Multilingual Automated Text Anonymization

Francisco Manuel Carvalho Dias

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Dr. Nuno João Neves Mamede
Dr. João de Almeida Varelas Graça

Examination Committee

Chairperson: Prof. Dr. Alberto Manuel Rodrigues da Silva
Supervisor: Prof. Dr. Nuno João Neves Mamede
Members of the Committee: Prof. Dr. Jorge Manuel Evagelista Baptista
Prof. Dr. Maria Luísa Torres Ribeiro Marques da Silva Coheur

June 2016

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 the implementation of a multilingual anonymization system for text documents in four languages: English, German, Portuguese and Spanish.

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.

Resumo

A partilha de dados sob a forma de texto é importante numa vasta gama de actividades. Porém, a partilha de dados suscita preocupações quanto a privacidade no caso em que os textos contêm informação sensível. A anonimização automática de texto é uma solução para a remoção das informações confidenciais contidas em documentos. No entanto, esta é uma tarefa desafiadora devido à forma não estruturada dos dados em forma de texto e da ambiguidade da língua natural.

Neste trabalho, apresentamos a implementação de um sistema de anonimização multilingue para documentos em quatro idiomas: Alemão, Espanhol, Inglês e Português.

Quatro métodos diferentes de anonimização foram avaliados e comparados. Dois métodos substituem a informação sensível por rótulos artificiais: supressão e etiquetação. Os outros dois métodos substituem a informação sensível por expressões textuais: a substituição aleatória e generalização.

A avaliação mostrou que o uso dos métodos de etiquetação e de generalização facilitam a leitura dos textos anonimizados, evitando alguns deslizes semânticos causadas pela remoção da informação original.

Keywords

Text Anonymization
Privacy
Named Entity Recognition
Coreference Resolution
Sanitization

Palavras-Chave

Anonimização de Texto
Privacidade
Reconhecimento de Entidades Mencionadas
Resolução de Co-referências
Sanitização de Dados

Acknowledgments

I would like to show my appreciation to my advisors Nuno Mamede and João Graça for their guidance, their confidence and for the help they have given me during this work. I am also very grateful to João Graça for the opportunity to work on this project and for helping me to head my work in the correct direction from the beginning. He was the major contributor to make this project come true.

I would like to thank professor Jorge Baptista from Universidade do Algarve for all the corrections and suggestions, which were very useful whilst writing this dissertation.

I would like to thank professor Luisa Coheur for the corrections and suggestions to the draft of this manuscript.

I would like to thank the Unbabel team for their reception and support, and for making possible this project.

I would like to thank professor David Matos for his advices about statistical significance.

I would like to thank the annotators for their patience and dedication to the task.

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

Finally, I would like to express my gratitude to the following persons in their native languages:

Quero agradecer à minha mãe, Umbelina Dias, por todo o apoio incondicional e compreensão ao longo destes anos de trabalho e por, mesmo à distância, ter dado todo o precioso suporte nos momentos mais difíceis.

Vreau să-i mulțumesc lui Alina pentru prietenie, înțelegere și sprijin în realizarea acestui proiect.

Lisbon, June 24th 2016
Francisco Dias

Table of Contents

Appendix

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Designation**
API	Application Programming Interface
CoNLL	Conference on Natural Language Learning
CRF	Conditional Random Fields
CRR	Co-reference Resolution
DCEP	Digital Corpus of European Parliament
F1	F1-score
HAREM	Shared Named Entity Recognition Evaluation for Portuguese
HIPAA	Health Insurance Portability and Accountability Act
HMM	Hidden Markov Models
I2B2	Informatics for Integrating Biology and the Bedside
IC	Information Content
ICU	Intensive Care Unit
JSON	JavaScript Object Notation
KB	Knowledge Base
LDC	Linguistic Data Consortium
LOC	Location
MSF	Morphosyntactic Features
MISC	Miscellaneous
ML	Machine Learning
MUC-6	Sixth Message Understanding Conference
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
ORG	Organization
PHI	Private Health Information
PER	Person
POS	Part-of-Speech
RDF	Resource Description Framework
STRING	Statistical and Rule-Based Natural Language Processing Chain

Abbreviation	**Designation**
SVM	Support Vector Machine
TSV	Tab-Separated Values
UTHealth	University of Texas Health Science Center
XIP	Xerox Incremental Parser
XML	Extensible Markup Language

List of Terms

Term	**Meaning**
Class	The category of an entity.
Corpus	A structured collection of texts in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.
De-identification	The process of masking or removing sensitive data.
Direct identifier	Exclusive information that explicitly identifies an entity (e.g. person names, organization names, phone numbers, ID numbers, among others)
Extra-linguistic object	The physical entity or concept referred by a linguistic term.
F1-score	An evaluation measure computed as the harmonic mean of *Precision* and *Recall*.
Gazetteer	Dictionary of named entities organized by classes.
Headword	Main term of noun phrases and compound-nouns.
Indirect identifier	Information that is not exclusive of an entity but combined with other indirect identifiers could be used to identify that entity (e.g. birthdate, ZIP code, among others).
Knowledge Base	Structured repository of information.
Mention	A group of named entities that refer the same extra-linguistic object.
Pseudonym	Alias to the name of an entity.
Redaction	Text anonymization performed manually by a human redactor.
Re-identification	The process of restoring sensitive data after being previously removed.
Sanitization	The same as **de-identification**.
Sensitive information	Private information that belongs to a person or organization.
Singleton group	A mention group with only one element.
Table of Solutions	Table containing the set of anonymized entities from a document and the solutions for restoring the original entities during the re-identification.

1 Introduction

"No word matters. But man forgets reality and remembers words."

Roger Zelazny, *Lord of Light* (1967)

DATA ANONYMIZATION is a process of masking or removing sensitive information 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.

An anonymization process that aims at the definitive deletion of sensitive information is usually called *redacion* (when performed by humans), *declassification* or *sanitization*. If this information is replaced by a specific label or entry, in order to be later included in the text, the process is called *de-identification* and the reverse process of inclusion is called *re-identification*. The most extensive use of anonymization systems is automatic de-identification of medical records, preserving the privacy of patients while making available their clinical information to outsiders.

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. One possible way of detecting sensitive information from the content of these documents is to identify text structures that constitute names or unique identifiers, known as *named entities* (NE), which represent real entities in the extra-linguistic universe.

1.1 Motivation

Text anonymization is useful when sharing information in the form of text. Information sharing is important for clinical and scientific research, decision making, and business studies. However, sharing information with third parties raises a concern for privacy breaches. United States of America and European Union have regulations on the use of data containing personal identifiable information.

In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) regulation [58] states that clinical data cannot be published unless it is de-identified. HIPAA specifies a list of seventeen categories of identifiers that should be removed prior to exposing a document to the public. In the E.U., the General Data Protection Regulation (GDPR) [15] proposes a unified law of privacy protection on the processing of personal data and on the free movement of such data, and advises to take seriously the privacy of personal data in order to avoid privacy breaches.

The problem of anonymization also became a concern to the Department of Defense of the United States. In 2010, DARPA released a request for information [10] on technologies for declassification in order to identify sensitive information and make it possible to disclose publicly a declassified version of this information.

Automated Human redaction of texts is a hard and time-consuming task. In addition, there is no guarantee that human redactors are able to perform better than an automated system. A study by Douglass *et al.* [13] showed the limitation of human redactors. During the study, human redactors were able to de-identify 98% of the entities over clinical texts against 100% when assisted by an automated algorithm in the same conditions. For all these reasons, nowadays automated text anonymization becomes necessary for information sharing.

1.2 Problems

In this work, we aim at the problems of text anonymization in a multilingual context with the objective of apply anonymization to a crowdsourcing translation service at Unbabel[1]. In a translation service, texts from customers are usually delivered to humans that translate the texts in other languages. In the case of a crowdsourcing service, these texts may be exposed to third parties through the Internet, lessening the control on the spreading of that data. Some kinds of text, like contracts or legal documents, may include sensible personal information that the customer does not intend to expose to third parties. As sharing this kind of information raises a concern on the customers privacy, the translation service must de-identificate the documents before make them available to the translators.

The task of text de-identification could be performed locally by human redactors but this would still expose the data to third parties and raise a privacy problem. A solution to this problem could be the implementation of an automated text anonymization system in order to perform de-identification over the documents and automatically restore all the identified entities after translating the text. Since a translation service usually deals with texts in several languages, such anonymization system must also be able to support multiple languages as well.

[1] Unbabel is an online translation service based on a community of human editors; URL: http://www.unbabel.com/

1.3 Objective

The objective of this work is to develop a multilingual automated text anonymization system which is to be included in the Unbabel's translation pipeline. The goals of this work are:

- Build a modular anonymization system. The modular structure of the system must be flexible in order to allow the replacement of the modules and the integration of new modules in the future;
- Support at least four languages: English, German, Portuguese, and Spanish. These languages have been chosen because they are some the most requested at the Unbabel services and also because they belong to different language families (two romance and two germanic languages);
- Evaluate and improve the performance of the modules of the anonymization system;
- Integrate this system in the STRING NLP chain (Mamede *et al* [29]), a hybrid Natural Language Processing (NLP) processing tool for Portuguese hosted at the INESC-ID Lisboa;
- Integrate this system in Unbabel's translation pipeline.

1.4 Outline of the Dissertation

This dissertation is divided into seven chapters as follows:

- In Chapter 2, we present some previous works developed in the context of Text Anonymization in chronological order with an emphasis on their structure and evaluation;
- In Chapter 3, we present the evaluation metrics and resources used in our work. Here, we describe in detail the tools and datasets that have been created for supporting the development and evaluation of this work;
- In Chapter 4, we present the architecture of our text anonymization system;
- In Chapter 5, we evaluate the performance of our system, comparing different combinations of modules and text anonymization methods;
- In Chapter 6, we provide an overview of the integration of our work into the STRING chain and the Unbabel's translation pipeline;
- Finally, in Chapter 7, we conclude this work and present future directions of research.

Appendixes of the thesis support this work with UML diagrams [39] of our implementation, an exhaustive list of results and detailed information about the production of the datasets used for evaluating our system. A list of outputs of the anonymization system is also presented in the Appendixes.

2 Related Work

"Nescire autem quid ante quam natus sis acciderit,
id est semper esse puerum."

Cicero, *De Oratore* (55 BC)

A NUMBER OF automatic text anonymization systems have been developed over the last years. Most of these systems have been applied to the anonymization of medical records [50, 59] so that clinical data may be published and used for research purposes without the risk of unauthorized access to the patients' identification.

A generic anonymization system is composed of four modules (Figure 2.1): (i) pre-processing, (ii) entity detection, (iii) entity voting or classification, and (iv) replacement. The *pre-processing* module extracts features from the text in order to be used in the next modules. Some of these actions can be Part-of-Speech (POS) tagging, stopword remotion, surface or sentence features extraction, among others.

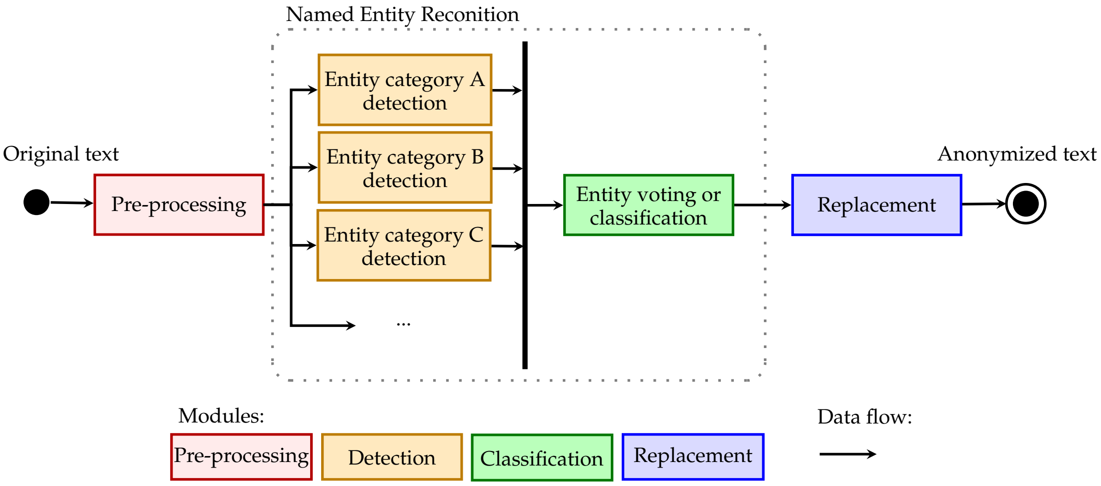


Figure 2.1: Flowchart of a generic anonymization system.

The *entity detection* module detects entity candidates in the text. This module can have one or more algorithms running in parallel, which can result in more than one classifiication for each entity. Then, the module of *entity voting and classification* refines the results from the previous module and determines the most probable classification for the entity.

The well-known process of entity detection in a text is Named Entity Recognition (NER). Most of the anonymization systems are based on NER tools [59], although some systems use algorithms based on Information Theory to detect candidates for anonymization. Anonymization systems based on a NER tool usually collapse detection and classification in just one module of NER but can use several classifiers in parallel.

The *replacement* module applies a substitution over the entities that have been detected. In some cases, the class of an extralinguistic entity is used as a substitution of the former entity.

Anonymization systems based on NER can be either rule-based (dictionaries, pattern-matching), model-based (Machine Learning (ML) models) or hybrid (combining rule-based and model-based techniques) [32, 59]. Rule-based systems have the advantage of requiring little or no training data [32] and the disadvantage of requiring the development of complex sets of rules in order to deal with a wide range of expressions and patterns to be detected. A system presented by Ferrández *et al*. [17] showed that rule-based methods could be used to achieve better precision while model-based methods could obtain better recall. Model-based systems use machine learning algorithms and need large annotated corpora for training [17, 33]. They have the advantage of having a less complex implementation and easy maintenance than rule-based systems. Most recent automated anonymization systems tend to be based on machine learning models [32], but many anonymization systems are still rule-based [3].

The performance of the anonymization systems uses to be evaluated for their recall (in some studies is called sensitivity [13]), precision and f1-score. In the previous anonymization shared tasks [50, 59] these metrics are applied at both token-level and instance-level evaluations. At *token-level*, the performance is measured having a token as the base unit of evaluation. At *instance-level*, performance is measured based on the correct detection and classification of each instance of an entity in the text. Some studies used accuracy [45, 51] to evaluate the performance and applied k-anonymity (Samatari & Sweeney [46]) as a metric of anonymization effectiveness [21] .

In this section, the automatic anonymization systems will be presented chronologically. The section covers the whole history of the development of automated anonymization systems from the first pattern-matching systems [51] until the state-of-the-art hybrid [17] and information-driven systems [47]. More anonymization systems are listed in revision works from Uzuner *et al*. [59], Meystre *et al*. [32] and Stubbs *et al*. [50].

In the end of this chapter, Table 2.1 presents the results reported from the evaluations of the systems presented in this section. Table 2.2 (on page 19) provides a summary of the characteristics of the text anonymization systems presented in this section, including methods, classifiers, features, detected classes and replacement methods.

2.1 Scrub

The Scrub was one of the first automated anonymization systems. It was introduced by Sweeney [51] in 1996 and used pattern-matching and dictionary approaches to anonymizing clinical documents.

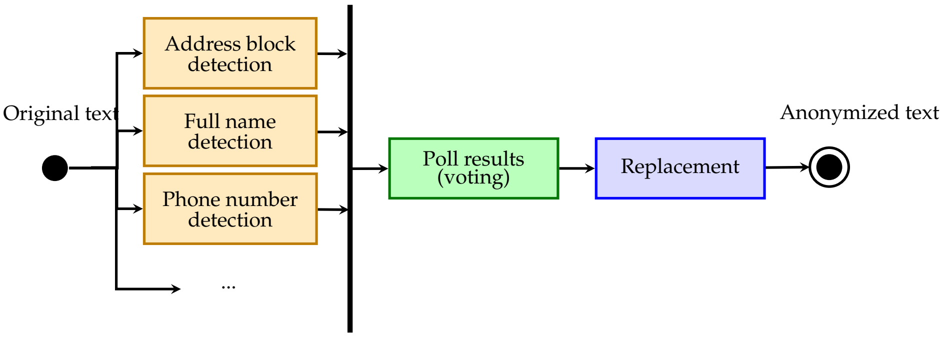


Figure 2.2: Diagram of Scrub system; adapted from the figure in Sweeney [51].

The Scrub system, depicted in Figure 2.2, ran multiple detection algorithms in parallel in order to detect different classes of entities. These algorithms may be organized in a cascade of precedences, having the more general entities to be detected first. In the end, the results for each algorithm were polled and voted. The voting process consisted of checking the results above a given threshold as compared to the result of the other algorithms. The system selects the result of that algorithm as the entity's class. Scrub detected 25 different classes, such as proper names, countries, locations, addresses, phone numbers, social security numbers, medical terms, among others. The detection of proper names was based on a dictionary lookup, and the detection of numerical entities used pattern-matching. It used an external tool in order to create fictitious replacement names. The replacement names for each match were stored inside a hashmap for further re-identification.

It was shown that using a threshold of 0.7, this system was able to detect 0.99 of the entities within a text. False-positives had been detected but not included in the evaluation results.

2.2 MEDTAG

In 2000, Ruch *et al.* [45] introduced a text anonymization system based on the MEDTAG NLP framework. MEDTAG was one of the first multilingual automated anonymization systems and was able to anonymize clinical documents in French and English. The MEDTAG system was rule-based and used a POS tagger and a lexicon of medical terms based on UMLS Metathesaurus [31]. The process involved three stages (Figure 2.3): first, the system ran a POS-tagger over the text; then classified the words into classes from the UMLS's vocabularies; and, finally, given the previous information, it determined whether a word was part of an entity to be anonymized.

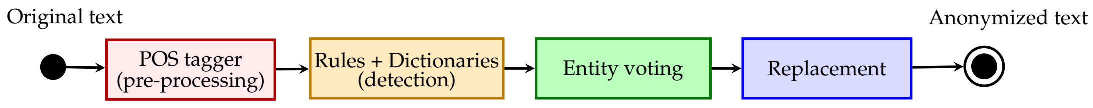


Figure 2.3: Diagram of MEDTAG system.

Using a set of rules coded by the author (approximately 40 rules) it was evaluated using 800 medical reports and scored an average accuracy of 0.97. No false-positives have been reported.

2.3 Concept-Match

Concept-Match was a system based on dictionaries developed by Berman [4] in 2003, using a remotion of stop words in a pre-processing step (Figure 2.4). This system parsed a text, removing all stop words (e.g. determinants and prepositions), and looked up the remaining words in health and biomedical vocabularies from the UMLS Metathesaurus [31] in order to find private health information (PHI) terms.

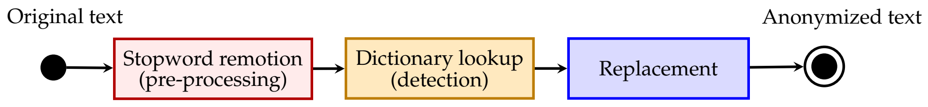

Original text
Stopword remotion (pre-processing)
Dictionary lookup (detection)
Replacement
Anonymized text

Figure 2.4: Diagram of Concept-Match system.

This system was evaluated with 567,921 phrases from the JHARCOLL public collection, a public corpus of medical texts. It was reported that this system scored a high recall and low precision due to false-positives, although no precise values have been provided. The removal of all words that do not match PHI terms makes this system very fast, de-identifying nearly 200 phrases per second. The Concept-Match[1] system has been made freely available on the Internet by the author.

2.4 i2b2 2006 Deid Challenge Systems

Part of the Informatics for Integrating Biology and the Bedside (I2B2) Challenge presented in 2006 [59], was dedicated to automatic de-identification of PHI from clinical texts. A total of seven teams with their de-identification systems participated in this challenge. Most of these systems were based on machine learning models.

Two datasets, one for training, containing 310,504 tokens, and another for testing, containing 133,623 tokens, were created from medical discharge summaries drawn from Partners Healthcare and using three human annotators. These datasets classify PHI into 8 different entity classes, featuring ambiguities

[1] URL: http://www.julesberman.info/parse.tar.gz

between entities or non-entities tokens, and out-of-vocabulary entities not present in dictionaries. These datasets can be requested at the i2b2 website upon data usage agreement.

The output of each system was evaluated based on precision, recall, and f1-score of the results for each class of entity and for the overall performance. The present section presents four systems that scored higher results on the competition.

2.4.1 MITRE

The system of Wellner *et al.* [62], later named MITRE, was based on NER methods (Figure 2.5) and included two already existent NER tools: Carafe[2], based on a Conditional Random Fields (CRF) [26] classifier, and LingPipe[3] based on Hidden Markov Models (HMM) classifiers. Regular expressions were also used in a post-processing step.

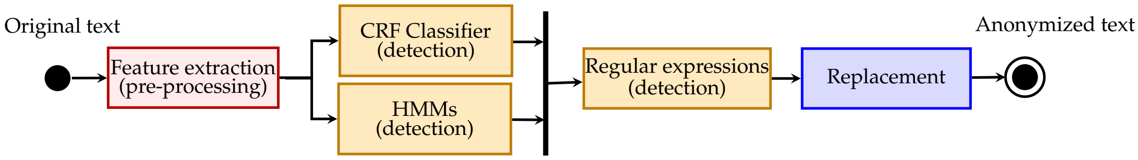

Original text
Feature extraction (pre-processing)
CRF Classifier (detection)
HMMs (detection)
Regular expressions (detection)
Replacement
Anonymized text

Figure 2.5: Diagram of MITRE system.

During the evaluation, the runs using Carafe achieved the best performance in the i2b2 challenge, scoring an f1-score of 0.983. The authors also added that the use *active learning*, using the feedback from human reviewers to add more training data, might improve the performance of the system. The MITRE system is freely available[4] with BSD license.

2.4.2 Szarvas *et al.*

Szarvas *et al*. [52] presented a system based on NER using Boosting [49] and C4.5 Decision Trees [44] from the Weka library. The NER tool was adapted to detect PHI using a set of features based on the surface word, frequency, and co-occurrence of words, dictionaries, and contextual information. No POS information was used.

Three different classifiers were trained and run in parallel to estimate if a given token belonged to a PHI (see Figure 2.6). A token is accepted as PHI if, at least, two classifiers voted positive. The system also used regular expressions to detect numeric entities.

