

The performance of this system cannot be compared with the performance of previous systems because the evaluation used different datasets and test conditions. Most of the evaluation of previous anonymization systems aimed at the performance of the NER tool, while this work considered also the performance of the co-reference resolution and anonymization methods.

Future work will aim at the implementation of an intelligent substitution method that could be able to generalize while

maintaining more grammatical traces of the entity (e.g. grammatical number, gender and case), using different knowledge bases, and taking advantage of more information provided by the NER module in order to perform a better NE linking with the entries of the KB. Future work related to the NER module will aim to improve the performance of the current NER tool and compare the performance of the anonymization system using other NER tools already developed for Portuguese.

ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

We would like to thank Alina Popa, Ana Gomes and Paulo Bispo for their collaboration during the annotation process and evaluation, and the reviewers for their comments that were helpful to improve this paper.

REFERENCES

[1] "GPO, U.S. 45 C.F.R. 46 Protection of Human Subjects 2008," October 2008. [Online]. Available: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr46_08.html

[2] European Commission, "Proposal for a general data protection regulation," January 2012. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012PC0011&from=EN

[3] L. Sweeney, "Replacing personally-identifying information in medical records, the Scrub system." in *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association, 1996, pp. 333–337.

[4] B. Wellner, M. Huyck, S. Mardis, J. Aberdeen, A. Morgan, L. Peshkin, Y. A., J. Hitzeman, and J. Hirschman, "Rapidly retargetable approaches to de-identification in medical records," *Journal of the American Medical Informatics Association: JAMIA*, vol. 14, no. 5, pp. 564–573, 2007.

[5] J. Gardner and L. Xiong, "HIDE: An integrated system for Health Information DE-identification," in *Computer-Based Medical Systems*. IEEE Computer Society, June 2008, pp. 254–259.

[6] I. Neamatullah, M. Douglass, L. Lehman, A. Reisner, M. Villarroel, W. Long, P. Szolovits, G. Moody, R. Mark, and G. Clifford, "Automated de-identification of free-text medical records," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, pp. 1–17, 2008.

[7] Ö. Uzuner, T. Sibanda, Y. Luo, and P. Szolovits, "A de-identifier for medical discharge summaries," *Artificial Intelligence in Medicine*, vol. 42, no. 1, pp. 13–35, Jan 2008.

[8] O. Ferrández, B. South, S. Shen, F. Friedlin, M. Samore, and S. Meystre, "BoB, a best-of-breed automated text de-identification system for VHA clinical documents," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 77–83, 2013.

[9] N. Mamede, J. Baptista, C. Diniz, and V. Cabarrão, "STRING: An hybrid statistical and rule-based natural language processing chain for Portuguese," 2012. [Online]. Available: http://www.inesc-id.pt/pt/indicadores/Ficheiros/8578.pdf

[10] *MUC6 '95: Proceedings of the 6th Conference on Message Understanding*. Stroudsdburg, PA, USA: Association for Computational Linguistics, 1995, Columbia, Maryland.

[11] D. Vrandečić, "Wikidata: A new platform for collaborative data collection," in *Proceedings of the 21st International Conference Companion on World Wide Web*, New York, NY, USA, 2012, pp. 1063–1064.

[12] P. Carvalho, H. Oliveira, C. Mota, D. Santos, and C. Freitas, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 1st ed. Linguateca, 2008, vol. 1, ch. 1, pp. 11–31. [Online]. Available: http://www.linguateca.pt/LivroSegundoHAREM/

[13] H. Najeh, D. Kolovratnik, J. Väeyrynen, R. Steinberger, and D. Varga, "DCEP - Digital Corpus of the European Parliament," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 26-31, 2014, pp. 3164–3171.

[14] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[15] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[16] G. G. K. J. Richard Landis, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.jstor.org/stable/2529310

[17] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, 1998, pp. 563–566.