During the evaluation, this system always scored an f1-score above 0.960 for all classes of entities. The overall f1-score was 0.976. The rich representation of features was reported by the authors as the major

[2] URL: `http://carafe.sourceforge.net/`
[3] URL: `http://alias-i.com/lingpipe/`
[4] URL: `http://mist-deid.sourceforge.net/`

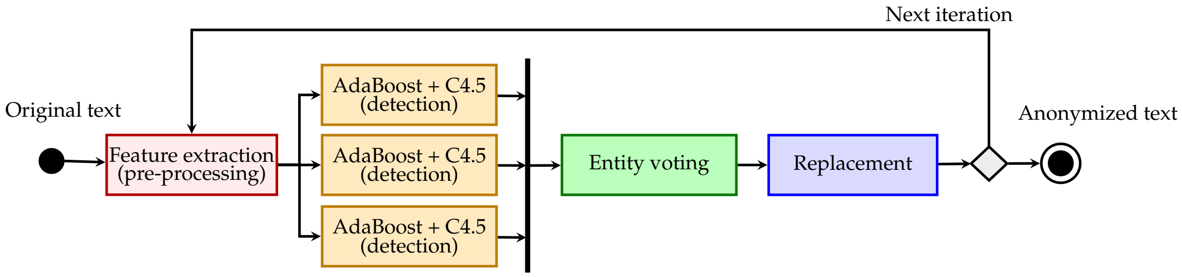


Figure 2.6: Diagram of Szarvas *et al.* system; three classifiers vote in parallel the detection of an PHI. Adapted from the figure in Szarvas *et al.* [52].

contributor to the success of the system. They also pointed out that the recall might be improved by tuning the parameters of C4.5 and Boosting.

2.4.3 Arakami *et al.*

Arakami *et al.* [1] developed a system based on a CRF classifier [26] to detect entities using three types of features: local features (morphology, POS, surrounding words), sentence features (position inside the sentence and inside the document) and extra-resource features (dictionaries of names and regular expressions to detect numeric entries). The architecture of this system is depicted by Figure 2.7. Label-consistency, a technique based on the idea that the same sequence of tokens of an entity might appear more than once inside a document, was also applied to improve the score.

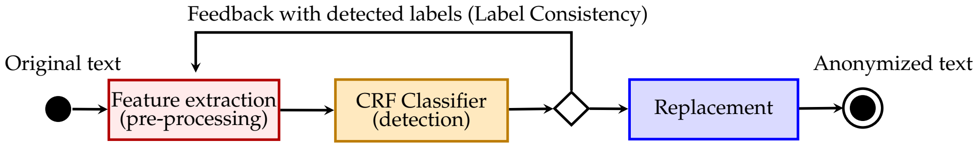


Figure 2.7: Diagram of the system from Arakami *et al.* [1].

The authors reported that the sentence features were the major contributors to detect some entities, such as identification numbers, dates and proper names, and that the use of larger dictionaries might improve the score. This system scored an above-average f1-score during the evaluation with the i2b2 corpus.

2.4.4 Hara

Hara [24] described a system based on Support Vector Machines (SVM) and regular expressions. Its processes involved four stages (Figure 2.8): pattern-matching to retrieve document titles, regular expressions to detect numeric entities, a classifier to detect textual entities inside a sentence and another classifier using YamCham[5], an open-source text chunker based on SVM to determine the class of the

[5] URL: http://www.chasen.org/~taku/software/YamCha/

entity. This classifier used POS and morphological features.

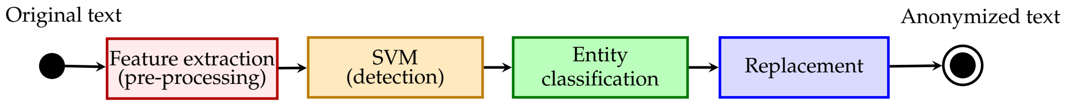


Figure 2.8: Diagram of the system from Hara [24].

The system achieved a higher score without using a sentence classifier.

2.5 HMS Scrubber

Beckwith *et al.* [3] developed in 2006 an open-source anonymization solution optimized for surgical pathology reports, since by that date no anonymization system was publicly available that was able to deal with different classes of PHI. The system, named HMS Scrubber, was developed using only open-source tools such as Java[6] and MySQL[7]. The HMS Scrubber removed sensitive PHI from surgical reports using pattern-matching and dictionaries (Figure 2.9). It replaced the entities with series of X's or a label displaying the class of PHI.

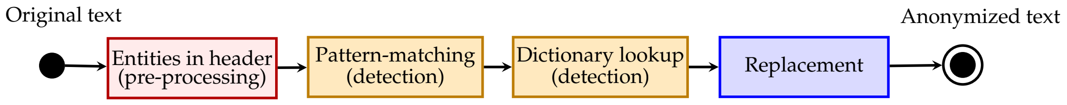

Original text
Entities in header (pre-processing)
Pattern-matching (detection)
Dictionary lookup (detection)
Replacement
Anonymized text

Figure 2.9: Diagram of HMS Scrubber system.

The evaluation of the system was made over a corpus of 1,800 surgical pathology reports and resulted in an overall recall of 0.983 and a precision of 0.424.

This system was made publicly available [8] with a BSD v3 licence and trained with the i2b2 Deid Challenge corpus (see Section 2.4).

2.6 HIDE

Gardner *et al.* [21] developed the Health Information DE-Identification (HIDE) framework for de-identification of PHI based on a CRF algorithm [26] model and the privacy model *k*-anonymity [46].

The system consisted of three components (see Figure 2.10) and aimed at extracting structured data from an unstructured data text. The first and second components were run iteratively. The first component extracts entities from the text using a trained CRF classifier [26] and the second component links the

[6] URL: http://www.java.com/
[7] URL: http://www.mysql.com/
[8] URL: https://open.med.harvard.edu/svn/scrubber/

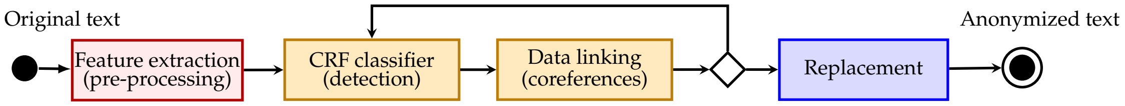

Original text
Feature extraction (pre-processing)
CRF classifier (detection)
Data linking (coreferences)
Replacement
Anonymized text

Figure 2.10: Diagram of HIDE system.

entities to previously recorded entities from the text. The third component run the anonymization of the structured data.

HIDE implemented three privacy levels for de-identification: full, partial and statistical de-identification. HIDE also distinguished between direct and indirect identifiers. *Direct identifiers* are defined as NEs that identify directly an individual. Examples of direct identifiers are proper nouns or identification numbers. *Indirect identifiers* identify some characteristics of the entity but are not able to identify, by themselves, an individual. Examples of indirect identifiers are age or gender.

Full de-identification removes all direct and indirect identifiers from the text while in partial de-identification only direct identifiers are removed. Although full de-identification is more secure, it replaces identifiers from the original text in such way that it makes it impossible for use in data mining. In order to preserve as much information as possible, with the guarantee that the probability of an entity being re-identified is kept under a given value, the system introduced the technique of statistical de-identification using k-anonymity.

The evaluation of the system was made over 100 reports from Emory Winship Cancer Institute and scored an overall accuracy of 0.982. An interesting evaluation made by the authors was about the effect of the k value of k-anonymity over the precision using the statistical de-identification. They have shown that increasing the k value boosts the privacy but decreases the precision of the data due to the anonymization process.

The system was made publicly available[9] with an MIT License.

2.7 MIT Deid

Neamatullah *et al.* [36] developed the MIT Deid package, a well-known system that was used to anonymize the MIMIC II database from PhysioNet [22]. MIMIC II database is a large dataset for research on ICU patient monitoring systems. MIT Deid was a dictionary and rule-based system (see Figure 2.11) and it used regular expressions to detect numerical entities. The system started by tokenizing the text. First, textual tokens were looked up in a dictionary. Then, pattern-matching was used to detect PHI using their context. An entity was replaced by a tag indicating its class.

Deid was evaluated using 2,434 nursing notes from MIMIC-II database previously edited to add more

[9] URL: https://code.google.com/p/hide-emory/

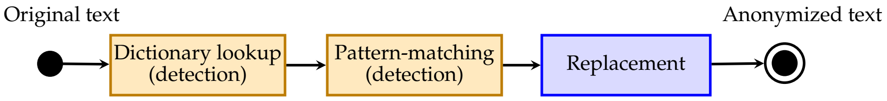


Figure 2.11: Diagram of MIT Deid system.

PHI to the text. It scored an average precision of 0.749 and recall of 0.967.

Deid package was made freely available [10] on the Internet by PhysioNet with a GNU v2 license.

2.8 Stat De-id

Uzuner *et al.* [60] developed Stat De-id in 2008, a system based on SVM. This system used a different SVM classifier to detect each class of entities (Figure 2.12). The system used some features such as morphology, POS of a token and its surrounding tokens, and the output of a Link Grammar parser[11], in order to determine a context of the whole sentence and also to deal with malformed sentences.

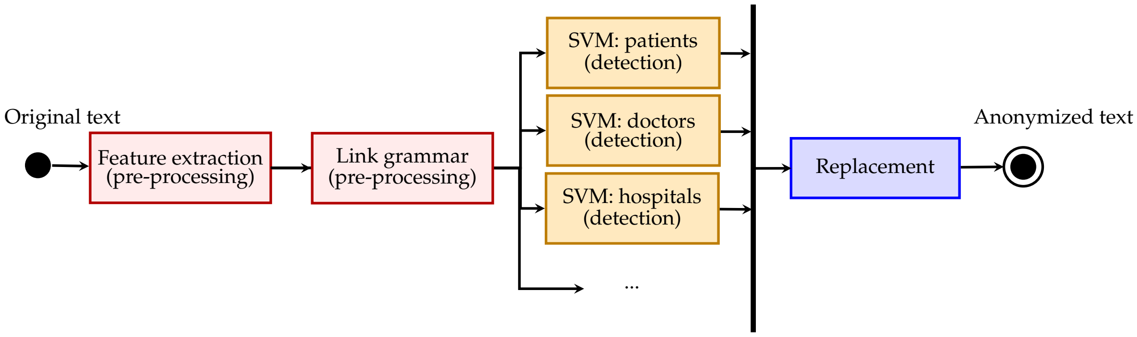

Original text
Feature extraction (pre-processing)
Link grammar (pre-processing)
SVM: patients (detection)
SVM: doctors (detection)
SVM: hospitals (detection)
...
Replacement
Anonymized text

Figure 2.12: Diagram of Stat De-id system.

The system was evaluated using 889 discharge summaries from Partners Healthcare and scored an average precision of 0.99, a recall of 0.97 and an f1-score of 0.98.

2.9 BoB

Ferrández *et al.* [17] presented Best-of-Breed System (BoB) in 2013, a new system with a hybrid design, using rules, dictionaries and machine learning methods as CRF [26] and SVM. The decision of developing a hybrid system was based on a research made by the authors which led to conclude that systems based on rules and dictionaries score a higher precision and those based on machine learning models score a higher recall.

[10] http://www.physionet.org/physiotools/deid/
[11] http://www.link.cs.cmu.edu/link/

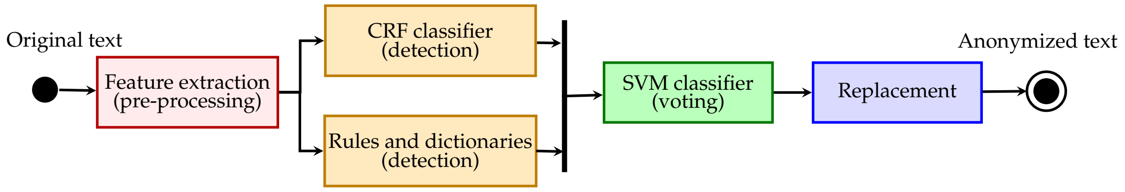

Original text
Feature extraction (pre-processing)
CRF classifier (detection)
Rules and dictionaries (detection)
SVM classifier (voting)
Replacement
Anonymized text

Figure 2.13: Diagram of Best-of-Breed system.

The system is quite complex and it is divided into three modules, as presented in Figure 2.13. The first module applies some pre-processing over the text using OpenNLP[12] tools.

The second module detects entities using rule-based and machine learning models in parallel. The rule-based model, based on pattern-matching and dictionaries, has the function of raising the recall of the system. This model uses patterns to detect numerical entities. The system also performs dictionary lookups using Lucene[13] to detect proper names and clinical terms. It introduces the idea of a fuzzy search of entities using Levenshtein distance, matching tokens when that distance is above a certain threshold. Surface features and POS are used for entity disambiguation using some heuristics. The machine learning model uses an CRF classifier [26] from Stanford NLP[14] trained for specific PHI from a corpus of 500 documents using contextual and morphological features.

The third module uses an SVM classifier, which can be applied to one-class or multi-class detection, to remove false-positives of the previous module, thus increasing the precision of the system.

This system was evaluated using different configurations over two corpora: a corpus of 800 diverse clinical documents (300 of them were used as the testing dataset) from Veterans Health Administration (VHA) and the i2b2 corpus (see Section 2.4). Its results were also compared to the MIT Deid [36] (see Section 2.7) using the same experimental setup.

The evaluation showed that BoB performs better than MIT Deid (Neamatullah *et al.* [36]) over the i2b2 corpus. The best configuration of BoB was achieved using three SVM classifiers for person names, numerical entities and ambiguous clinical terms, and a linear SVM classifier for locations and organizations. The authors reported that further work could aim at adding more patterns in order to detect more entities and test other machine learning models.

2.10 Sanchez et al.

Sánchez *et al.* [47] described a new automatic anonymization method that can be applied to a general type of textual structure and is able to preserve the utility of a text after the *de-identification* process,

[12] http://opennlp.apache.org/
[13] http://lucene.apache.org/core/
[14] http://nlp.stanford.edu/software/index.shtml

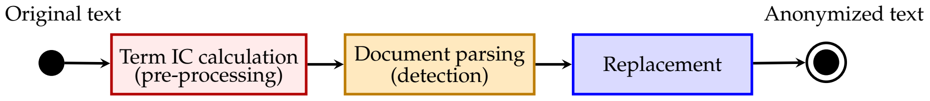

Original text
Term IC calculation (pre-processing)
Document parsing (detection)
Replacement
Anonymized text

Figure 2.14: Diagram of the system from Sánchez *et al.* [47].

based on the idea of Information Content (IC) of a term. Given the term t and its probability *p(t)* of appearing within a corpus, the IC of that term is computed as described by Equation 2.1.

$$\mathrm{IC}(t) = -log_2\ \mathrm{p}(t) \tag{2.1}$$

Instead of relying on NER methods to detect entities, this method suppresses terms that provide more IC about the text to a supposed attacker. A term can be a noun or a noun phrase (NP). A way of quantifying ICs consists in measuring the specificity of a term: the more specific a term is, the more information about the document it provides, while the more general a term is, the less information it provides. In order to have an accurate value of the IC of a term, a very large corpus is needed. As the Web could be considered as acting as a large corpus, Sánchez *et al.* use the hit counts from a web search engine to calculate its probability of appearance. Then, all terms with an IC above a computed sanitization threshold should be anonymized. This system applies suppression and generalization of terms based on WordNet [34]. The flowchart of the system is presented by Figure 2.14.

The system was evaluated for its precision, recall, f1-score and utility, although no average values have been reported by the authors. The *utility* is the ratio of IC of the anonymized document over the IC of the original document.

2.11 i2b2/UTHealth NLP Shared Task

Part of the i2b2 / University of Texas Health Science Center (UTHealth) NLP Shared Task presented in 2014 [50] was dedicated to automatic de-identification of private health information (PHI) from longitudinal narratives. A total of 10 university teams participated in this challenge. All of these systems were based on machine learning models supported by regular expressions.

The datasets for this challenge consisted of 1,304 medical records annotated with the 7 major classes of PHIs suggested by the Health Insurance Portability and Accountability Act (HIPAA). These datasets can be requested at the i2b2 website upon data usage agreement. The output of each system was evaluated based on precision, recall, and f1-score for each class of entity and for the overall performance. The systems were compared based on the resulting f1-score. The systems that achieved highest scores were based on a model-based CRF entity detection.

The present section presents three systems that scored higher performance results on the competition.

2.11.1 Nottingham

Yang & Garibaldi [64], representing the University of Nottingham, developed a hybrid system using a CRF classifier, a dictionary, and regular expressions. The system used a wide set of token-based and 3-token contextual features: word, lemma, POS, syntactic chunk tags, word form, capitalization. It also uses positional and sentence structure clues in order to detect weel-structured expressions. The dictionary and regular expressions were used to detect rare PHI. Using already detected PHIs, the system infers a list of potential PHIs that could occur simultaneously in the same document to be detected in a second pass. The structure of this system is depicted in Figure 2.15.

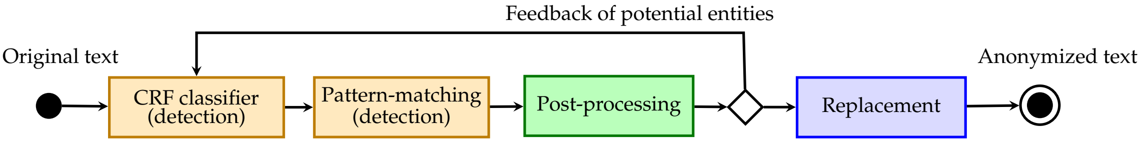

Feedback of potential entities
Original text
CRF classifier (detection)
Pattern-matching (detection)
Post-processing
Replacement
Anonymized text

Figure 2.15: Diagram of the system from Yang & Garibaldi [64].

This system scored an f1-score of 0.936, the highest score from that NLP Shared Task.

2.11.2 Harbin Grad

Liu *et al.* [28], representing the Harbin Institute of Technology, developed a hybrid system using two CRF classifiers and regular expressions. The first CRF classifier used features at context and token-level such as the word, POS, bag-of-words, word form, capitalization, among others. The second CRF classifier used features at character-level in order to avoid token boundary errors. To achieve this, the sentences were split into characters. This classifier used the same set of features as the first CRF classifier. Well-structured entities, such as phone numbers and email addresses, are detected by a pattern-matching classifier. The structure of this system is depicted in Figure 2.16.

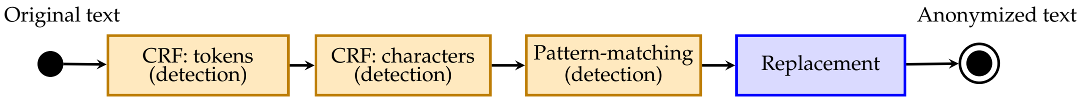


Figure 2.16: Diagram of the system from Liu *et al.* [28].

2.11.3 UNIMAN

Dehghan *et al.* [11], representing the University of Manchester, developed a hybrid system composed of a dictionary and rule-based classifier in a first stage, and a set of CRF classifiers in a second stage. In the same way as the Nottingham system, this system uses a second pass in order to inferre more candidates for PHI. The system uses a set of CRF models built in parallel. Each model is specialized on detecting each class of PHI.

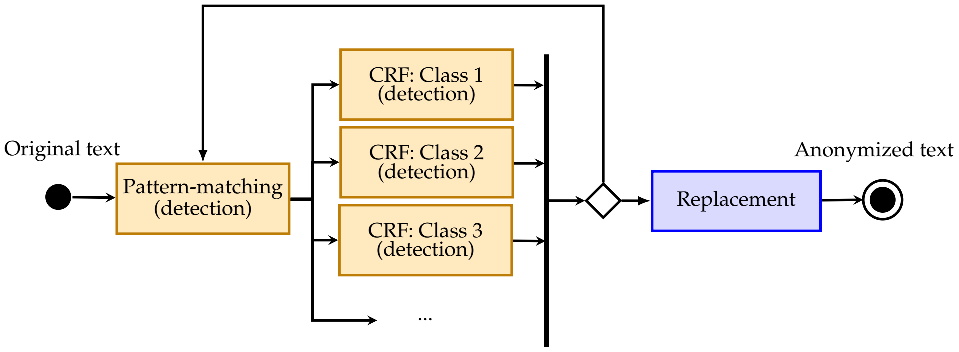


Figure 2.17: Diagram of the system from Dehghan *et al.* [11].

The structure of this system is depicted in Figure 2.17.

2.12 Dias et al.

Dias *et al.* [12] presented in 2016 a system for anonymization for Portuguese texts (Figure 2.18). This system implemented four anonymization methods and took advantage of the coreference resolution over entities' instances in order to provide a consistent labeling of mentions.

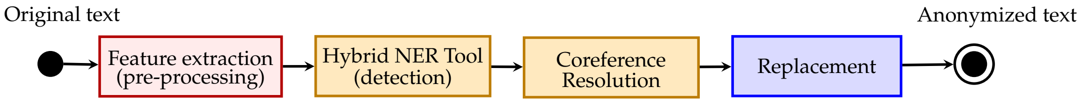

Original text
Feature extraction (pre-processing)
Hybrid NER Tool (detection)
Coreference Resolution
Replacement
Anonymized text

Figure 2.18: Diagram of the system from Dias *et al.* [12].

Apart from the common evaluation of the performance aiming at the detection of entities, this study also aimed at comparing the quality of the anonymized text for each anonymization method.

2.13 Summary

In this section, we presented some related work in the area of text anonymization. Most of these works aimed to anonymize clinical documents in order to remove PHIs. These systems can either use pattern-matching, model-based, or hybrid approaches. Most recent systems are hybrid, i.e., use pattern-matching in conjunction with model-based techniques. CRFs were the most used ML algorithm for the model-based classifiers.

Two text de-identification competitions were held in the past: the i2b2 Deid Challenge in 2006 and the i2b2/UTHealth NLP Shared Task in 2014. In these competitions, all participants teams were encouraged to developed de-identification systems that were evaluated with the same datasets.

Table 2.1 presents the performance results of the related works introduced in this section, grouped by their testing datasets. Table 2.2 presents a summary of the features of these related works.

System name	**Authors**	**Dataset**	**Precision**	**Recall**	**F1-score**
MITRE	Wellner *et al.* [62]	i2b2 2006	0.992	0.975	0.983
(Szarvas)	Szarvas *et al.* [52]	i2b2 2006	0.989	0.964	0.976
(Arakami)	Arakami *et al.* [1]	i2b2 2006	0.983	0.967	0.975
(Hara)	Hara [24]	i2b2 2006	n/a	n/a	0.920
BoB	Ferrández *et al.* [17]	i2b2 2006	0.878	0.921	0.899
MIT Deid	Neamatullah *et al.* [36]	i2b2 2006	0.734	0.489	0.587

(a) Systems evaluated using the i2b2 2006 datasets.

System name	**Authors**	**Dataset**	**Precision**	**Recall**	**F1-score**
(Nottingham)	Yang & Garibaldi [64]	i2b2/UTHealth	0.9815	0.9414	0.9611
(Harbin Grad)	Liu *et al.* [28]	i2b2/UTHealth	0.9722	0.9250	0.9480
(UNIMAN)	Dehghan *et al.* [11]	i2b2/UTHealth	0.9564	0.9366	0.9464

(b) Systems evaluated using the i2b2/UTHealth NLP Shared Task datasets.

System name	**Authors**	**Dataset**	**Precision**	**Recall**	**F1-score**
MEDTAG	Ruch *et al.* [45]	(800 reports)	0.97	1.00	0.98
HMS Scrubber	Beckwith *et al.* [3]	(1800 reports)	0.424	0.983	0.592
MIT Deid	Neamatullah *et al.* [36]	MIMIC-II	0.749	0.967	0.844
Stat De-id	Uzuner *et al.* [60]	(889 reports)	0.99	0.97	0.98
BoB	Ferrández *et al.* [17]	VHA	0.845	0.925	0.884

(c) Systems evaluated using assorted datasets.

Table 2.1: Comparison of the reported results from the evaluation of the anonymization systems presented in Section 2.

Authors	System name	Year	Methods and Classifiers	Features	Multilingue	Classes of entities							Anonymization Methods	Free
						Person	Location	Organization	Date	ID Number	Phone number	Clinical term		
Sweeney [51]	Scrub	1996	Rule-based: Regular expressions, Dictionary	Rules, dictionary lookup		✓	✓	✓	✓	✓	✓		Suppression	
Ruch *et al.* [45]	MEDTAG	2000	Rule-based: Dictionary	Rules, dictionary lookup	✓	✓		✓	✓	✓	✓	✓	Suppression	
Berman [4]	Concept-Match	2003	Rule-based: Dictionary	Dictionary lookup		✓	✓	✓	✓	✓	✓	✓	Suppression Tagging Generalization	
Wellner *et al.* [62]	MITRE	2006	NER: CRF or HMM; Regular expressions	Morphology, token, surrounding tokens (3) Dictionary classification		✓	✓	✓	✓	✓	✓		Suppression	✓
Szarvas *et al.* [52]	(Szarvas)	2006	NER, Boosting and C4.5	Morphology, token frequencies Dictionary, document titles		✓	✓	✓	✓	✓	✓		Suppression	
Arakami *et al.* [1]	(Arakami)	2006	Pattern-matching, CRF	Morphology, token, surrounding tokens (5) POS, POS of surrounding tokens Dictionary classification		✓	✓	✓	✓	✓	✓		Suppression	
Hara *et al.* [24]	(Hara)	2006	Pattern-matching, SVM	Morphology, token, lemma Document titles		✓	✓	✓	✓	✓	✓		Suppression	
Beckwith *et al.* [3]	HMS Scrubber	2006	Pattern-matching, dictionary	Dictionary lookup		✓	✓	✓	✓	✓	✓		Suppression	✓
Gardner *et al.* [21]	HIDE	2008	CRF	Morphology, token, lemma, POS		✓			✓	✓			Suppression Generalization	✓
Neamatullah *et al.* [36]	Deid Package	2008	Pattern-matching, dictionary	Dictionary lookup		✓	✓		✓	✓	✓		Tagging	✓
Uzuner *et al.* [60]	Stat DE-id	2006	SVM	Morphology, Token, Surrounding tokens (2) POS, Punctuation, Link grammar Dictionary, document titles		✓	✓	✓	✓	✓	✓		Tagging Generalization	
Ferrández *et al.* [17]	BoB	2013	Hybrid: Rule-based, Dictionary CRF, SVM			✓	✓	✓	✓	✓	✓		Suppression Generalization	
Sánchez *et al.* [47]	(Sánchez)	2013	Information Content	Tokens, POS Web lookup		✓	✓						Suppression Generalization	
Yang & Garibaldi [64]	(Nottingham)	2014	Hybrid: Regular expressions, Dictionary CRF	Token, POS, context, word form, capitalization Chunking		✓	✓	✓	✓	✓	✓		Suppression	
Liu *et al.* [28]	(Harbin Grad)	2014	Hybrid: Regular expressions, Dictionary CRF	Token, POS, context Bag-of-words		✓	✓	✓	✓	✓	✓		Suppression	
Dehghan *et al.* [11]	(UNIMAN)	2014	Hybrid: Regular expressions, Dictionary One CRF per entity class	Token, POS, context Position in line		✓	✓	✓	✓	✓	✓		Suppression	
Dias *et al.* [12]	(Dias)	2016	Hybrid NLP Chain	Tokens, POS Coreference Resolution		✓	✓	✓					Suppression Tagging Generalization	

Table 2.2: Summary of characteristics of several text anonymization systems presented in Chapter 2, listed chronologically.

3 Metrics and Resources

"I believe in observation, measurement, and reasoning, confirmed by independent observers. I'll believe anything, no matter how wild and ridiculous, if there is evidence for it."

Isaac Asimov, *The Roving Mind* (1983)

THIS CHAPTER introduces some tools utilized to assist the development and evaluation of this work: evaluation metrics, datasets, and new services and auxiliary tools. The performance of each module of this work was evaluated using the metrics that are introduced in Section 3.1. These metrics aim at understanding how this system performs given a specific set of inputs and also at comparing the performance of this system with other systems.

Some modules of this work need very specific corpora or datasets that cannot be found freely available on the Internet, or that do not exist at all. For that reason, new datasets were created aiming at the development and evaluation of these modules. These datasets are presented in Section 3.2. In addition, we developed a tool for manual annotation, as the annotation process proved to be a difficult and time-consuming task. Section 3.3 presents other tools were also developed to provide data to the system.

3.1 Metrics

Evaluation metrics in NER are normally based on precision, recall, and f1-score [56]. In our study we use different units of evaluation in order to study distinct types of errors:

1. Token-Level Evaluation - The evaluation unit is a token. This is the most common type of evaluation on NLP and the metrics are defined in the Section 3.1.1.
2. Instance-Level Evaluation - The evaluation unit is the instance of an entity. This type of evaluation targets the instances of entities that are correctly detected on their extent and class and underrates entities with partial errors in the classification.

The overall value of the metrics (precision, recall, and f1-score) can be computed in using different averaging strategies (Stubbs *et al*. [50]): macro-average and micro-average.

1. Macro-average - The metrics are calculated from the average results of the whole dataset, giving the same weight to each class.
2. Micro-average - The metrics are calculated from the average results for each document of the dataset, giving the same weight to each document.

3.1.1 Metrics for Token-level Evaluation

The metrics of precision (Equation 3.1) and recall (Equation 3.2), where the number of classified entries is indicated as true-positives (TP), true-negatives (TN), false-positives (FP) and false-negatives (FN), are defined as:

$$\text{precision} = \frac{TP}{TP + FP} \tag{3.1}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3.2}$$

F1-score (equation 3.3) is the harmonic mean of precision and recall, and is defined as:

$$\text{f1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3.3}$$

3.1.2 Metrics for Instance-level Evaluation

The metrics for instance-level were used on MUC-6 [7] and i2b2 shared tasks [50, 59] on de-identification. Each entity can be classified as: (i) correct (C), if all the tokens were correctly detected and classified; (ii) misclassified (M), if all the tokens were correctly detected but at least one was incorrectly classified; (iii) spurious (S), if tokens that do not belong to the entity were inserted; or (iv) deleted (D), if all the tokens were not detected or correctly classified. Based on this classification of entities, we define precision (Equation 3.4) and recall (Equation 3.5) as:

$$\text{precision} = \frac{C}{C + M + S} \tag{3.4}$$

$$\text{recall} = \frac{C}{C + M + D} \tag{3.5}$$

F1-score is defined as in the equation 3.3.

3.1.3 Metrics for Coreference Resolution

The B^3 algorithm (Bagga & Baldwin [2]) was used to measure the performance of coreference resolution tasks. B^3*-precision* is defined in Equation (3.6) and B^3*-recall* is defined in Equation (3.7) where P_m is the predicted chain of mentions and T_m is the true chain of mentions, for each entity m in a mention contained in a document d from a dataset D. N is the total number of entities mentioned in D.

$$\text{B}^3\text{-precision} = \frac{1}{N}\sum_{d\in D}\left(\sum_{m\in d}\frac{|P_m \cap T_m|}{|P_m|}\right) \tag{3.6}$$

$$\text{B}^3\text{-recall} = \frac{1}{N}\sum_{d\in D}\left(\sum_{m\in d}\frac{|P_m \cap T_m|}{|T_m|}\right) \tag{3.7}$$

The metrics of B^3-precision and B^3-recall are hereinafter called *B^3-score*.

3.1.4 Metrics for Information Retrieval

Availability is defined in Equation (3.8) as the ratio between number of entities with a parent entity *P* in a given Knowledge Base (KB) and the total number of entities, *N* :

$$\text{availability} = \frac{\#\,(\,N \cap P\,)}{\#\,N} \tag{3.8}$$

Relevance is defined in Equation (3.9) as the ratio of relevant substitutions compared to the total number of entity substitutions made for a document:

$$\text{relevance} = \frac{\#\text{ relevant substitutions}}{\#\text{ entity substitutions}} \tag{3.9}$$

3.1.5 Inter-Annotator Reliability Coefficients

Manually annotated corpora are based on the judgment of human annotators that may produce different responses for the same text. One of the conditions for a corpus to be considered valid is its *reliability*, i.e., the consistency of the responses among the annotators.

In our work, we compute three reliability coefficients: (i) Average pairwise percent agreement, (ii) Cohen-κ and (iii) Fleiss-κ. These coefficients were computed using the ReCal3: Reliability Calculator (Freelon [20]).

The *average pairwise percent agreement* is calculated as the percentage of observations that a pair of annotators agrees.

Cohen-κ (Cohen [8]), defined in Equation (3.10), measures the reliability of an annotation task based on pairs of annotators, where p_o is the observed proportion agreement and is p_e the expected mean proportion of agreement. These proportions can be computed from a *coincidence table* considering the annotations from a pair of annotators.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.10}$$

Fleiss-κ (Fleiss [19]) is an extension of the *Cohen-κ* to measure the reliability of any number of annotators.

3.1.6 Statistical Significance

We used the approximate randomization method (Chinchor [7], Noreen [37]) in order to assess the statistical significance between the performance of two classifiers. We use this method because the distribution of the classifications is unknown and they are nominal. We applied stratified shuffling to the classification results from each classifier and created NS distinct pairs of pseudo-results (with NS = 9,999 as in [7]) and counted how many NGE times each pseudo-result produces better performance than the unshuffled results. This way, we test if the differences between results from a pair of classifiers could be achieved by chance. Finally, we compute the significance level s as presented in Equation 3.11. If the significance level is lower than a threshold value α, we can say that the results of a pair of classifiers are significantly different for a given value α. We used an arbitrary threshold value of α = 0.1 that, according to [37], guarantees a significance level p < 0.01.

$$s = \frac{NGE + 1}{NS + 1} \tag{3.11}$$

The null hypothesis is that the values of metrics of performance (precision, recall and f1-score) are approximately the same, as they could be achieved by chance, therefore, the performance of the classifiers should not be treated as significantly different.

3.2 Datasets

3.2.1 Corpora for Evaluation of the Anonymization System

We evaluated the anonymization system with a dataset collection of text documents in four languages (English, German, Portuguese, and Spanish), from different text domains and styles. The system performance was evaluated with two different datasets for each language.

- English datasets are based on the CoNLL 2003 English testing datasets (Kim *et al.* [56]) and reports from the Digital Corpus of European Parliament (DCEP) corpus [23];
- German datasets are based on the CoNLL 2003 German testing datasets [56] and reports from the DCEP corpus [23];
- Portuguese datasets are based on the golden collection of Segundo HAREM (Carvalho *et al.* [6]) and reports from the DCEP (Najeh *et al.* [23]);
- Spanish datasets are based on the CoNLL 2002 Spanish testing datasets (Tjong Kim Sang [55]) and reports from the DCEP corpus [23].

	English		German		Portuguese		Spanish	
Dataset	**CoNLL**	**DCEP**	**CoNLL**	**DCEP**	**HAREM**	**DCEP**	**CoNLL**	**DCEP**
# Tokens	9,253	10,508	3,469	4,824	23,342	11,497	7,734	6,322
# Documents	50	25	25	15	50	25	25	15

Table 3.1: Distribution of tokens in each dataset.

Table 3.1 lists the dimension of each dataset. These corpora have been chosen because they are rich in NEs and are separated into documents. Corpora from HAREM, CoNLL 2002 and CoNLL 2003 datasets contain free-styled texts from newspapers, with a wide variety of NEs. On the other hand, DCEP reports contain formal text and a set of NEs specific from parliamentary reports.

Named Entities

The HAREM golden collection, CoNLL 2002 and CoNLL 2003 datasets already provide annotations for NEs. The documents from the DCEP were manually annotated for NEs by human annotators in order to create a golden corpus. During the annotation task, we followed the CoNLL 2003 annotation guidelines, except that we did not use the miscellaneous (MISC) class. Table 3.2 lists the number of NEs contained in each corpus, showing that both corpora for the same language present a very different profile of NEs distribution by the 3 classes here considered.

	English		German		Portuguese		Spanish	
Dataset	**CoNLL**	**DCEP**	**CoNLL**	**DCEP**	**HAREM**	**DCEP**	**CoNLL**	**DCEP**
Person	293	106	76	34	607	60	149	19
Location	361	130	95	38	399	43	130	32
Organization	259	431	132	305	257	509	240	393
Overall	913	667	303	377	1263	612	519	444

Table 3.2: Distribution of NE instances by class in each dataset.

The NE tagset from all corpora were normalized into the MUC-6 3-class annotation: *Person*, *Location*, and *Organization*. Where the dataset provides multiple annotation alternatives for each NE, we always chose the last alternative from the list.

Manual Annotation

The documents from the DCEP dataset were annotated at token-level for NE, and all the datasets were annotated for coreferences between entities. The annotation task was performed by a team of human annotators. The annotation team was composed of Portuguese students, foreign exchange students, and language teachers. There was at least one participant with native proficiency for each language. Table 3.3 shows the distribution of annotators by language.

	English	German	Portuguese	Spanish
Native	1	1	3	1
Bilingual	2	1	1	1

Table 3.3: Annotation team composition, by language and proficiency.

First, we created the golden corpora for NE as follows. The documents for annotation had been randomly selected from each dataset, split into subsets of the same size and distributed among 2 or 3 annotators, depending on the size of the dataset. A common subset of 5 documents was distributed to all annotators as presented in Figure 3.4.

Number of documents in the dataset	50	25	15
Number of Annotators	3	2	2
Documents in each subset	15	10	10
Documents in the common subset	5	5	5
Documents for each Annotator	20	15	10

Table 3.4: Distribution of annotators and documents depending on the size of the dataset.

The documents were delivered to the annotators using our Unannotator tool (Section 3.3.3). The annotators were instructed to follow the annotation guidelines of CoNLL 2003 for the manual NE annotation and the guidelines presented in Appendix E for coreference. We exemplified the process of annotation of some documents before starting the annotation task. We used the ReCal3 tool (Freelon [20]) in order to assess the inter-annotator reliability coefficients. Table C.8 presents the coefficients of the inter-annotator agreement.

Dataset	English DCEP	German DCEP	Portuguese DCEP	Spanish DCEP
Average agreement[1]	95.9%	95.5%	97.9%	97.2%
Average Cohen-κ	0.843	0.671	0.913	0.889
Fleiss-κ	0.841	0.671	0.913	0.889
Observed agreement	0.959	0.955	0.980	0.972
Expected agreement	0.742	0.862	0.766	0.746

Table 3.5: Agreement between annotators in each annotation task by language.

According to Landis & Koch [27], the strength of agreement during the annotation of NEs can be considered substantial in the German documents ($\kappa \in [0.61; 0.80]$) and almost perfect in all other cases ($\kappa > 0.80$). Portions of our annotated datasets are publicly available at https://www.l2f.inesc-id.pt/~fdias/mscthesis/corpora.

[1] Average pairwise percent agreement

3.2.2 Silver-Standard Corpora for Training NER Classifiers

During the evaluation of our system, we used auxiliary NE classifiers trained with corpora from the same text domain as part of the evaluation corpora (Section 3.2.1). As the DCEP (Najeh *et al.* [23]) datasets were used to evaluate all the language configuration of the system, we decided to use subsets of the DCEP corpora, different from the evaluation datasets in order to avoid overffitting. We also added used some WMT-News (Tiedemann [54]) in order to improve the size of the training corpora and enrich the training dataset with more NEs.

We have chosen these corpora because they contain reports and news commentaries. These types of text describe events and actions between entities, and for that reason they usually have high density of NEs. However, these datasets are not annotated for NEs and the process of creation of a golden-standard requires time and human resources. As the objective of this corpora is to train an auxiliar NER classifier to study the influence of using of this classifier over the performance, it is not required to produce a classifier with the best performance as possible. For that reason, we used the approach of Ehrmann *et al.* [14] to create automatically a multilingual NE annotated corpora based on parallel corpora.

A dataset with 7,500 parallel sentences was annotated using a semi-automated process. Sentences shorter than 80 characters were not considered, in order to discard wrong alignments and document headers.

Language pair	**Parallel corpora**	**# sentences**
English ↔ Portuguese	DCEP	1,500
English ↔ German	DCEP, WMT-News	3,000
English ↔ Spanish	DCEP, WMT-News	3,000

Table 3.6: Language pairs and associated parallel corpora used to create the annotated corpora.

The annotation of the corpora was made using a similar approach as Ehrmann *et al.* [14] and consisted of the following steps:

1. Suggestions for annotation were automatically generated by running a pair of NER tools (the same presented in Section 4.2.2, on page 37) over the parallel corpus and marking all NEs in both aligned sentences;
2. The NEs that presented the same surface form and same class in both aligned sentences were automatically annotated, and those with different classes in both aligned sentences were annotated in both sentences with the class that appear in the English sentence;
3. The NEs detected in only one of the sentences were automatically propagated to the other sentence in the case of same surface or high similarity distance between surfaces using the Levenshtein distance;

4. Other cases, such as different NE length in both sentences, have been solved manually in few cases. Any other NEs were not annotated.

3.3 Auxiliary Tools

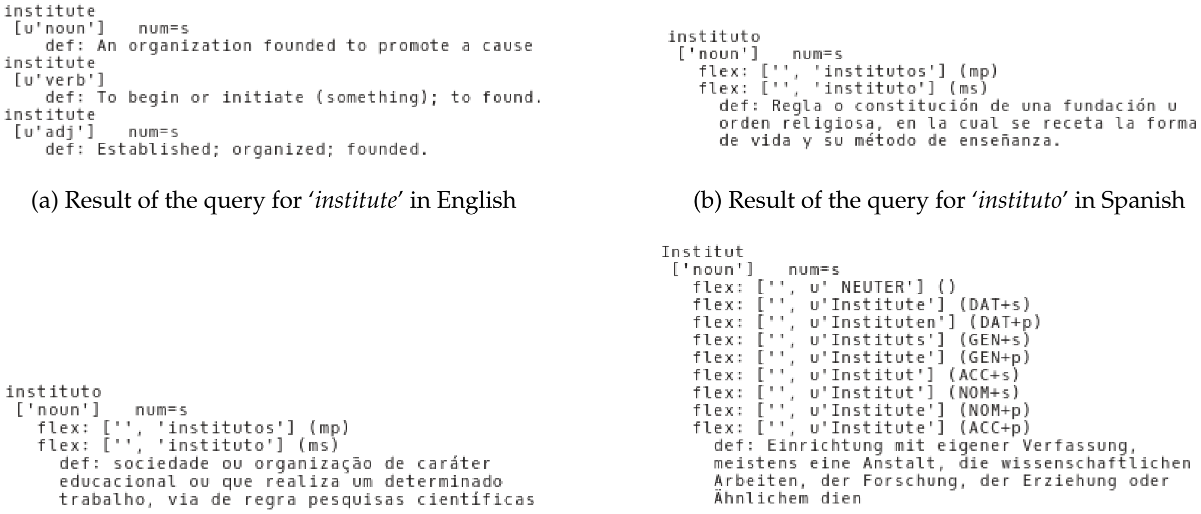
3.3.1 Wiktionary Query Service

This service provides an API to access a multilingual dictionary based on Wiktionary, which can be accessed either locally or remotely. Wiktionary [63] is a collaborative, multilingual dictionary that is freely available online. We developed a service that queries Wiktionary for terms in a given language and returns the information in the form of structured entries. The advantages of using this service include:

Standardization - Dumps from Wiktionary entries may present different structures, depending on the term or the language, which means different implementations; this service provides a standard Application Programming Interface (API) to access entires from Wiktionary;

Local dictionary - Wiktionary can be either accessed remotely or locally. Local access uses a serialized file or database to store data and has the advantage of improving access time and reducing network traffic. Both types of access use the same API.

The disadvantages of this service are: not all information about a term, such as the etymology, may be available through this service; and the service will not return any information if there is a change in the format of the dumps from Wiktionary.



```
institute
 [u'noun']   num=s
    def: An organization founded to promote a cause
institute
 [u'verb']
    def: To begin or initiate (something); to found.
institute
 [u'adj']   num=s
    def: Established; organized; founded.
```

(a) Result of the query for ‘*institute*’ in English

```
instituto
 ['noun']   num=s
   flex: ['', 'institutos'] (mp)
   flex: ['', 'instituto'] (ms)
     def: Regla o constitución de una fundación u
     orden religiosa, en la cual se receta la forma
     de vida y su método de enseñanza.
```

(b) Result of the query for ‘*instituto*’ in Spanish

```
instituto
 ['noun']   num=s
   flex: ['', 'institutos'] (mp)
   flex: ['', 'instituto'] (ms)
     def: sociedade ou organização de caráter
     educacional ou que realiza um determinado
     trabalho, via de regra pesquisas científicas
```

(c) Result of the query for ‘*instituto*’ in Portuguese

```
Institut
 ['noun']   num=s
   flex: ['', u' NEUTER'] ()
   flex: ['', u'Institute'] (DAT+s)
   flex: ['', u'Instituten'] (DAT+p)
   flex: ['', u'Instituts'] (GEN+s)
   flex: ['', u'Institute'] (GEN+p)
   flex: ['', u'Institut'] (ACC+s)
   flex: ['', u'Institut'] (NOM+s)
   flex: ['', u'Institute'] (NOM+p)
   flex: ['', u'Institute'] (ACC+p)
     def: Einrichtung mit eigener Verfassung,
     meistens eine Anstalt, die wissenschaftlichen
     Arbeiten, der Forschung, der Erziehung oder
     Ähnlichem dien
```

(d) Result of the query for ‘*Institut*’ in German

Figure 3.1: Results of queries to Wiktionary for the word ‘institute’ in four languages. The query returns a structure with the POS, grammatical number, gender and case, and the a short definition for each dictionary entry.

Figure 3.1 presents an output example from this service. If a term exists in Wiktionary, it returns a structure containing the following information about the term:

1. POS: noun, proper noun (proper), verb, adjective (adj), adverb (adv), pre- or postposition (adp);
2. Grammatical number: singular (s), plural (p), dual (d);
3. Grammatical gender: masculine (m), feminine (f), neuter (n), genderless (o);
4. Grammatical case: nominative (NOM), accusative (ACC), genitive (GEN), dative (DAT);
5. List of flexions of the term for combinations of the features (2), (3) and (4);
6. Brief description of the meaning of the term;
7. Translation of term in other languages for each meaning in (6).

3.3.2 Wikidata Query Service

The objective of this service is provide an API to access a KB. In this work, we use two KBs:

- Wikidata (Vrandečić [61]), via HTTP request;
- A local KB that can be based on Wikidata dumps or imported from an Resource Description Framework (RDF) file. The RDF format is a standard for data exchange on the Internet that stores relations in the form of triples, containing two objects and a link.

Wikidata is a collaborative KB, freely available online[2]. Given an object, Wikidata provides a list of relationships, called statements, involving that object. *Statements* are relations between two objects in the KB given a property. As Wikidata provides multilingual entries for each object, it is possible to use it in a multilingual system.

A query to this service consists of a term and a language. Optionally, the service also accepts a list of properties and a parent object filter. By default, the list of statements properties by the service is '*subclass of*', '*instance of*', '*member of*' and '*part of*' . The service calls the Wikidata *item disambiguation search*[3] in order to list all entries whose *title*, *pseudonyms* or *acronyms* match a given term in the KB. Then, the list of results is filtered for results in the requested language, and only one of the results is returned.

This query service provides 2 heuristics to choose the result of the query:

1. Lowest index first: each entry has a unique numeric index. It returns the entry with the lowest index, for a given language;
2. Filter by a parent: searches for results that have the same entry as a parent object. This heuristic is used as a strategy to filter ambiguous entries using a 'superclass', e.g., "*Paris*" can refer multiple entities such as the capital of France, a mythologic entity or a music album. In that case, the

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

[3] URL: https://www.wikidata.org/wiki/Special:ItemDisambiguation

superclass can help to disambiguate each term. If more than one result is still available, the service chooses the entry with the lowest index (heuristic 1).

If no result is found for that language, the service returns the first result for English language. An example of the result of a query to Wikidata about the entry 'Portugal' is shown in Figure 3.2.

```
['Q45', 'member state of the European Union', 'sovereign state', 'country', 'member
state of the United Nations']
['Q1045833', 'asteroid', 'JPL Small-Body Database']
['Q14110517', 'family name', 'Toponymic surname', 'Jewish name']
```

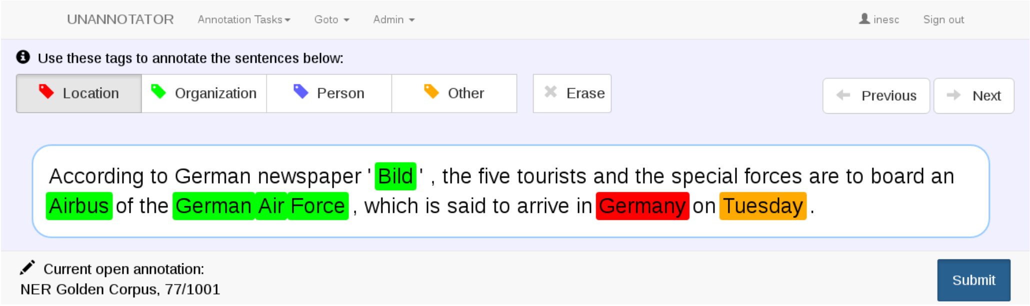
Figure 3.2: The result of querying Wikidata for the term *'Portugal'* contains three objects with the same name and identified with a unique id number: a country, an asteroid and a surname. The query returns a list of statements *'instance of'* for each object.

A query to Wikidata returns: (i) a unique index of the term (e.g. Q45 for the term *"Portugal"* as a country), (ii) a list of statements for a given list of properties, and (iii) a lists translations of the term in other languages, if available.

The results of the queries to Wikidata are stored in the local KB, avoiding multiple queries for the same term and unnecessary web traffic.

3.3.3 Unannotator Annotation Tool

In order to speed up the process of annotation, we developed an application for touch screen devices, called Unannotator, to assist in the manual annotation task. This application was developed in Python [43] using a local PostgreSQL [42] database to store the annotations. The application runs as a web server and is accessed through a web browser. It can be used by multiple users at the same time.



Use these tags to annotate the sentences below:

According to German newspaper ' Bild ' , the five tourists and the special forces are to board an Airbus of the German Air Force , which is said to arrive in Germany on Tuesday .

Current open annotation:
NER Golden Corpus, 77/1001

Figure 3.3: Annotation tool interface in a NE annotation task. One sentence is shown at a time. Each class of NE is highlighted with a different color.

In Unannotator, an annotation task consists of an annotation tagset and a corpus to annotate. The corpus is composed of sentences or documents. Each sentence or document is shown to the annotator, one at a time. An annotation tagset is the set of tags that an annotator can use to classify each word from the sentences, e.g. NE classes, as presented in Figure 3.3. A pair of sentences from parallel corpora can also be annotated simultaneously.

This application considers 3 roles, which correspond to different types of users:

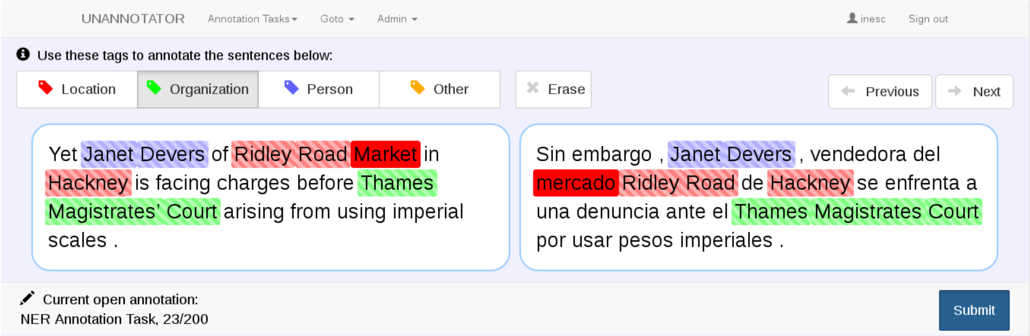
- **Administrators** - Users that are able to create new annotation tasks, import and export data, assign tasks to users, and annotate any data;
- **Annotators** - Users that have privileges to access tasks and annotate text;
- **Revisors** - Users that have privileges to access and change annotated data from other users, and rate their work.

In order to strengthen the use of an annotation guideline, a brief guideline is presented at the center of the interface when a user opens a task. Some textual usage hints are presented in the interface.

By default, there are four types of annotation task:

- **Simple annotation** - One sentence is presented to the user, which is suitable for annotation of POS or NEs. An example of this annotation task in presented in Figure 3.3;
- **Parallel corpora annotation** - Two aligned sentences are presented to the user at the same time, which is suitable for annotation of expressions in parallel corpora such as the annotation of NEs in parallel corpora. An example of this annotation task in presented in Figure 3.4;
- **Co-reference annotation** - The document is presented with highlighted NEs and the user must group the NEs that refer to the same object. An example of this annotation task in presented in Figure 3.5;
- **Evaluation** - A set of sentences is presented to the user to evaluate or mark errors for each sentence, which is suitable for evaluating translations and relevance of replacements of NEs. An example of this annotation task in presented in Figure 3.6.

Texts can be imported into the annotation task as a plain text file with no annotations, or a TSV file containing pairs of token-annotation. Files in the CoNLL format are also supported, although only the last annotation from each line is assigned to the token. The annotations are exported as a TSV file.



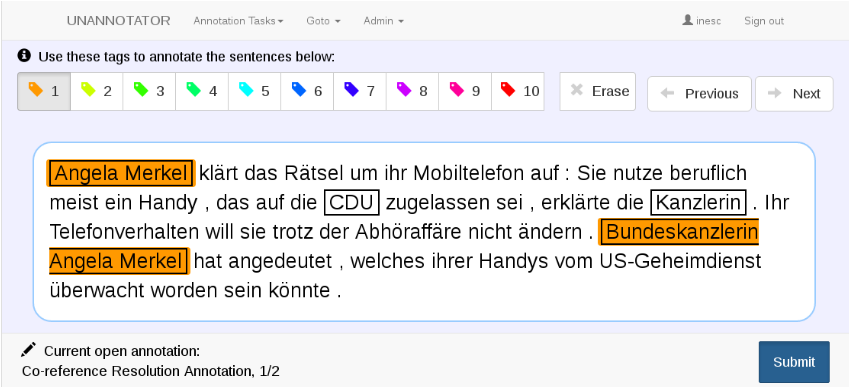
Use these tags to annotate the sentences below:

Yet Janet Devers of Ridley Road Market in Hackney is facing charges before Thames Magistrates' Court arising from using imperial scales .

Sin embargo , Janet Devers , vendedora del mercado Ridley Road de Hackney se enfrenta a una denuncia ante el Thames Magistrates Court por usar pesos imperiales .

Current open annotation:
NER Annotation Task, 23/200

Figure 3.4: Annotation tool interface featuring a pair of aligned sentences in English and Spanish. Each class of NE is highlighted with a different color.

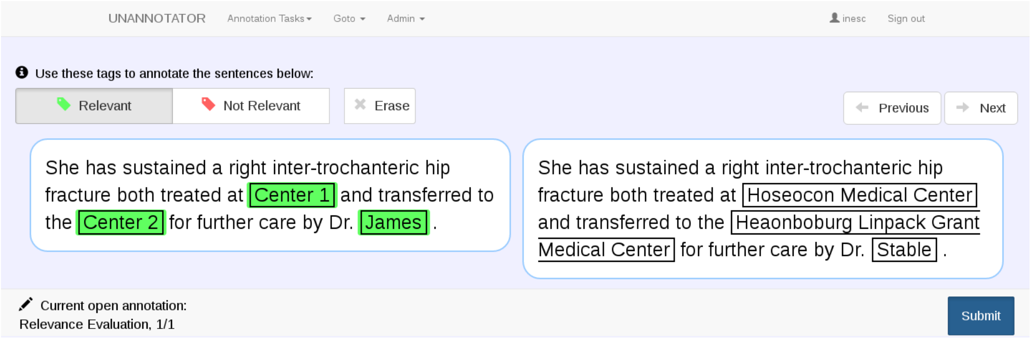


Use these tags to annotate the sentences below:

Angela Merkel klärt das Rätsel um ihr Mobiltelefon auf : Sie nutze beruflich meist ein Handy , das auf die CDU zugelassen sei , erklärte die Kanzlerin . Ihr Telefonverhalten will sie trotz der Abhöraffäre nicht ändern . Bundeskanzlerin Angela Merkel hat angedeutet , welches ihrer Handys vom US-Geheimdienst überwacht worden sein könnte .

Current open annotation:
Co-reference Resolution Annotation, 1/2

Figure 3.5: Annotation tool interface during a coreference annotation.



Use these tags to annotate the sentences below:

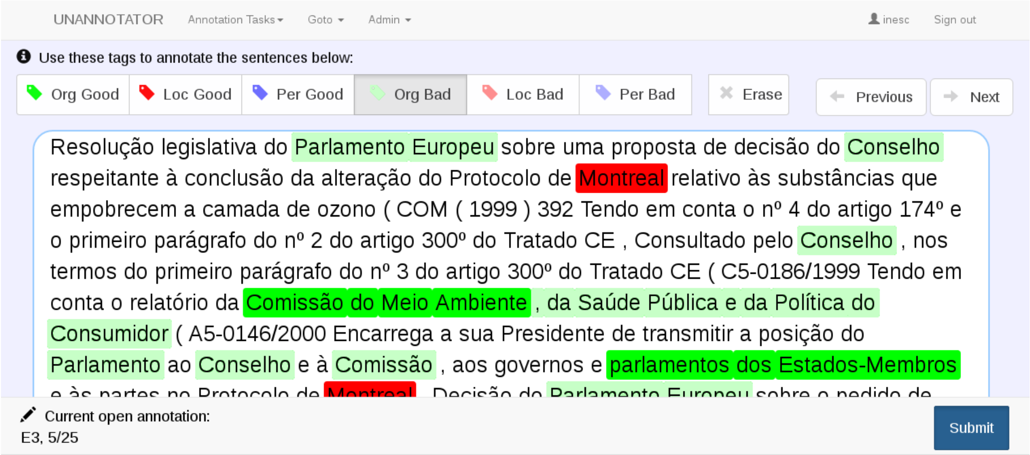
She has sustained a right inter-trochanteric hip fracture both treated at Center 1 and transferred to the Center 2 for further care by Dr. James .

She has sustained a right inter-trochanteric hip fracture both treated at Hoseocon Medical Center and transferred to the Heaonboburg Linpack Grant Medical Center for further care by Dr. Stable .

Current open annotation:
Relevance Evaluation, 1/1

Figure 3.6: Annotation tool interface during an evaluation of the relevance of substitutions.

From a user's perspective, a task of annotation consists of choosing a tag from a menu and then to press the tokens to annotate them. The annotation tags are usually based on a pallette of colors assigned to the tokens in the text, but numeric tags can be used as well. These pallettes of colors can be changed by the administrator in the case of colorblind annotators or limited representation of colors by the screen. The result of the annotation task can be exported as a tab-separated file.



Use these tags to annotate the sentences below:

Resolução legislativa do Parlamento Europeu sobre uma proposta de decisão do Conselho respeitante à conclusão da alteração do Protocolo de Montreal relativo às substâncias que empobrecem a camada de ozono (COM (1999) 392 Tendo em conta o nº 4 do artigo 174º e o primeiro parágrafo do nº 2 do artigo 300º do Tratado CE , Consultado pelo Conselho , nos termos do primeiro parágrafo do nº 3 do artigo 300º do Tratado CE (C5-0186/1999 Tendo em conta o relatório da Comissão do Meio Ambiente , da Saúde Pública e da Política do Consumidor (A5-0146/2000 Encarrega a sua Presidente de transmitir a posição do Parlamento ao Conselho e à Comissão , aos governos e parlamentos dos Estados-Membros e às partes no Protocolo de Montreal . Decisão do Parlamento Europeu sobre o pedido de

Current open annotation:
E3, 5/25

Figure 3.7: Comparing the result of the NER module with the golden corpus in a sentence from the Portuguese DCEP dataset.

This application was also used to visualize differences between two annotation files (Figure 3.7), for example, for comparing the annotation from two human evaluators or comparing the result from a module of our system with a golden corpus.

3.3.4 Extraction of Headwords

A *headword* is the major element of a noun phrase or compound-noun. Noun phrases can be represented as a graph of grammatical dependencies, having the headword as their center. This headword is a noun that determines the syntactic and semantic categories of the noun phrase (Zwicky [65]).

In our work, we often use headwords in order to characterize and represent NEs. The headwords of the NEs can be automatically provided by some external NER tools such as STRING (Mamede *et al.* [29]) for Portuguese texts. However, when using other NER tools, the system needs a function that extracts the headword of a NE for a given language. For that reason, we have implemented a simple headword algorithm that is able to deal with the four languages used in our system. The Figure 3.8 presents the pseudocode of the implementation of the headword extraction.

```
Use: hwlist                                              ▷ List of known headwords
function HeadwordExtraction(NE, lang)
    for token in NE do
        if lang GERMAN then
            token ← splitting(token)
        end if
        if token ∈ hwlist(lang) then
            return token
        end if
        if lang ∈ (GERMAN, ENGLISH) then
            if next(token) ∈ PREPOSITION then            ▷ Prepositional expressions
                return token
            else if previous(token) ∈ GENITIVE then         ▷ Genitive expressions
                return token
            end if
        end if
    end for
    if lang ∈ (GERMAN, ENGLISH) then
        return last(NE)
    else if lang ∈ (PORTUGUESE, SPANISH) then
        return first(NE)
    end if
end function
```

Figure 3.8: Implementation pseudocode of our multilingue headword extraction algorithm.

Language	**Named Entity**	**Headword**	**Heuristic**
Spanish	**Museo** de Historia de Barcelona	Museo	First word
English	Atlantic **Airlines**	Airlines	Last word
English	French **Academy** of Sciences	Academy	Prepositional
German	Botanischer **Garten** Dresden	Garten	Genitive flexion
German	Beobachter**mission** der Vereinten Nationen	Mission	Prepositional + Splitting

Table 3.7: Examples of application of the heuristics for headword extraction in different languages.

Our implementation uses a set of heuristics to extract a headword candidate. We have implemented three heuristics for NEs in English and German texts: (i) the first token followed by a preposition is a headword; (ii) the first token preceded by a noun or adjective in the genitive case; or, (iii) the last token from the NE. We have implemented one heuristic for NEs in Portuguese and Spanish: the first token from the NE. The NEs in German are split before being processed by the heuristics. The splitting process is performed by jWordSplitter [35], an open-source splitter for German compound words, that returns an array of tokens. We consider the last token of the array as the main part of the compound word. We created and evaluated this set of heuristics based on a list of NEs extracted from Wikidata (Vrandečić [61]). Table 3.7 lists examples for all the heuristics implemented in our algorithm.

4 Text Anonymization

"It's a beautiful thing, the destruction of words. Of course the great wastage is in the verbs and adjectives, but there are hundreds of nouns that can be got rid of as well."

George Orwell, *Nineteen Eighty-Four* (1949)

THE TEXT ANONYMIZATION SYSTEM is responsible for detecting sensitive information in a text document and applying a method of anonymization in order to mischaracterize that information. This chapter describes the structure of the anonymization system implemented, together with the anonymization methods, giving reasons for our decisions.

In Section 4.1, we set forth the architecture of our anonymization system and Section 4.2 details the implementation of each module. In Section 4.3, we describe four anonymization methods implemented in our work. Section 4.4 provides information about the data flow inside our system.

4.1 Architecture

Our anonymization system is based on the modular monolingual system by Dias *et al*. [12]. The modules of the original system were improved in order to implement a multilingual system and to solve some major issues reported by the authors. This implementation presented some challenges.

The first challenge was to extend a monolingual system into a multilingual system. As the original system was composed of modules that were specific to only one language, the system should be well structured and flexible enough in order to allow modules to be exchanged.

The second challenge is the major concern of every anonymization system: to raise the performance of the detection of sensitive information. For that reason, we implemented some techniques used in previous studies and evaluated the contribution of each technique to the performance of the detection.

Our system is composed of a pipeline with five modules: (i) Pre-processing, (ii) NER, (iii) Second-pass detection, (iv) Coreference Resolution (CRR), and (v) Anonymization. Each of the modules was implemented independently and the information flow along the modules is presented in Figure 4.1.

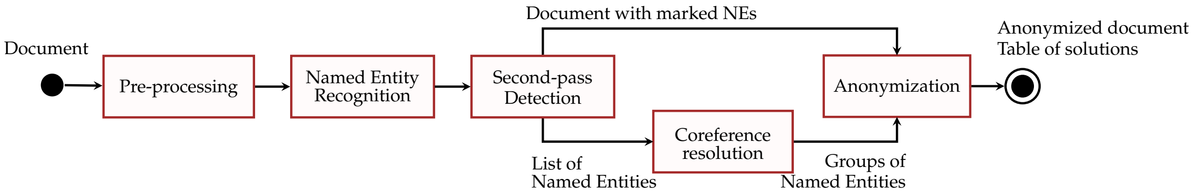


Figure 4.1: Flowchart of the anonymization pipeline implemented in our work.

This pipeline receives a text document and returns an anonymized version of the same document and the respective *table of solutions*. The *table of solutions* contains all the replacements made in the document and their positions inside the text.

4.1.1 Services

The modules of our system are divided into two main services: NER and Anonymization. Each service groups together all the implementations needed to provide one specific task from our system. These services are arranged in a modular structure. This modular arrangement facilitates the subsequent modifications to the system, such as to add a new anonymization method, or to recognize entities from a new language, by simply changing the modules of the pipeline. This arrangement into independent services also makes it possible to provide both services as web services. The UML [39] models of both services are presented in the Appendix A.

In this work, we use NER techniques in order to detect sensitive information. This decision is based on the following reasons:

- NEs contain direct identifiers with the most probable data to contain private information;
- NER tools are available for several languages, and some of these tools are based on machine learning methods, which makes them easier to train;
- Previous works in text anonymization based on NER techniques achieved promising results (Wellner *et al.* [62], Uzuner *et al.* [60], Yang & Garibaldi [64], among others);
- The module of entity detection is completely independent of the anonymization process and can be later replaced by another tool, keeping the modular structure of the pipeline;
- NER classifiers provide information about the entity category which is useful for some of the anonymization methods that were implemented (see Section 4.3).

We chose to implement a hybrid approach in our system because hybrid anonymization systems produced better results in most recent works (Uzuner *et al.* [60], Yang & Garibaldi [64], among others).

4.1.2 Dealing with multiple languages and configurations

In order to process texts in different languages, the structure of the anonymization system must be flexible and must change according to the language. Another major concern is raised by the fact that modules from different languages should use the same standards for data representation, e.g. all NER modules should use the same tagset.

The structure of the system is defined in a configuration file. This *configuration file* is a JSON [9] file containing the options for the all the modules of the system. The root of this file is divided into language keys and each key defines the configuration of modules of the system for a given language. The following list shows some of the options that can be defined in the configuration file:

- Switch modules on or off, e.g. the use of CRR or Second-pass detection (Figure 4.1);
- Combine different NER tools;
- Define a conversion table between the tagset of a NER tool and the standard tagset of the system;
- Assign the classes of sensitive information that should be processed by the system;
- Assign the anonymization methods to be applied to each class of sensitive information;
- Assign the classes that implement each method;
- Define other options comprising each specific module of the method, e.g. the default options for each anonymization method.

The full list of configurations is presented in Appendix B.

4.2 Modules

4.2.1 Pre-processing

The module of pre-processing prepares the text in order to be processed by the pipeline, which consists in the normalization of the text and extraction of features. The normalization changes the character encoding to UTF-8 and tokenizes the text using a tokenization tool provided by Moses (Koehn *et al*. [25]). The extraction of features consists of determining the morphosyntactic features (MSF)s (such as number, gender, and case) at word level to be used by subsequent modules.

4.2.2 Named Entity Recognition

The function of the NER module is to detect candidate expressions that convey sensitive information within the text. This module receives a normalized text and returns a list with the NEs contained in the text. Each NE is identified by its surface form, position within the text, and class.

In the monolingual system by Dias *et al.* [12], the authors suggested that some issues could be solved by increasing the recall of the NER classifier in different domains of text. In that direction, we propose a new design for the NER module, composed of a set of NER classifiers in parallel and a voting component.

In our system, the task of NER is performed in two stages as illustrated in Figure 4.2: (i) a set of NER classifiers in parallel; and (ii) a voting stage that receives the output from the NER classifiers and votes the classifications. The function of the model-based NER tool is to detect and classify common NEs by using already available tools. The function of the pattern-matching component is to detect more specific and rare patterns of entities.

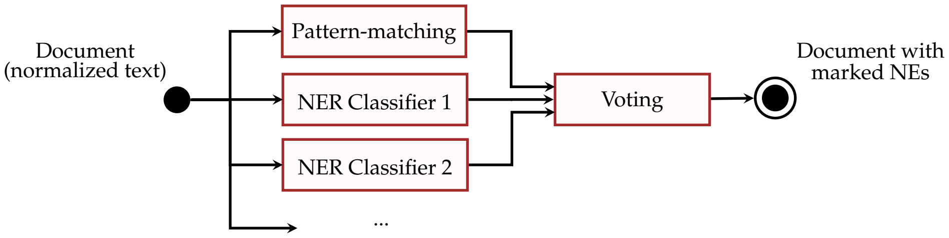


Figure 4.2: Flowchart of the NER module.

Pattern-matching

The pattern-matching component detects tokens or sequences of tokens that match a fixed pattern. The use of pattern-matching can be extended by the user to any entity classes with a well-defined pattern. Examples of these classes can be country names, time expressions, or identification numbers. In this component, patterns can be described by gazetteers or regular expressions. These patterns are stored in comma-separated files, containing a pattern and the respective NE class.

Entries from the gazetteer and regular expressions can also be used to check for false positive detections. By default, any NE contained inside a previously detected NE is not labeled by this component.

NER Classifiers

In this work we use two main NER classifiers depending on the language:

- The Stanford CRF classifier (Finkel *et al.* [18]), also known as Stanford NER[1], which is an open-source implementation of the CRF algorithm (Lafferty *et al.* [26]). This classifier is used for NER in texts in the following languages: English, Spanish, and German. This CRF classifier uses the following models trained in NE-annotated corpora for each language:

[1] Available at http://nlp.stanford.edu/software/CRF-NER.shtml

→ English, trained with the CoNLL-2003 English training datasets [56].

→ German, trained with the CoNLL-2003 German training datasets [56] extended by Faruqui & Padó [16].

→ Spanish, trained with AnCora dataset (Taulé *et al.* [53]).

- The STRING chain (Mamede *et al.* [29]), a hybrid, statistical and rule-based NLP chain for Portuguese. This tool is hosted at INESC-ID L2F[2] and it is able to recognize named entities in Portuguese texts. Although STRING features a larger tagset of NE classes with high granularity, our implementation uses only the three classes of NEs from MUC-6 [7] (person, location, and organization).

Along with the main NER classifier, more classifiers can be combined in parallel, as in the works of Uzuner *et al.* [60] and Dehghan *et al.* [11]. These classifiers can be used to improve the performance of the main NER classifier or to detected new classes of entities. In the cases where new models have to be trained, we chose to use the Stanford CRF classifier because:

- The CRF algorithm provides state-of-the-art results on entity detection (Yang & Garibaldi [64], Liu *et al.* [28], Dehghan *et al.* [11], among others);
- The Stanford CRF implementation made by Finkel *et al.* [18] is open source;
- The classifier provides a feature factory and is reasonably well-documented;
- The classifier is bundled with NER models trained for some languages (including English, German and Spanish);
- It is possible to train new models for the classifier.

```
"nerservice": "local",
"localner": {
  "model": "dewac.ser.gz",
  "classifier": "stanford-ner.jar"
},
"parallel": {
  "state": 1,
  "paralleltags": ["ORGZ", "PERZ", "LOCZ"],
  "classifiers":[
    {
      "url": "http://localhost:1025/ner",
      "classifier": "dcep-deutsch"
    }
  ]
}
```

Figure 4.3: Snippet of the configuration file containing the registration of a local NER classifier (above) and a remote classifier in parallel (below).

[2] INESC ID Lisboa, Spoken Language Systems Laboratory; URL: https://www.l2f.inesc-id.pt

Voting

The voting stage receives a list of classifications for each token, as classified by the pattern-matching and NER tools, and returns one classification. This stage was implemented using a template-method pattern and can be extended in the future. The default voting action classifies tokens in cascade, sorting the NER classifiers in a cascade of precedences as defined in the configuration file. In the case of ambiguous classification, it is chosen the classification from the first NER classifier in the cascade of precedences.

4.2.3 Second-pass Detection

The function of the *second-pass* detection module is to improve the performance of the NER module by applying a post-processing stage to the matched tokens and handling unmatched tokens. Previous works (Arakami *et al.* [1], Szarvas *et al.* [52] and Yang & Garibaldi [64]) used a post-processing stage in order to mitigate errors of detection and classification. As in our work the performance of the anonymization relies on NER, it is also important to add a module that corrects the results of NER and generates additional candidates.

Our implementation of the second-pass is based on the idea introduced by Arakami *et al.* [1] that, sometimes, a NE appears twice or more within a document, but a NER classifier could not be able to classify uniformly all the instances of that NE in different contexts. The second-pass detection consists of two stages:

- **Short form detection** - Based on the idea of the generation of coreferences in the work of Yang & Garibaldi [64]; a NE can present a short form (e.g. "*Parliament*" is the short form of the NE "*European Parliament*") which might not be detected by the NER classifier; we implemented a solution that creates a list of NE headwords detected by NER and look them up in the document;
- **Label-consistency** - Based on the idea of *label-consistency* from the work of Arakami *et al.* [1]; depending on the context, two appearances of a NE might be classified with different labels; we implemented *label-consistency* based on the majority label assigned to a NE within the document.

4.2.4 Coreference Resolution

A *coreference* consists of two linguistic expressions that refer the same extra-linguistic object. The Coreference Resolution (CRR) groups together all expressions that refer to the same extra-linguistic object, within a document .

During an anonymization process, some methods replace the NEs different expressions, causing the text to lose its specificity relatively to the original entities. As NEs within a text refer to the entities of interest, it is important to ensure that coreference is kept in order to maintain the meaning of the text. In order to preserve the context of the original document, it is important to replace all the occurrences of NEs that refer to the same extra-linguistic object by the same expression. Previous works, such as Gardner *et al.* [21] and Yang & Garibaldi [64], also took advantage of a CRR.

In this work, we aimed exclusively at the coreference between NEs. Although some CRR software already exists, we chose to implement a simple module from scratch because that CRR software is either not free, not multilingual or perform complex anaphoric resolution that is not required in this case.

This module receives a list of NEs and returns a *coreference chain* containing the NEs grouped by *mentions*. A mention is a group of entities that refer to the same extra-linguistic object. Our approach is similar to Bontcheva *et al.* [5], a rule-based tool. Some rules were extended in order to handle several languages and to utilize the output of the Second-pass Detection module to improve the coreference. This module uses the following rules for coreferences:

- Any entities matching the same surface;
- Any entities matching the headword stems and matching the same surface in the rest of the words (e.g. "*Europäische Parlament*" and "*Europäische Parlaments*"); stemming is only utilized in German documents and is performed by the Snowball German stemming algorithm (Porter [41]);
- Any entities whose surfaces differ only by a title (e.g. "*Mr. John Doe*" and "*John Doe*");
- Abbreviations and acronyms of the full expression of an entity (e.g. "*Europäische Parlament*", "*EP*" or "*E.P.*"). Abbreviations are only grouped when there is only one possible full expression for all abbreviations, otherwise, it is considered ambiguous and grouped in distinct mentions;
- Surfaces only differing by a term designating the extra-linguistic type of object (e.g. "*International Business Machines Corporation*" and "*International Business Machines*");
- Partial matching of the surface of the entities, including titles and abbreviations, in the cases there are no ambiguities (e.g. "*John Doe*", "*Mr. Doe*", "*John D.*", and "*John*");
- Possessive forms in English (e.g. "*John Doe*" and "*John Doe's*");
- Prepositional expressions (e.g. "*University of Cambridge*" and "*Cambridge University*").

The CRR module was implemented using a template-method pattern in order to provide different methods of coreference depending on the language and NE class. This module is composed of the abstract class `Coreferencer` that provides the basic methods for CRR and a set of classes that extend this abstract class. Each extended class implements a variation of the CRR method targeting a specific NE class and language (Appendix A). The configuration file of the CRR module assigns which implementation must be called for performing CRR over a NE class.

By default, the CRR is only performed over NEs whose classes are assigned in the configuration file. All other NEs are considered grouped in singleton mentions. Figure 4.4 shows a snippet of an example configuration of classes for the CRR in English.

```
"en": {
    "coreference": {
        "PERZ": "basicpercoref.EnglishPERCoref",
        "ORGZ": "basicorgcoref.EnglishORGCoref",
        "LOCZ": "basicloccoref.EnglishLOCCoref"
    }
}
```

Figure 4.4: Snippet of the configuration file showing the assignment of the coreference module classes.

4.2.5 Anonymization

The anonymization module applies an anonymization method to the entities detected by the previous modules. 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 entities, their replacements, and positions in the document.

This module applies an anonymization method depending on the class of the entity. These methods are defined in the *configuration file* of the system. By default, the anonymization process assigns a unique replacement expression to each mention by using the coreference chains of entities. This means that all the entities that mention the same extra-linguistic object, even if the entities present different surfaces, are replaced by the same expression. If no coreference chains are provided to this module, it assigns different replacements to each mention.

In order to implement some methods of anonymization, the anonymization module also needs access to external dictionaries of entities and knowledge bases. The methods implemented in our work are presented in the next section.

4.3 Anonymization Methods

Different methods of anonymization can be applied to text depending on the purpose of the anonymization. The anonymization methods can be divided into 3 groups: suppression, tagging, and substitution.

Suppression is a simple way of anonymizing a text that consists in the suppression of the NE using a neutral indicator that replaces the original text, e.g. ‘XXXXXX’. Most of the text anonymization systems implement this method.

Tagging consists of the replacement of the NE by a label that could indicate its class or a unique identifier. All the entities that refer to the same object within the document can be replaced by the same label, as provided by the output of the CRR module. In our work, one of the implementations of this method was made by concatenating the class given by the NER tool and a unique numeric identifier, e.g. `[**Organization123**]`.

Substitution is the replacement of a NE by another entity. There are several implementations for this replacement. Some possible implementations are: *random substitution* by adding a random entity from a dictionary, *swapping* NEs inside a document, *intelligent substitution* by an entity with the same features of the original NE (eg. morphosyntactic features), and *generalization* by an entity from the same class.

In this system, we have implemented four methods of anonymization: (i) suppression, (ii) tagging, (iii) random substitution, and (iv) generalization. The following sections describe the two substitution methods implemented in this work.

4.3.1 Random substitution

Random substitution replaces an NE by another random NE from the same class and MSF. This method was implemented using a *default list* containing random entities of each class for each language. The MSFs of an NE are determined by the NE headword. The headword is the major word from a NE. These features are determined by querying the dictionary for the NE headword. If no entity is found in the dictionary, some default MSFs are assumed for the entity (masculine, singular and nominative case). Given these MSFs, an entry of the same class and same features is looked up in the *default list* of entities. If no suitable entity is found, the tagging method is used instead. The headword is extracted from a NE using the implementation presented in the Section 3.3.4 on page 33.

Determining the grammatical gender is important for replacing NEs that refer to persons by another NE of the same gender, e.g. replacing *John* by *Peter*, or replacing *Mary* by *Anna*. In some highly inflected languages, all nouns, including proper names have MSFs, and determiners or modifiers must show agreement with them (an example is depicted in Figure 4.5).

Der US-Astronaut Scott Kelly kehrt nach 340 Tagen auf die Erde zurück.

(Translation: the US astronaut Scott Kelly returns back to Earth after 340 days.)

Figure 4.5: Sentence[3] in German showing the agreement between the entities (underlined) and respective determiners. The predicate 'kehrt ~ zurück' also agrees with the grammatical number of the subject 'Scott Kelly'.

[3] Extracted from `http://www.welt.de/wissenschaft/article152792060/`

In order to maintain the natural grammatical concordance of a text after the replacement of the entities, it is important to replace an entity by another of the same class and the same MSFs. Our implementation uses two sources of information to determine the MSFs of a NE depending on the language:

- **STRING** - When processing documents in Portuguese, the STRING chain is able to provide the MSFs of a NE;
- **Wiktionary Query Service** (Section 3.3.1) - This service queries for a word in Wiktionary [63] and if the word exists in the dictionary, it returns some information about the word such as the POS, the grammatical number and gender, among others.

The entries with the same MSFs are then looked up in the *default list*. One of these entries is chosen randomly to be the substitution entry. Figure 4.6 presents the pseudocode of the implementation and Table 4.1 shows some examples of replacements based on the MSFs of the entities.

```
Use: default_list                                   ▷ List with default entities by class
Use: dictionary
function RandomSubstitution(entity, class, language)
    hw ← headword(entity)                           ▷ Gets the head of the NE
    features ← query-dictionary(hw, language, dictionary)
    if features = ∅ then
        features ← DEFAULT                          ▷ it uses default grammatical features
    end if
    if ⟨class, features⟩ ∈ default_list then
        return lookup-list(class, features, default_list)
    else
        return tagging(class)                       ▷ Uses tagging method
    end if
end function
```

Figure 4.6: Implementation pseudocode of random substitution method.
The function receives as arguments an entity, the entity's class and the language code.

In order to avoid two different entities being replaced by the same random entry, each term is checked for collision with the surface form of other entries. As an extra precaution against duplicates with entities of the document, each term is also searched in the body of the document for any expression with the same surface form. In the case of collision, the method looks up another entry from the *default list*.

The *default list* contains tuples of entries with a term representing a NE, its language, NE class, and MSFs (we use the number, gender, and case). Table 4.2 exemplifies the structure of the *defaults list*.

In this table, we are able to notice that, for some languages like German, there is a wide range of variations in MSFs of the entities, meanwhile in English, the entities vary only in grammatical number. In this

Original entity		**Replacement**
Aeroporto (airport) *masculine, singular*	➤	**Recinto** (venue) *masculine, singular*
Frankfurts (from Frankfurt) *masculine, singular, genitive*	➤	**Wegs** (from the way) *masculine, singular, genitive*

Table 4.1: Example of replacements of entities, in Portuguese and German, by others with the same morphosyntactic features.

Language	**Class**	**Number**	**Gender**	**Case**	**Term**
Portuguese	*location*	*singular*	*masculine*	-	'recinto'
Spanish	*location*	*singular*	*feminine*	-	'arena'
English	*location*	*singular*	-	-	'venue'
German	*location*	*singular*	*neuter*	*nominative*	'Wahrzeichen'

Table 4.2: Example of a *default list* with entries in distinct languages.

example (Table 4.2), we display only the nominative case of "*Wahreichen*" (landmark) that is ambiguous with its accusative and dative forms.

4.3.2 Generalization

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

In order to implement the method of generalization, this module needs to access a KB in order to retrieve the superclasses of a given entity. The current implementation uses Wikidata (Vrandečić [61]) as a KB. Wikidata is a free collaborative KB, available online[4]. Our implementation uses the Wikidata Query Service (see Section 3.3.2) to query the KB.

By querying Wikidata for an entity, it is possible to retrieve a list of entity statements that may include: superclasses, pseudonyms, acronyms, and membership, among others. Names of superclasses can be used as a generalization of an entity (example in Figure 4.7). The superclass of a given entity is determined by the entry pointed by its statement '*subclass of*'. When not available, the classes pointed by the statements '*instance of*', '*member of*' or '*part of*' are used as a superclass. Entities with the class `Person` are not generalized. Instead, the system uses the random substitution method presented above (4.3.1).

Given an entity, the method looks up for any entry in the KB whose name, pseudonym or acronym matches the entity. Then, it filters these entries by the class of the entity, and checks if the class is a

[4] URL: https://www.wikidata.org

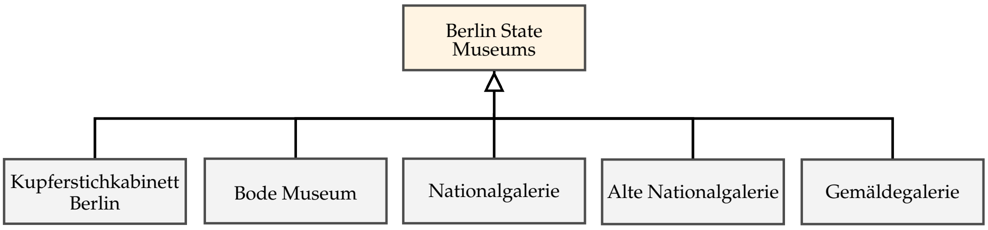


Figure 4.7: Example of hierarchy in Wikidata, with 5 museum entities linked to 'Berlin State Museums'. In this case, 'Berlin State Museums' can be used as generalization of any other 5 entities.

```
Use: KB                                                        ▷ Knowledge base
function Generalize(entity, category, lvl)
    if category = PERSON then
        return RandomSubstitution(entity, category)
    else
        $e \leftarrow$ query-KB(entity, category, KB)
        if $e = \emptyset$ then                          ▷ Then uses only the headword
            $h \leftarrow$ headword(e)
            $e \leftarrow$ query-KB(h, category, KB)
        end if
        if $e \neq \emptyset$ then
            $root \leftarrow$ root-KB(category, KB)
            while $e \neq root$ and anon-measure(e, KB) $<$ lvl do
                $e \leftarrow$ get-parent-entry(e, KB)
            end while
            return e
        else
            return RandomSubstitution(entity, category)
        end if
    end if
end function
```

Figure 4.8: Implementation pseudocode of generalization method. The function receives as arguments an entity, its category and the desired level of anonymity.

superclass of the current entry. In order to guarantee a given level of anonymization, the implementation is able to choose a superclass that satisfies an anonymity measure, e.g. a minimum number of child entities of the current generalization.

Wikidata provides multilingual statements for each entity, making it possible to use it in several languages. Figure 4.8 presents the pseudocode of the implementation.

By default, the generalization method is not applied to entities that represent persons, being used the random substitution method instead.

In order to avoid collision between the surface form of two entities or expressions in the text, each entry is checked for collision with other entries and searched for duplicates in the text of the document. In the case of collision, the method uses the random substitution method (4.3.1) to choose another substitution.

4.4 Flow of Data

The data flow in our pipeline consists of a JSON object [9], hereinafter referred to as *job*, with a fixed structure that is transported along the pipeline. The fixed structure of this job makes it possible to exchange the modules of the pipeline without the need of creating data interfaces between the new modules. Each module receives and processes the data of this object. In the end, the module returns this same object combined with the results.

At the beginning of the pipeline, the job has its simplest format containing the following information:

- A string containing the text to be anonymized.
- The language of the text.
- An anonymization method to be applied to the text.
- A pair of delimiter strings to the anonymized information.
- A configuration file.

More detailed description about the JSON job is provided in the Section B.2 of Appendix B.

4.5 Implementing new Anonymization Methods

New anonymization methods can be added to the system by providing an implementation in a class that extends the abstract class `AnonimizationMethod`. The new anonymization method receives a job and a language and returns the same job with a new key "anon" containing the result of the anonymization method. The UML diagram in Appendix A shows the method signature for this class and the relationships with other classes of the anonymization package.

In order to make the method available in the system, the new anonymization method must be registered in the configuration file by creating a new key with the label of the method. Figure 4.9 shows an example of the registration of a new anonymization method.

The JSON schemas of the configuration file are presented in the Section B.1 of Appendix B.

```
"METHOD_LABEL": {
    "name": "New Anonymization Method",
    "class": "newmethod.NewAnonymizationMethod",
    "options": "..."
}
```

Figure 4.9: Registering a new anonymization method in the configuration file.

4.6 Summary

This chapter provides a detailed description about the architecture of the text anonymization system. The objective of this system is to detect and remove sensitive information from a text using some anonymization methods. The sensitive information is detected using one or more NER classifiers. This anonymization system was implemented in a modular structure. All modules from this system can be replaced or extended in order to implement new functionalities in the future.

5 Evaluation

"The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents."

H. P. Lovecraft, *The Call of Cthulhu* (1928)

THIS CHAPTER presents the procedures and results of the evaluation of the text anonymization system. This evaluation aims to study the system's performance when using different combinations of modules and anonymization methods, which were presented in the previous chapter, and also to compare the performance in the detection of sensitive information using datasets from the previous de-identification shared tasks.

The main objective of an anonymization system is to detect as much sensitive information as possible. For that reason, the evaluation of previous studies (Chapter 2) was directed at assessing the performance on the detection of sensitive information, in a similar way as in the evaluation of a NER tool. However, in our study, we also try to evaluate how readable and adquate seems an anonymized text. In order to approach a measure to the readability and adquacy of a text, we asked human readers to rate the output of the system.

In Section 5.1 we describe the experimental setup of the evaluation. Section 5.2 presents the main results, which will be discussed in Section 5.5. Moreover, the results of the evaluation using the i2b2 de-identification shared tasks datasets (Stubbs *et al.* [50] and Uzuner *et al.* [59]) are presented in Section 5.4. Finally, Section 5.3 presents a brief evaluation of the time complexity of the system.

Due to space limitations, some of the results and examples used during the discussion are presented in the appendixes of this dissertation .

5.1 Experimental Setup

The evaluation of the anonymization system consisted in an intrinsic evaluation of the performance of the NER, CRR and anonymization modules. With this evaluation we intend to measure: (i) how much sensitive information is detected by the system (Section 5.2.1), (ii) how the system keeps track of references to entities referred by the sensitive information in an anonymized text (Section 5.2.2), and (iii) how readable and adquate seems an anonymized text to a human reader (Section 5.2.3).

We evaluate the performance of the system using two golden-standard datasets per language in order to study the performance over different text domains and different writing styles. These datasets are detailed in the Section 3.2.

The performance of NER was evaluated based on the metrics of precision, recall and f1-score. The evaluation was made at token-level (Section 3.1.1) in order to measure the overall performance on the detection. The same evaluation was also made at instance-level (Section 3.1.2) in order to study the performance of the detection of entities on their full extent.

We evaluated intrinsically the NER module by feeding the module with the unannotated text from the golden-standard datasets and comparing the result with the annotations of the golden-standard. Six different experimental setups were used in order to compare the performance of the NER task separately in different configurations. We use the f1-score to compare performances. However, we prioritize the recall over the precision as the recall measures how much sensitive information is detected by the system. We performed statistical analysis (Section 3.1.6) of the differences between the performance of the baseline and each configuration, using the approximate randomization method (Noreen [37]) and a significance level $p < 0.01$. We present the results for the following configurations:

1. **Baseline** - The configuration of the system from Dias *et al.* [12] was the baseline to evaluate our system, consisting of a NER classifier (Section 4.2.2), without pattern-matching and without second-pass detection.
2. **Baseline + Pattern-matching** - The baseline system combined with a gazetteer. We used the following gazetteers for each language:

   → The gazetteers from CoNLL 2003 shared tasks (Kim & De Meulder [56]) were used for English and German documents. The English gazetteer contains 7,345 entries and the German gazetteer contains 5,066 entries.

   → Repentino (Sarmento *et al.* [48]) gazetteer was used for Portuguese documents. Ths gazetteer contains 380,176 entries.

   → A gazetteer from Wikiner (Nothman *et al.* [38]) was used for Spanish documents. This gazetteer contains 2,807 entries.
3. **Baseline + Second-pass Detection** - The baseline system combined the Second-pass Detection module (Section 4.2.3).
4. **Baseline + NER classifier in parallel** - The baseline's NER tool in parallel with a secondary NER classifier trained with silver-standard corpora based on DCEP and WMT-News datasets with annotated NEs (Section 3.2.2).
5. **Baseline + All the previous configurations** - The baseline system combined with all the modules: Pattern-matching, Second-pass Detection and secondary NER classifier in parallel.

We evaluated intrinsically the CRR module performance by feeding the module with lists of entities from the golden-standard datasets and comparing the result with the golden-standard chains of entities. The performance of CRR was evaluated based on the metrics of B^3-score of the coreference chains (Section 3.1.3).

We evaluated intrinsically the performance of the anonymization based on substitution methods by feeding the anonymization module with lists of entities from the golden-standard datasets and rating the results manually. The evaluation was performed by human readers, as presented at the Table 3.4 on page 26. The performance of the anonymization was evaluated based on the availability and relevance of the substitution performed by the anonymization method (Section 3.1.4).

Some of the discussions of results are supported by the list of results in appendix. The complete list of results in terms of precision, recall and f1-score for each configuration and averaging strategy can be consulted in the Appendix C. Appendix D lists some outputs of the anonymization system for texts in each language.

5.2 Results

5.2.1 Named Entity Recognition

This section summarizes the results of the evaluation of the NER module using comparative graphs of the performance of each combination. The following convention for the abbreviations used in the labels of the graphs:

BL - Baseline
BL+PM - Baseline combined with Pattern-matching
BL+2P - Baseline combined with the Second-pass Decision module
BL+2NER - Baseline combined with a second NER classifier in parallel
BL+All - Baseline combined with all the previous modules

For each dataset, we present the macro-averaged token-level overall performance considering a 3-class tagset and the significance test for each metric. We also compare the performance (f1-score) using three different averaging strategies in order to understand how the NER module performes in both document-level and dataset-level, and at entity-level detection.

5.2.1.1 English

The following graphs present the performance of the NER in English documents in 6 different configurations for each dataset: CoNLL 2003 dataset in Figure 5.14 and DCEP dataset in Figure 5.2. Figure 5.3 shows the performance of NER for different averaging strategies. Table 5.1 presents the significance tests between the performance of each combination and the baseline.

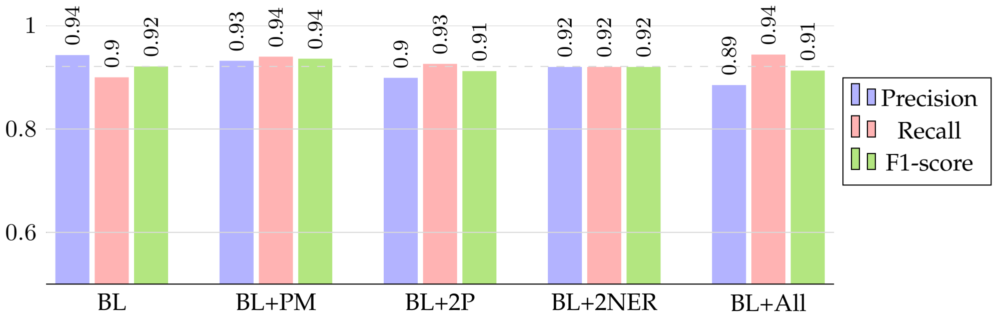


Figure 5.1: Performance of the macro-averaged token-based classification by NER module in the English CoNLL 2003 dataset.

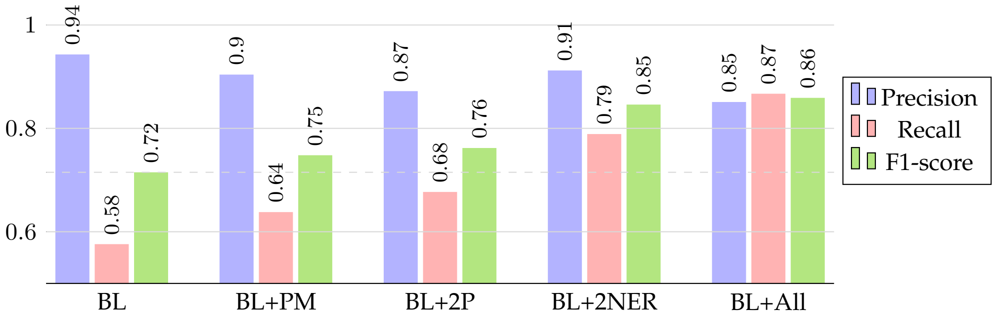


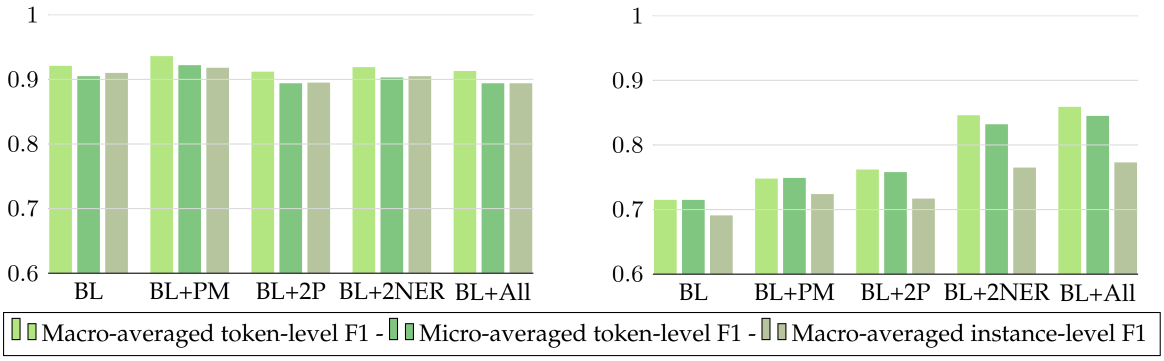
Figure 5.2: Performance of the macro-averaged token-based classification by NER module in the English DCEP dataset.

Dataset	**BL+PM**	**BL+2P**	**BL+2NER**	**BL+All**
CoNLL	F	-	-	-
DCEP	F	F	R, F	R, F

Table 5.1: Significance tests for the performance of each configuration against the baseline, considering $\alpha < 0.1$ (labels: P = precision, R = recall, F = f1). Labels mark only the performance metrics that are statistically different.

In average, the recall increased when modules are combined with the baseline configuration. However, the precision decreases which is caused by false-positives added by each configuration. The improvement of performance (f1-score) for combining modules was significant in the DCEP dataset (table 5.1).

- **Baseline** - The majority of false-positives occur on `organizations` entities (Table 5.2). Comparing with the performance on CoNLL dataset, this configuration achieves a lower recall in the DCEP dataset for all the classes, most notably a considerable amount of undetected `organizations` spe-



(a) Performance of NER in the CoNLL dataset.

(b) Performance of NER in the DCEP dataset.

Figure 5.3: Comparison between f1-score when using macro-averaged token-level, micro-averaged token-level and macro-averaged instance-level averaging strategies of the NER module in English datasets.

cific to the text domain of DCEP (Table 5.2) such as "*Committee on Foreign Affairs*").

		System output			
		LOC	PER	ORG	O
Golden Standard	LOC	387	1	6	18
	PER	5	472	5	17
	ORG	20	15	313	40
	O	18	2	13	7753

		System output			
		LOC	PER	ORG	O
Golden Standard	LOC	163	0	1	26
	PER	2	183	2	51
	ORG	24	0	544	428
	O	5	0	60	8995

Table 5.2: Confusion matrixes for the baseline configuration in CoNLL (left) and DCEP datasets (right).

- **Pattern-matching** - The use of a gazetteer increased significantly the f1-score in both datasets due to an improvement of the recall in almost all classes (Table C.1, page 96). The majority of false-positives in both datasets were caused by `organizations` (classified as `locations`), lowering the precision of the `organization` class.
- **Second-pass Detection** - This module increased significantly the f1-score in the DCEP dataset due to an improvement of the recall for all NE classes (Table C.1, page 96). Results from CoNLL dataset did not vary significantly from the baseline.
- **Second NER classifier in parallel** - The NER classifier in parallel increased significantly the performance in the DCEP dataset. Values of recall, and consequently the f1-score, were favored by the detection of `organizations` specific to the text domain of DCEP. We must note that the parallel classifier was trained with data from the same domain as the DCEP dataset (Section 3.2.2).
- **All modules** - Combining all previous modules favored significantly the recall of the configuration, mainly on `organizations` (Table C.1, page 96). However, it combines the false-positive detections from all previous modules reducing the precision. Even so, it increased significantly the f1-score of the NER.

Both macro-averaged and micro-averaged token-level, and entity-level averaging follow the same tendency when testing with different configurations (Figure 5.3).

5.2.1.2 German

The following graphs present the performance of the NER in German documents in 6 different configurations for each dataset: CoNLL 2003 dataset in Figure 5.4 and DCEP dataset in Figure 5.5. Figure 5.6 shows the performance of NER for different averaging strategies. Table 5.3 presents the significance tests between the performance of each combination and the baseline.

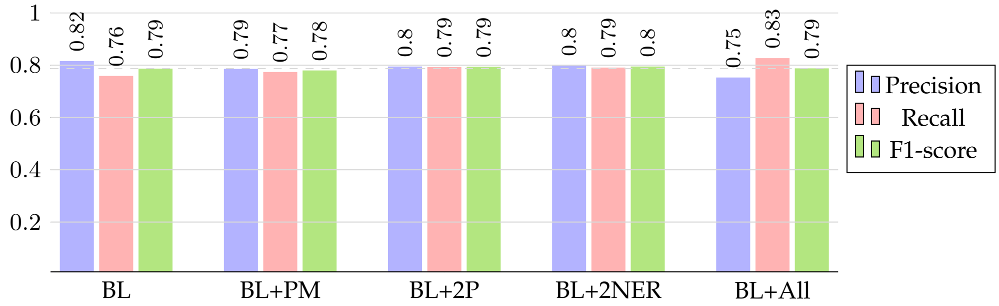


Figure 5.4: Performance of the macro-averaged token-based classification by NER module in the German CoNLL 2003 dataset.

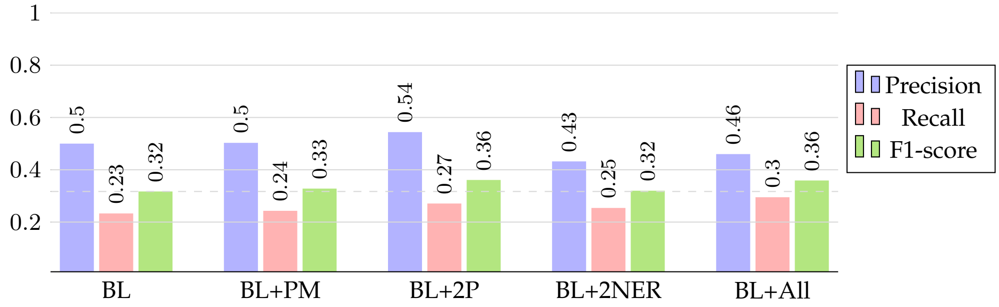


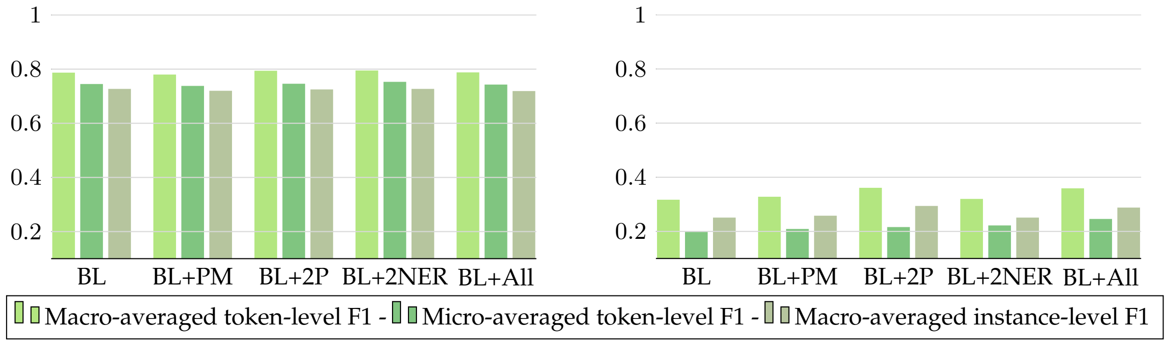
Figure 5.5: Performance of the macro-averaged token-based classification by NER module in the German DCEP dataset.

Dataset	**BL+PM**	**BL+2P**	**BL+2NER**	**BL+All**
CoNLL	-	F	-	F
DCEP	-	F	F	-

Table 5.3: Significance tests for the performance of each configuration against the baseline, considering $\alpha < 0.1$ (labels: P = precision, R = recall, F = f1). Labels mark only the performance metrics that are statistically different.

In average, the performance was higher in the CoNLL dataset, possibly motivated by the fact that this dataset belongs to the same text domain of the NER tool training corpus (Section 4.2.2).

- **Baseline** - A considerable number of false-negatives occur in `organizations` and `locations`, mainly on DCEP dataset (Table 5.4) resulting in a lower recall in this dataset. Some `organization` entities that are specific to the text domain of DCEP and were not detected by the baseline NER, such as "*Weltgesundheitsorganisation*", are the reason for this result.
- **Pattern-matching** - This module did not vary significantly the performance when compared with the baseline.



(a) Performance of NER in the CoNLL dataset using different averaging strategies.

(b) Performance of NER in the DCEP dataset using different averaging strategies.

Figure 5.6: Comparison between macro-averaged token-level, micro-averaged token-level and macro-averaged instance-level performance of the NER module in German datasets.

		System output			
		LOC	PER	ORG	O
Golden Standard	LOC	80	3	10	18
	PER	3	60	5	35
	ORG	17	8	166	44
	O	5	6	12	2775

		System output			
		LOC	PER	ORG	O
Golden Standard	LOC	39	1	1	14
	PER	0	26	0	32
	ORG	0	0	29	467
	O	4	0	80	4021

Table 5.4: Confusion matrixes for the baseline configuration in CoNLL (left) and DCEP datasets (right).

- **Second-pass Detection** - This module increased significantly the f1-score in both datasets due to the detections of short-form mentions such as "*Union*" and "*Zentralbank*" that were not detected in the baseline configuration, that resulted in an improvement of the recall for all classes (Table C.2, page 97).
- **Second NER classifier in parallel** - The use of a NER classifier in parallel resulted in a significantly poorer performance when compared to the baseline, in the DCEP dataset. A considerable number of acronyms were misclassified as `organizations`, increasing the number of false-positives, such as "*CNS*" (6 occurences) , "*EG*" (10 occurrences) and "*KOM*" (31 occurrences). However, there was not a significant variation of the performance in the CoNLL dataset.
- **All modules** - Combining all previous modules results in a significantly different performance (f1-score) in the CoNLL dataset, although little noticeable. There was no a significant performance variation in the DCEP dataset.

Both macro-averaged and micro-averaged token-level, and entity-level averaging follow the same tendency when testing with different configurations (Figure 5.6).

5.2.1.3 Portuguese

The following graphs present the performance of the NER in Portuguese documents in 6 different configurations for each dataset: HAREM dataset in Figure 5.7 and DCEP dataset in Figure 5.8. Figure 5.9 shows the performance of NER for different averaging strategies. Table 5.5 presents the significance tests between the performance of each combination and the baseline.

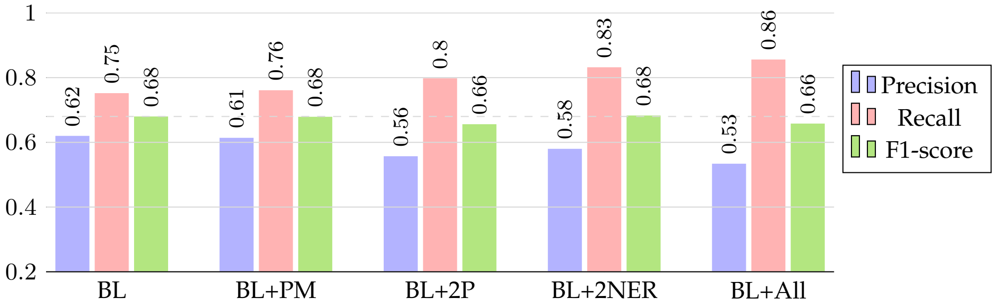


Figure 5.7: Performance of the macro-averaged token-based classification by NER module in the Portuguese HAREM dataset.

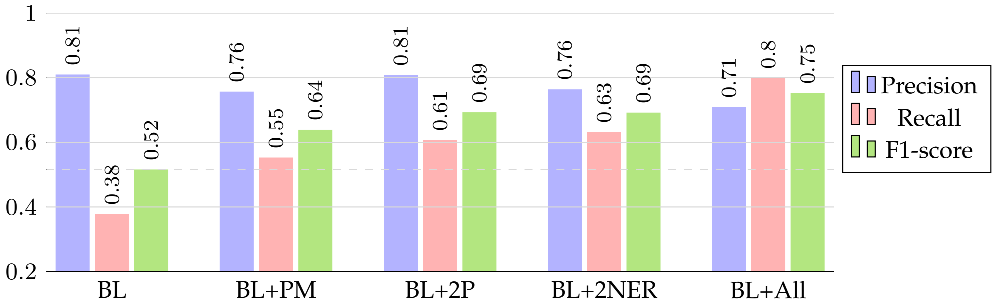


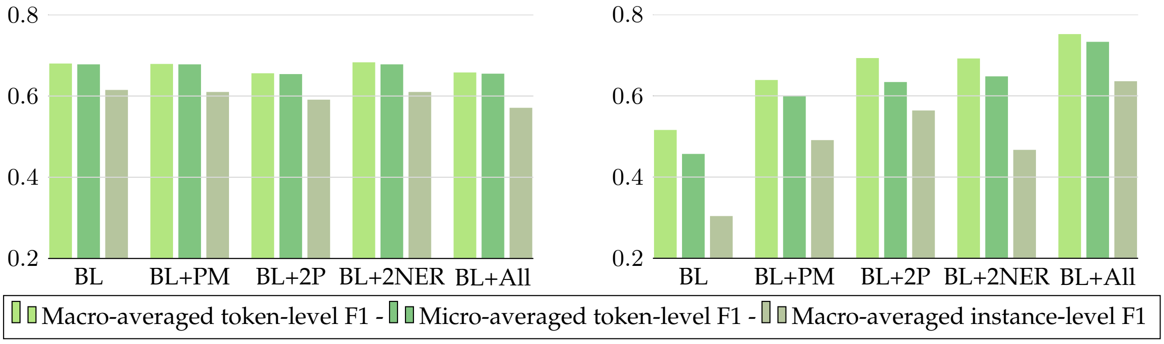
Figure 5.8: Performance of the macro-averaged token-based classification by NER module in the Portuguese DCEP dataset.

Dataset	**BL+PM**	**BL+2P**	**BL+2NER**	**BL+All**
HAREM	-	P, R, F	P, R, F	P, R, F
DCEP	F	R, F	R, F	R, F

Table 5.5: Significance tests for the performance of each configuration against the baseline, considering $\alpha < 0.1$ (labels: P = precision, R = recall, F = f1). Labels mark only the performance metrics that are statistically different.

Overall, the recall increases when modules are combined with the baseline configuration. In DCEP dataset, the overall precision is higher than the recall, caused by a higher number of false-negatives than false-positives.

- **Baseline** - The precision is higher in DCEP dataset than in HAREM for `organizations` and `locations` because there is a lower proportion of false-positives in DCEP dataset for these two classes (Table 5.6).
- **Pattern-matching** - The use of a gazetteer improved significantly the performance in DCEP dataset due to the increase of true-positive detections of `organizations`, mostly entities that belong



(a) Performance of NER in the HAREM dataset using different averaging strategies.

(b) Performance of NER in the DCEP dataset using different averaging strategies.

Figure 5.9: Comparison between macro-averaged token-level, micro-averaged token-level and macro-averaged instance-level performance of the NER module in Portuguese datasets.

Golden Standard	System output: LOC	PER	ORG	O
LOC	387	1	6	18
PER	5	472	5	17
ORG	20	15	313	40
O	18	2	13	7753

Golden Standard	System output: LOC	PER	ORG	O
LOC	163	0	1	26
PER	2	183	2	51
ORG	24	0	544	428
O	5	0	60	8995

Table 5.6: Confusion matrixes for the baseline configuration in HAREM (left) and DCEP datasets (right).

to text domain of DCEP and were not detected by STRING, such as "*Parlamento Europeu*" and "*Europol*". However, the use of the gazetteer added some false-positives in the case of ambiguous terms such verbal form "*Tendo*" (having) and the conjunction "*Como*" (as) both classified as cities.

- **Second-pass Detection** - Adding the Second-pass detection raises significantly the recall of all classes in both datasets (Table C.3, page 98), mainly the recall of `organizations`, due to the detection of partially mention entities. Although the performance (f1-score) was significantly improved in the DCEP dataset, it performed poorer in the HAREM dataset due to a significant drop in the precision motivated by the propagation of false-positive detections (e.g. the classification of "*Reforma Protestante*" as `person` propagated this false-positive throughout the text).
- **Second NER classifier in parallel** - The use of a NER classifier in parallel raised significantly the recall of all classes more than any other module, but also added some false-positives (mostly `persons` being classified as `locations`) lowering significantly the precision in the HAREM dataset (Table C.3, page 98). This module improves significantly the performance (f1-score) in both datasets, although it is less noticeable in the HAREM dataset.
- **All modules** - The baseline system with the combination of all the previous modules results in a significantly higher recall but also lower precision. The contributions of all the modules motivated the high recall, but all the false-positive detections also were accumulated, resulting in a lower precision. The performance (f1-score) is improved significantly in the DCEP dataset and is lowered significantly in the HAREM dataset.

Both macro-averaged and micro-averaged token-level, and entity-level averaging follow the same tendency when testing with different configurations (Figure 5.9).

5.2.1.4 Spanish

The following graphs present the performance of the NER in Spanish documents in 6 different configurations for each dataset: CoNLL 2002 dataset in Figure 5.10 and DCEP dataset in Figure 5.11. Figure 5.12 shows the performance of NER for different averaging strategies. Table 5.7 presents the significance tests between the performance of each combination and the baseline.

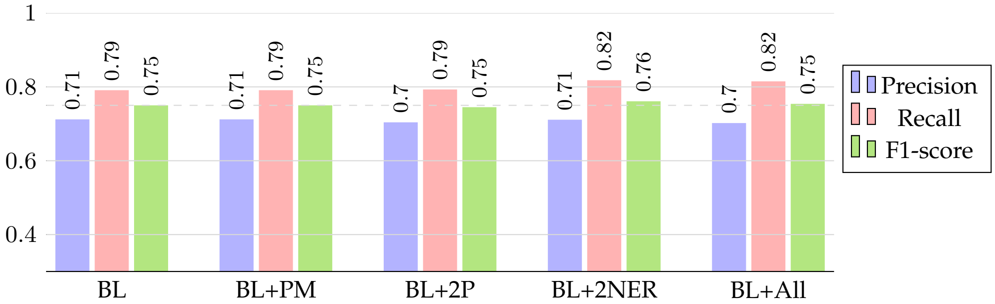


Figure 5.10: Performance of the macro-averaged token-based classification by NER module in the Spanish CoNLL 2002 dataset.

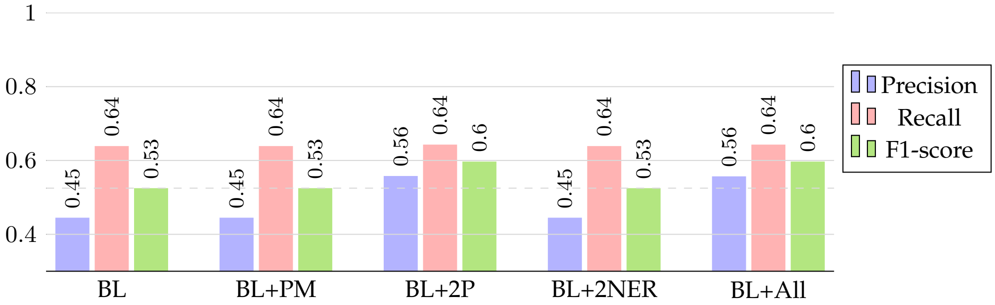


Figure 5.11: Performance of the macro-averaged token-based classification by NER module in the Spanish DCEP dataset.

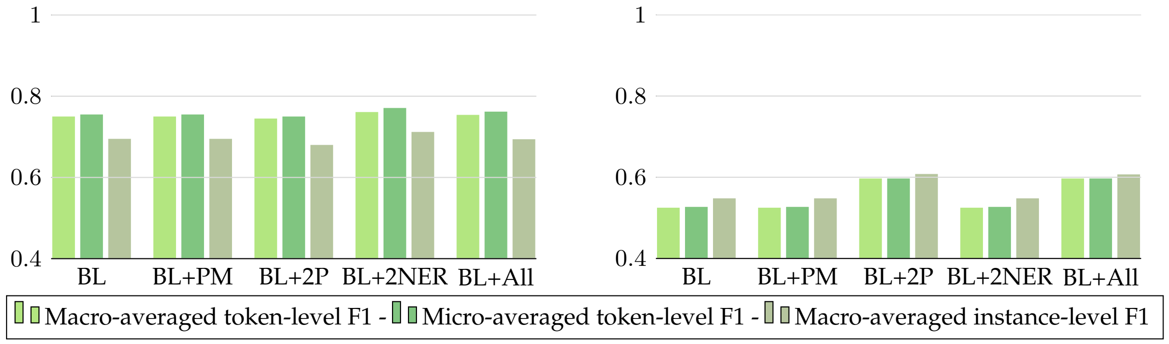
Dataset	**BL+PM**	**BL+2P**	**BL+2NER**	**BL+All**
CoNLL	-	-	F	-
DCEP	-	P, F	-	P, F

Table 5.7: Significance tests for the performance of each configuration against the baseline, considering $\alpha < 0.1$ (labels: P = precision, R = recall, F = f1). Labels mark only the performance metrics that are statistically different.

In average, the performance was higher in the CoNLL dataset. In both datasets, the overall recall is higher than the precision caused by a higher number of false-positives than false-negatives.

(Section 4.2.2).

- **Baseline** - The overall performance is lower in the DCEP dataset due to a high proportion of false-positives in the class `organization` (Table 5.8), such as the expression *“Reglamento (CE) n^{o} 2533/98”*, probably caused by its different text style.
- **Pattern-matching** - This module did not vary significantly the results when compared with the baseline.



(a) Performance of NER in the CoNLL dataset using different averaging strategies.

(b) Performance of NER in the DCEP dataset using different averaging strategies.

Figure 5.12: Comparison between macro-averaged token-level, micro-averaged token-level and macro-averaged instance-level performance of the NER module in Spanish datasets.

		System output			
		LOC	PER	ORG	O
Golden Standard	LOC	90	4	42	13
	PER	3	245	13	20
	ORG	31	11	238	118
	O	6	10	112	6552

		System output			
		LOC	PER	ORG	O
Golden Standard	LOC	10	0	26	35
	PER	0	4	20	0
	ORG	2	0	518	265
	O	1	5	609	4760

Table 5.8: Confusion matrixes for the baseline configuration in CoNLL (left) and DCEP datasets (right).

- **Second-pass Detection** - This module increased significantly the precision and the f1-score in DCEP dataset due to the detections of short-form mentions of `organizations` such as "*Parlamento*". Results from CoNLL dataset did not vary significantly from the baseline.
- **Second NER classifier in parallel** - The use of a NER classifier in parallel resulted in a significantly different performance (f1-score) in the CoNLL dataset when compared with the baseline, although little noticeable.
- **All modules** - The baseline system with the combination of all the previous modules results in a significantly higher precision and f1-score in the DCEP dataset. The main contribution to this performance improvement was made by the short-form mentions of the second-pass detection module. Results from CoNLL dataset did not vary significantly from the baseline.

Both macro-averaged and micro-averaged token-level, and entity-level averaging follow the same tendency when testing with different configurations (Figure 5.12).

5.2.2 Coreference Resolution

The CRR module was evaluated for its performance on resolving coreference relations between entities. Figure 5.13 compares the performance of the CRR module for each dataset with the baseline performance of the CRR module assuming that all mentions belong to singleton groups, i.e., the scenario where no references were found between entities.

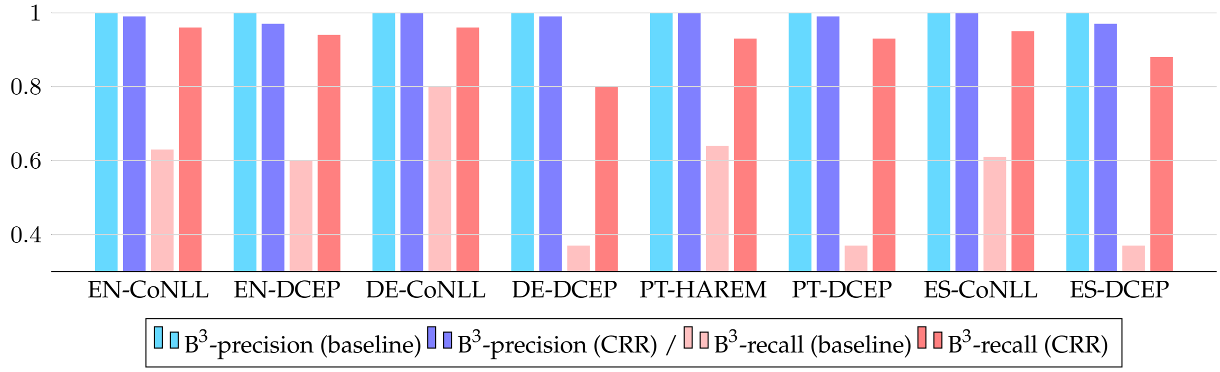


Figure 5.13: Performances of the CRR module and the baseline CRR in terms of macro-averaged B^3-scores of the entity chains.

We set the baseline of the CRR evaluation as the scenario where no references were found, therefore all chain of mentions contain only one entity. We measured the performance, in terms of B^3-score, for the baseline scenario and compared with the performance of our CRR module. The objective of a high-performace CRR module is: (i) maintain the B^3-precision as close as possible to the baseline, as this scenario reflects the highest precision possible (in the case of our system, the highest value is 1.0); (ii) render the B^3-recall as higher as possible, as a lower value than the baseline would mean that a considerable amount of singleton mentions are being incorrectly grouped with other mentions.

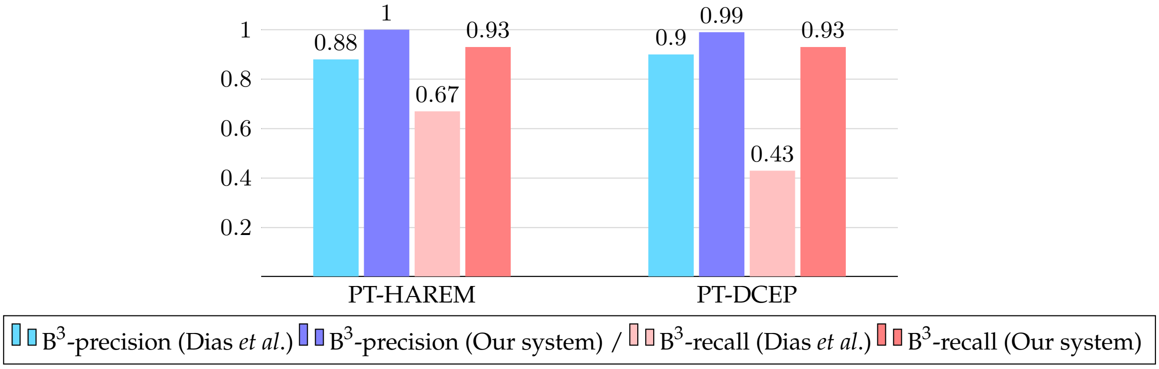


Figure 5.14: Comparison between the performances of our CRR module and Dias *et al.* CRR module in terms of macro-averaged B^3-scores of the entity chains.

Our implementation of CRR showed improvements in the Portuguese datasets when compared to the results of Dias *et al.* [12]. The reason for this is that the implementation of [12] only uses surface forms of entities to link coreferences between mentions, while our implementation uses a much richer set of rules, that are able to resolve abbreviations, such as "*World Health Organization*" and "*WHO*", and short forms of mentions, such as "*Kingdom of Sweden*" and "*Sweden*", or "*Erin Purcell*" and "*Purcell*". The label-consistency technique (Section 4.2.3) also aimed at correcting some inconsistent classifications reported by the authors.

In general, the value of B^3-precision of the CRR module did not vary substantially with the use of different corpora. Two variations of B^3-recall could be noted in the German and Spanish DCEP datasets.

5.2.3 Anonymization

Automated anonymization methods were evaluated based on:

- The ability of human readers to find coreferences between anonymized entities. This is measured using the B^3-score. Some of these results were already presented in the previous section (Section 5.2.2).
- The *availability* of the replacements of a NE provided by a Knowledge Base (KB). The metric of *availability* is introduced in the Section 3.1.4, page 23. We assume that a NE class always have an *availability* of 1.
- The *relevance* of the replacements provided by the method. The metric of *relevance* is introduced in the Section 3.1.4, page 23. We assume that a substitution by the NE class has a *relevance* of 1.

5.2.3.1 Suppression

The suppression method removes the entities from the text, leaving the interpretation of the text to the human readers. In order to understand how a human is able to retrieve some of the original information from an anonymized text with suppressed entities, two human evaluators were asked to find coreference relations between entities in texts anonymized using the suppression method. The average result is compared with the CRR baseline in Figure 5.15 in terms of B^3-score.

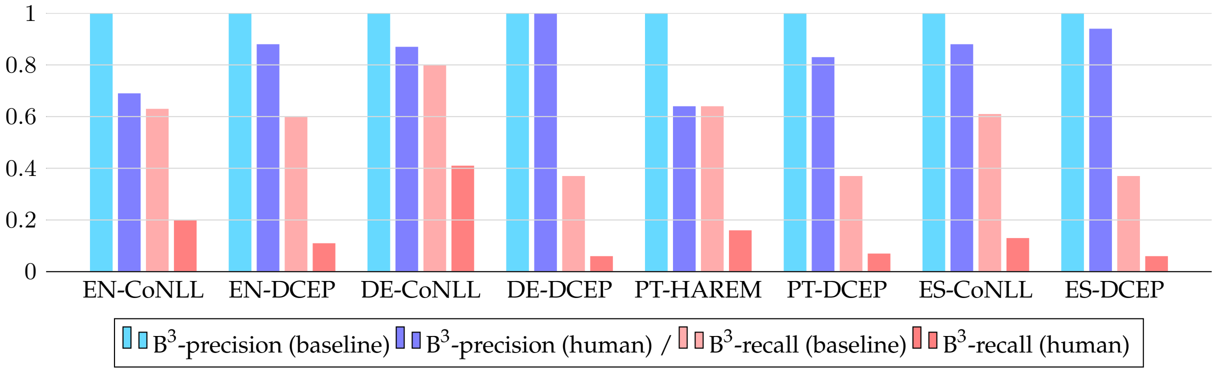


Figure 5.15: Comparison of the performances of the CRR baseline and CRR carried out manually by humans in an anonymized text in terms of macro-averaged B^3-scores of the entity chains.

The lower B^3-recall values of the human-made CRR, when compared to the baseline (Figure 5.15), suggests that human readers were not able to group correctly some (if not most) of the mentions of suppressed entities. However, the sentence D.4.b, on page 105, provides an example of a text in Spanish whose suppressed entities are easily grouped. It was noted that the surrounding words acted like a major clue to detect coreference 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. As this anonymization method suppresses the entities, no *availability* and *relevance* could be computed.

5.2.3.2 Tagging

As the tagging method replaces each mention by a distinct label, the ability of a reader to find coreferences between entities is given by the performance of the CRR module (Figure 5.13). Each label is composed of the class of the entity and a unique identifier and is always provided by the NER module, so we can consider that the *availability* of this method is 1. If we consider that the class of an entity is always a relevant substitution, the *relevance* of this method is given by the overall 3-class performance of the NER module (as presented in Section 5.2.1).

5.2.3.3 Random Substitution

The random substitution method replaces all the occurrences of a mention by the same entry. For that reason, the ability of a reader to find coreferences between entities is also given by the performance of the CRR module (Figure 5.13). The random substitution method always provides a replacement entry from the *default list* (Section 4.3.1). If we consider that the entries of the *default list* cover all the combinations of MFSs and classes of the entities for a given language, we can consider that the *availability* of this method is 1. The *relevance* of the replacements was rated by two human readers for each dataset, and the results are presented in the Table 5.9. The inter-rater agreement for the relevance rating is presented in the Table C.8 on page 100.

	English		German		Portuguese		Spanish	
Entity Class	**CoNLL**	**DCEP**	**CoNLL**	**DCEP**	**HAREM**	**DCEP**	**CoNLL**	**DCEP**
Location	0.45	0.11	0.50	0.05	0.26	0.17	0.69	0.22
Organization	0.23	0.21	0.17	0.13	0.20	0.39	0.11	0.02
Average	0.36	0.16	0.30	0.12	0.24	0.37	0.31	0.04

Table 5.9: Relevance of the replacements by the random substitution method, by dataset and entity class.

5.2.3.4 Generalization

The generalization method replaces all the entities from a mention by a distinct entry. For that reason, the ability of a reader to find coreferences is given by the performance of the CRR module (Figure 5.13). This anonymization method queries a Knowledge Base (KB) in order to provide a replacement entry that generalizes a given entity. The *availability* was computed as the rate of generalizations given by the KB and the total number of entities (Equation 3.8 on page 23). The *availability* of generalizations for this method is presented in Table 5.10.

	English		German		Portuguese		Spanish	
Entity Class	**CoNLL**	**DCEP**	**CoNLL**	**DCEP**	**HAREM**	**DCEP**	**CoNLL**	**DCEP**
Location	0.96	0.99	0.95	1.00	0.91	1.00	0.88	1.00
Organization	0.83	0.96	0.90	0.96	0.82	0.68	0.77	0.98

Table 5.10: Distribution of generalization availability by dataset and entity class.

Entries assigned by this anonymization method were rated for the relevance of the replacement of the NEs. This rating was made by two human readers, who assigned a binary score (Equation 3.9 on page 23) to each replacement entry. This score is based on the proper fit of the entry into the context. The average result of the *relevance* rating are presented in Table 5.11. The inter-rater agreement for the relevance rating is presented in the Table C.8 on page 100.

	English		German		Portuguese		Spanish	
Entity Class	**CoNLL**	**DCEP**	**CoNLL**	**DCEP**	**HAREM**	**DCEP**	**CoNLL**	**DCEP**
Location	0.84	0.93	0.73	0.67	0.52	0.81	0.63	0.90
Organization	0.60	0.73	0.73	0.48	0.53	0.45	0.58	0.41
Average	0.75	0.77	0.73	0.50	0.52	0.49	0.60	0.44

Table 5.11: Distribution of generalization relevance by dataset and entity class.

5.3 Performance

We measured the time performance of each module over the evaluation golden corpora for each module configuration. The results presented in Table 5.12, show that the implementation of the NER module has an quite acceptable time complexity except when gazetteers are used. Gazetteers are loaded into memory each time the NER module receives a new dataset, causing a delay as some of the gazetteers are quite large (e.g. Repentino (Sarmento *et al.* [48]) gazetteer has 380,176 entries). The combination of all the modules have similar time performance due to the loading of gazetteers.

We measured the time consumed at the anonymization task over the evaluation golden corpora for

Configuration	Total Time (s)	Avg Time Doc (ms)
Baseline	6.9	29.7
BL+PM	1,773.0	7,708.7
BL+2P	101.8	442.8
BL+2N	12.2	53.2
BL+All	1,915.2	8,326.0

Table 5.12: Time performance of the NER module.

each anonymization method. The results presented in Table 5.13, show that the access to the external dictionary and KB make the substitution methods much slow than suppression and tagging.

Anonymization Method	Total Time (s)	Avg Time Doc (ms)
Suppression	7.1	30.9
Tagging	7.3	31.7
Random Substitution (+ Wiktionary access)	1,976.4	8,593.0
Generalization (+ Wikidata access)	9,764.2	42,453,0

Table 5.13: Time performance of the anonymization module.

5.4 Evaluation on i2b2 Datasets

We evaluated the performance of the baseline configuration of our system using the corpora from the i2b2 2006 Deidentification Challenge (Uzuner *et al.* [59]) and the i2b2/UTHealth 2014 Deidentification Challenge (Stubbs *et al.* [50]), and compared the performance of our system with others participants in the challenges. Each corpus is composed of a training and a testing dataset.

These datasets contain free-text medical records written in English but with a different text type from those previously used in the evaluation. The i2b2 2006 training dataset contains 669 medical records annotated for 8 major classes of PHI (age, date, doctor, hospital, id, location, patient and phone) divided into 25 subcategories. A total of 220 medical records were used as the testing dataset. The i2b2 2014 training dataset contains 1,304 medical records annotated for 7 major classes of PHI (age, contact, date, id, location, name and profession) divided into 25 subcategories.

The sensitive information contained in the PHIs were replaced by random surrogates with the particularity of having some entities replaced by out-of-vocabulary surrogates, such as "*Dr. Stable*" or "*Mr. A*", which cannot be found in gazetteers.

We trained our NER module based on a CRF classifier [18] with the i2b2 training datasets using the following features: word-form, class, current token, 3-token context, word shape (capitalized words, acronyms, and numbers).

We used the metrics of precision, recall and f1-score to evaluate and compare the performance of our system. In both datasets, we used the categorization into major classes of PHIs and micro-averaged token-level evaluation. The Table 5.14 presents the results of our system using the i2b2 2006 and 2014 datasets.

System	Precision	Recall	F1-score
Wellner [62]	0.987	0.975	0.981
Szarvas [52]	0.993	0.969	0.980
Arakami [1]	0.991	0.949	0.970
Hara [24]	0.961	0.938	0.949
(Our system)	0.964	0.922	0.943

(a) Evaluation with i2b2 2006 datasets

System	Precision	Recall	F1-score
Nottingham [64]	0.9900	0.9640	0.9768
Harbin-Grad [28]	0.9776	0.9629	0.9702
UNIMAN [11]	0.9825	0.9567	0.9694
(Our system)	0.9179	0.8498	0.8825

(b) Evaluation with i2b2 2014 datasets

Table 5.14: Comparison of the result our system with results from some participants of the i2b2 shared tasks using micro-averaged token-based binary classification.

This evaluation showed that our system achieves a very acceptable performance when using medical records in English, performing better (in terms of f1-score) than the average result from both i2b2 shared tasks. However, the top systems of these shared tasks achived a higher performance as they use a wider set of features such as the position within the document, POS, among others.

5.5 Discussion

In this section we discuss the results obtained in the previous sections of this chapter. The section is divided into three main topics: NER, CRR and anonymization.

5.5.1 Named Entity Recognition

Size of evaluation datasets - One of the concerns during the creation of the golden-standard datasets was their size in order to obtain statistically significant results during the evaluation. The English and Portuguese datasets (Table 3.2 on page 25) showed to have enough size to obtain statistically significant differences between configurations for most metrics. On the other hand, the German and Spanish datasets, containing less than one-half of the number of entities than the other two datasets, were still enough to compare configurations based on the f1-score.

NER performance depends on the text domain - The overall results of the NER module suggest that the performance of the detection of sensitive information strongly depends on the domain of the text. The NER classifiers for English and German were trained with data derived from the CoNLL 2003 training datasets, therefore, the NER performance with the CoNLL test datasets achieved a significantly higher performance than in DCEP. The lower performance of the NER in the DCEP datasets when compared to CoNLL was the result of a significantly lower recall in organization entities (Tables C.1 and C.2), which were very specific to the DCEP domain (some examples have been presented along Section 5.2). Such difference was much more pronounced in the German DCEP dataset. On the other hand, the difference of performance between the two datasets (CoNLL and DCEP) in the Spanish documents was less pronounced, as the NER classifier was trained with AnCORA (Taulé *et al.* [53]), a corpora with a different text style. Nevertheless, organization entities that were very specific to the domain of DCEP still caused a significant number of false-negatives, resulting in a lower recall.

Micro-averaged versus Macro-averaged - We compared macro- and micro-average token-level results in order to understand the distribution of average NER performance along every single document. In general, the micro-averaged f1-score follows the tendency of the macro-averaged f1-score in all configurations. Small drifts to that tendency, specially on DCEP dataset, occured due to an odd performance of the NER coused by documents with different distribution of NE classes. Micro-averaged f1-score is lower than macro-averaged. As the tendencies are very similar, from now on we refer only the macro-averaged in this discussion.

Token-level versus Instance-level - In general, all configurations result in the same or better instance-level performance (f1-score) than the baseline. This performance follows the same tendency as the token-level performance except for some parallel NER configurations.

Usage of different tagsets - The Portuguese rule-based NER classifier also showed significantly higher precision on DCEP than in HAREM. One of the main causes for this difference was the conversion between the NE tagset from HAREM to our standard tagset (Section 4.1.2). In the original HAREM dataset there is the possibility of multiple classifications for a NE and the granularity of tagset is different. In our HAREM dataset, these multiple classifications were removed and converted into a 3-class tagset. For those reasons, the classification of some organizations and locations could be considered incorrect during the tests. This issue raises a concern on the conversion between different tagsets in a modular system as our anonymization system.

Gazetteers improve recall - The use of gazetteers showed to improve the average recall, and it improved significantly the performance (f1-score) in some datasets. This suggests that the use of gazetteers might improve the performance of NER, although this conclusion cannot be extended to every gazetteer. The use of gazetteers also adds false-positives due to the ambiguity between entities and non-entities' expressions (as described in Section 5.2.1.1). This ambiguity could result in a poorer average precision, although this variation was not significant in our experiments. It is important that the gazetteers do not contain this type of ambiguous expressions in order to avoid false-positives. Table 5.15 shows a result where the use of a gazetteer allowed to detect a NE "*Bundesrepublik Deutschland*" not detected by the baseline configuration.

September 1976 zur Einführung allgemeiner unmittelbarer Wahlen der Mitglieder des [Europäischen Parlaments] , unter Hinweis auf Artikel 46 des Grundgesetzes der [Bundesrepublik Deutschland] , unter Hinweis auf seinen Beschluss vom 16 .

(a) Result of the BL configuration.

September 1976 zur Einführung allgemeiner unmittelbarer Wahlen der Mitglieder des [Europäischen Parlaments] , unter Hinweis auf Artikel 46 des Grundgesetzes der **[ORG Bundesrepublik Deutschland]** , unter Hinweis auf seinen Beschluss vom 16 .

(b) Result of the BL+PM configuration.

Table 5.15: Comparison of the result of the baseline configuration and gazetteer in a sample text extracted from the German DCEP. Entities are surrounded by square brackets. Detected entities are highlighted in bold typeface.

Second-pass detection propagates (in)correct classifications - The second-pass detection improved significantly the performance (f1-score) of NER in some datasets due to the detection of short-form NEs that were not detected by the NER module (Sections 5.2.1.1, 5.2.1.2 and 5.2.1.3). The poorer performance in the HAREM dataset (Table 5.7) suggests that this technique should be used in conjunction with NER classifiers with 'good enough' precision in order to avoid the propagation of false-positives. Table 5.16 compares a result from the baseline configuration and the second-pass detection configuration. In this example, the headwords *Comité* and *Conselho* detected by the NER module (highlighted in Figure 5.16a) were propagated and successfully marked two undetected entities (Figure 5.16b).

Resolução do [Parlamento Europeu] referente à Comunicação da Comissão ao [Conselho] , ao [Parlamento Europeu] , ao [Comité Económico e Social] e ao **[ORG Comité das Regiões]** sobre o Plano de Acção da [Comissão para as Competências e a Mobilidade] Trabalhos preparatórios da reunião do **[ORG Conselho Europeu]** de 24 e 25 de Outubro

(a) Result of the BL configuration.

Resolução do [Parlamento Europeu] referente à Comunicação da Comissão ao **[ORG Conselho]** , ao [Parlamento Europeu] , ao **[ORG Comité Económico e Social]** e ao **[ORG Comité das Regiões]** sobre o Plano de Acção da [Comissão para as Competências e a Mobilidade] Trabalhos preparatórios da reunião do **[ORG Conselho Europeu]** de 24 e 25 de Outubro

(b) Result of the BL+2P configuration.

Table 5.16: Comparison between results of the BL and BL+2P configurations in a sample text extracted from the Portuguese DCEP. Entities are surrounded by square brackets. Detected entities are highlighted in bold typeface.

NER classifiers in parallel - The use of a NER classifier in parallel trained with a silver-standard DCEP corpora (Section 3.2.2) improved significantly the performance (f1-score) of NER in some DCEP datasets (Tables 5.2 and 5.8). However, this technique did not raise the performance significantly when compared with other simpler techniques as the second-pass. In addition, these NER classifiers are dependent on the text domain as they did not improve the performance when using other datasets than DCEP. One solution could be the use of several NER classifiers, each one specific to a single text domain, and then to combine the results of all the classifiers.

Best NER configuration - Based on the results, the best configuration is a NER classifier trained with corpora from the same text domain as it showed to improve significantly the performance in most of the datasets. However, as the text domain is not predictable and training a new NER classifier takes time and resources, a solution for these issues could be combining the baseline configuration with a Second-pass detection, has also proved to result in an improvement of the f1-score for all the configurations with statistically significantly different results ($p < 0.01$). In addition, this configuration resulted in a better performance (f1-score) at instance-level in most DCEP datasets.

5.5.2 Coreference Resolution

CRR performance depends on the text domain - A tendency to a lower B^3-recall of the CRR on all DCEP datasets suggests that the text domain also influences the performance of the coreference. This lower performance, noticed in all DCEP datasets, particularly in German texts, is caused by long-forms and short-forms of entities being grouped incorrectly in different mentions such as "*Republik Lettland*" and "*Lettland*".

CRR performance is satisfactory - Previous studies have been made about coreference resolution. However, none of those studies aimed exclusively at the coreference of NEs or proper nouns. For that reason, we cannot compare directly the results of our implementation of CRR with these previous studies. Even so, we consider acceptable the performance of the CRR module for the purpose our work, as the CRR scored a minimum B^3-precision of 0.97 for the datasets of our study. A high value of B^3-precision means that the CRR tends to group mentions correctly. The performance of CRR was influenced by the accuracy of the headword extraction. For that reason, an improvement in the extraction of the headword, and also the use of synonyms and abbreviations of the headword, may improve the overall performance of the CRR.

5.5.3 Anonymization

Along this section, we use the following text sample in order to provide examples that illustrate the result for each anonymization method:

> Among other points , **Parliament** states its readiness to evaluate proposals for a general correction mechanism based on the principle of **Community** solidarity .

Suppressing entities hinders text understanding - The suppression method does not maintain the information about the entities and the context around them 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 (Figure 5.15). The B^3-recall of human-made coreference resolution is lower than the baseline in most datasets, indicating that human readers were not able to recognize successfully most of the links between entities. The B^3-precision displays a low value for all datasets, suggesting that human readers also tend to group distinct mentions into the same group. The performance of this method is dependent on the performance of the binary NER classification (tables of Section C.1 on page 95). In order to ensure better anonymization using the suppression method, the NER module must achieve a higher recall. The following text is an example of the result of the suppression method:

```
Among other points , ****** states its readiness to evaluate proposals for a
general correction mechanism based on the principle of ****** solidarity .
```

Tagging is not natural but is satisfactory - The tagging method facilitates the resolution of references between NEs for a human reader. Although the replacement labels may allow a reader to identify the class of the anonymized entity, the result is a text containing labels instead of NEs, which does not resemble a natural language text. The performance of this module is strictly dependent on the performance of the NER and CRR modules. The CRR module has been shown to run with an acceptable performance (Section 5.5.2), however, some resolution errors occur due to an inconsistent classification of entities that mention the same object. For that reason, it is important to increase both precision and recall of the NER classification in order to improve the performance of this anonymization method. The following text is an example of the result of the tagging method:

```
Among other points , ORGANIZATION1 states its readiness to evaluate proposals for a
general correction mechanism based on the principle of ORGANIZATION2 solidarity .
```

The random substitution hinders relevance - The random substitution method provides a simple solution for anonymization with an output that resembles a natural language text. However, the relevance of the replacements was low because entities were chosen randomly (see Table 5.9), so they were often replaced by an entity outside the context (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 generic or vague enough, to avoid semantic drifts. The following text is an example of the result of the random substitution method:

```
Among other points , Entreprise 1 states its readiness to evaluate proposals for a
general correction mechanism based on the principle of Society 1 solidarity .
```

In inflected languages, this method seems to find replacement entities with the same MSFs with an very acceptable performance (example in Spanish in Sentence D.4.d). Some errors may occur due to an incorrect headword extraction causing a disagreement in the grammatical gender (see example in Portuguese in Sentence D.3.d). The replacement "*Unternehmen*" in the German text (Sentence D.2.d) also agrees with the dative case of the original entity.

Generalizations - The generalization method provides an output similar to a natural language text. Entities are replaced by their superclasses in a Knowledge Base (KB), resulting in a considerably higher relevance than that of the random substitution method. Entities that designate locations achieved a good relevance score, as they were easier to query in the KB. Ambiguous entities usually raise NE linking issues because an entity may have several entries in the KB with the same name. Some examples of irrelevant substitutions are presented in Table 5.17.

Language	Original	Substitution
English	Newfoundland	Dog breed
German	Belfast	Town in den Vereinigten Staaten
Portuguese	Centro de Astronomia	Região de França
Spanish	Mérida	Película de animación

Table 5.17: Examples of irrelevant substitutions caused by Named Entity Linking issues.

The following text is an example of the result of the generalization method:

```
Among other points , legislature states its readiness to evaluate proposals for a
 general correction mechanism based on the principle of social group solidarity .
```

One of the issues of this method is that the substitutions often show no grammatical agreement with their contexts, or they lack determiners. Another issue that occurs frequently in generalizations of `locations` is the appositive use of proper names (toponyms) to classifiers, which leads to the repetition of the context of the entity. The sentence *"the southern province of* ***Guangdong****"* was automatically anonymized as *"the southern province of* ***province of China****"*. In this case, the NER module detected *"Guangdong"* as an entity, when the classifier *"province of ..."* should also be included. Incorrect generalizations may also occur when no translation is available for a given generalization. The sentence D.3.e on page 105 shows an example where *"comunidades no* ***Alto Rio Negro****"* was replaced by *"comunidades no* ***en:diocese of the Catholic Church****"* due to the unavailability of a translation of the generalization to Portuguese. Finally, the poor time performance of this method may make it not suitable for anonymizing large numbers of documents, long documents or in time-critical applications.

5.5.4 Overall

Anonymization methods - 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 of a natural text to a human reader. However, this method is limited by the recall of the KB and its time performance. Even so, generalization performs much better than random substitution. There is no anonymization method that fits all scenarios. 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.

Still not a perfect process - Nowadays, state-of-the-art automated text anonymization is far from being a perfect process, however, the performance of the detection of sensitive information could be useful to speed up the process of manual de-identification at a post-edition step.

Indirect identifiers - Even in an anonymized text, some entities can still be retrieved through an analysis of the indirect identifiers by an informed reader. The indirect identifiers may act as clues to re-identify some entities, being the worst-case scenario when a fully informed reader is able to identify all the anonymized information from a text. The following examples are presented in the Appendix D: (i) in the German text (Section D.2), the expression *"SPD-Forum 2010"* is an indirect identifier that allows to reader to retrieve some of the anonymized entities; (ii) in the Spanish text (Section D.4), the context of *"tenista brasileño"*, *"romance (...) con la modelo"* and *"posar desnuda para (...) la revista"* allows an informed reader to recognize unequivocally almost all the entities from the text.

Comparing anonymization with previous studies - Finally, the performance of the anonymization methods of this system cannot be directly compared with 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. The comparison of the performance of NER with the results from the i2b2 challenges showed that our system achieved above-average results in the detection of sensitive information.

5.6 Summary

This chapter presented the experiments conducted in order to evaluate the text anonymization system. The evaluation aimed at 3 different modules: NER, CRR and anonymization. The evaluation was performed using a collection of 8 datasets of texts, 2 for each language (English, German, Portuguese and Spanish). The evaluation showed that the NER module performs better when the NER classifier was trained with corpora belonging to the same text domain as the documents. Also, the second-pass detection was a solution that makes possible to raise the performance in all datasets. The CRR performs well when compared with the baseline. The anonymization methods were evaluated based on the adequacy of the output as a natural text and against the coreference resolution task. The 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 removal of the original information.

6 Integration

"Cyclops, you asked my noble name, and I will tell it; but do you give the stranger's gift, just as you promised. My name is Nobody."

Homer, *Odyssey* (ca. 850 BC)

OUR ANONYMIZATION SYSTEM can operate as a stand-alone service of automated text sanitization or it can be integrated into other systems. In this chapter, we present three ways if integrating the system into a larger NLP system: a re-identification system, an anonymization module for the STRING chain and the Unbabel's translation pipeline.

6.1 Re-identification

Re-identification is the process of restitution of the original entities into an anonymized text. A re-identification service receives an anonymized text and a *table of solutions* and restitutes the original entities into the text. We implemented this service, based on regular expressions, by searching for the position of each entity inside an anonymized text and replacing it by the original entity. Re-identification can be performed in one of the following conditions:

1. Given the position of the entity inside the text that is specified in the table of solutions;
2. Given a different placeholder (e.g. a tag) for each entity that is specified in the table of solutions;
3. Given the order of appearance of a unique placeholder (e.g. a suppression placeholder) in the text;

These conditions ensure that all entities in a text are replaced by the correct entities because the text is translated by humans, therefore, the human redactors can change the order or surface of the NEs to the point of not being recognized by the system. For that reason, the use of unique markups around the entities or the same order of appearance is important for a good performance of this module.

6.2 Integration in the STRING chain

One of the final objectives of our work is to implement our anonymization system in the STRING processing chain (Mamede *et al*. [29]), at INESC-ID Lisboa L2F, in order to provide the chain with a new functionality, enabling the automated anonymization of text documents in Portuguese.

The STRING chain is composed of a pipeline of modules, illustrated in Figure 6.1. These modules perform the basic operations from a NLP system such as text pre-processing (tokenization, text normalization and segmentation into sentences), POS tagging (both rule-based and model-based), chunking and dependency extraction using the XIP parser. Apart from these operations, the STRING chain is also able to perform operations of NER and anaphoric resolution.

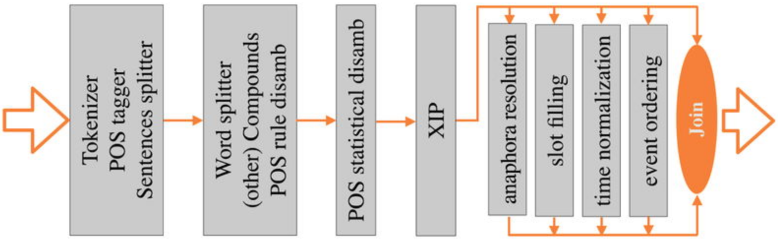


Figure 6.1: STRING chain architecture. Diagram adapted from STRING website[1].

The output of the processing chain is an Extensible Markup Language (XML) file containing the text split into sentences. The XML output of the STRING contains sentences divided into LUNIT elements, which contain a sentence. Each sentence element contains a tree of nodes of type NODE, identified with a unique index, having each token as an element leaf, and a list of dependencies.

```
<NODE num="25" tag="NOUN" start="52" end="59">
   <FEATURE attribute="MASC" value="+" />
   <FEATURE attribute="SG" value="+" />
   (...)
</NODE>

<DEPENDENCY name="NE">
   <FEATURE attribute="WATERCOURSE" value="+" />
   <FEATURE attribute="GEOGRAPHIC" value="+" />
   <FEATURE attribute="LOCATION" value="+" />
   <PARAMETER ind="0" num="25" word="Rio Tejo" />
</DEPENDENCY>
```

(a) A portion of XML output from STRING, with a tree node at the top and a dependency link below.

```
{
    "begin": "52",
    "end": "59",
    "entity_class": "LOCZ",
    "gen": "M",
    "num": "S",
    "surface": "Rio Tejo",
    "uid": 2
},
```

(b) A portion of JSON input corresponding to the output in (a).

Figure 6.2: Conversion of the output of STRING chain to the format of the input of the anonymizer.

In this integration, we used the STRING chain as pre-processing step and NER modules of the anonymization pipeline (Figure 4.1 on page 36). Other modules, such as pattern-matching and second-pass detection, can also be used in conjunction with STRING. The result of STRING's NER comes in the form of dependencies, containing the identifier of a node from the tree and some classes marked as attributes. A sample from the XML output of the STRING chain is shown in Figure 6.2a.

We created a converter between the schemas of the STRING XML output and the anonymizer JSON job (Section 4.4), and we fed the JSON job directly to the anonymization module of our system, as portraited

[1] URL: https://string.l2f.inesc-id.pt/w/index.php/Architecture

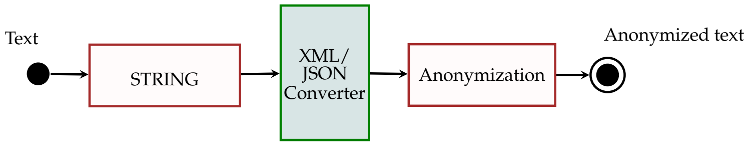


Figure 6.3: Diagram of the integration of the anonymization module to STRING using a converter.

by Figure 6.3. This convertor uses the position of the NEs inside the text, the MFS, surface form and class of the entity provided by STRING. The class of the NE was convert from the entity classification directives adopted at STRING (Oliveira [40]), which is divided into a category, type and subtype, into the 3-class categorization used in our work. The comparison between the NE classification directives of STRING and our system are listed in Table 6.1.

STRING		Our System
Category	**Type**	**3-Class**
*	INDIVIDUAL	PER
*	COLLECTIVE	ORG
LOCATION	∼ VIRTUAL	LOC

Table 6.1: NE classification directives of STRING and equivalent 3-class classifications in our system.

Figure 6.2 shows an example of a conversion of a STRING output to a JSON job.

The STRING chain is also able to perform anaphora resolution (Marques [30]). The result of the anaphora resolution contains pairs of *antecedents* and *mentions* (or anaphoras). By filtering anaphoras whose both antecedents and mentions contain NEs, it is possible to convert the anaphoras into the coreference chain of entities to be used by the anonymization system. Otherwise, the CRR module of our anonymization system can be used instead.

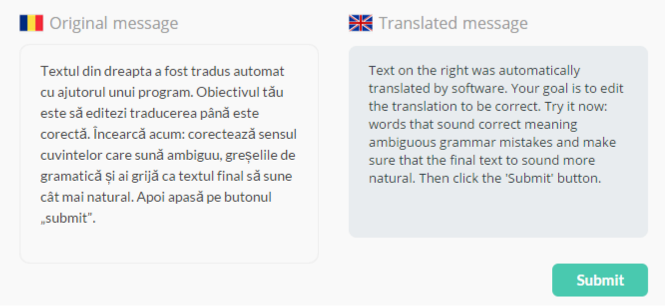
The anonymization module receives the JSON *job* and executes the selected anonymization methods. For security reasons, the random anonymization and generalization methods do not access external resources provided by the Wiktionary Query Service (Section 3.3.1) and the Wikidata Query Service (Section 3.3.2). For that reason, the morphosyntactic features (MSF) are provided by the STRING output and the KB are stored locally in a PostgreSQL database [42].

The local KB is derived from a portion of the dumps of Wikidata. The multilingual dump of Wikidata was filtered in order to keep only Portuguese entries. The KB stored in the database contains the relationships (statements) between objects and the properties '*subclass of*', '*instance of*', '*member of*' and '*part of*' for each object. The UML diagram of the database is presented in Appendix B.

The anonymization module is invoked by adding the argument `-anon` to the running parameters of the STRING script executable. The configuration of the anonymization system is defined in a configuration file. Different configuration files can be defined by calling the STRING with the argument `-anonconfig` `file.json`. By default, the anonymization system will output a plain text, however, the format of the output can also be selected in the configuration file. A demonstration of the text anonymization system can be tested at the STRING website: `https://string.l2f.inesc-id.pt`.

6.3 Integration in the Unbabel pipeline

Unbabel provides an online platform for collaborative translation. Through this platform, customers start translations tasks by sending their texts to be translated into other languages. The original texts are pre-processed and automatically translated using a Machine Translation (MT) system. Then, the pair of original and MT translated texts are sent to human translators, whose task is to correct the MT output in order to create an accurate translation. Figure 6.4 shows an example of a translation task.



Textul din dreapta a fost tradus automat cu ajutorul unui program. Obiectivul tău este să editezi traducerea până este corectă. Încearcă acum: corectează sensul cuvintelor care sună ambiguu, greșelile de gramatică și ai grijă ca textul final să sune cât mai natural. Apoi apasă pe butonul „submit".

Text on the right was automatically translated by software. Your goal is to edit the translation to be correct. Try it now: words that sound correct meaning ambiguous grammar mistakes and make sure that the final text to sound more natural. Then click the 'Submit' button.

Figure 6.4: An example of the Unbabel's translation task interface.
This example shows the translation of a mockup text from Romanian to English.

This integration aims at providing a service to Unbabel's customers that allow their text documents to be anonymized before being distributed among the human translators. The simple idea behind this service is to de-identify the text documents using our text anonymization system, to distribute an anonymized version of the texts among the translators, and to store locally the *table of solutions* for that anonymization task. Then, the translated texts returned by translators is re-identified using the re-identification service (Section 6.1). The anonymized text contains placeholders instead of the original entities that are stored in the *table of solutions*. However, this integration poses four challenges:

Firstly, the automated text anonymization process is not perfect, and it will not be able to detect correctly all the entities and may incorrectly identify others (Section 5.2.1). This raises a concern as it may expose partially some sensitive information about the customers. A solution to this problem could be to provide an additional interface to the customers, that may allow them to review and correct the result of the anonymization process.

Secondly, the placeholders of the NEs should be kept along the pre-processing, automatic translation and human translation in order to successfully re-identify the text. Both pre-processing and MT tasks apply automated changes to the text and this processing may change the surface form of the

NE placeholders. A solution to this problem could be to apply a markup tag to the placeholders in order to inform the translation pipeline not to process that segment of text. Some MT systems, like Moses (Koehn *et al.* [25]), a state-of-the-art statistical MT system, provide a markup tag to inform the MT decoder not to translate a segment of text as the following mockup:

```
This is a <ne translation="Named Entity">Named Entity</ne> that will not be translated.
```

However, as the state-of-the-art MT systems are based on statistical phrase-based techniques, the inclusion of these markups may lead to incorrect translations due to incorrect alignments. In order to keep the placeholders during the human translation, the translation task interface (as in Figure 6.4) could provide some mechanism to avoid that placeholders are edited by the translators.

Thirdly, the NEs removed during the de-identification phase of the document may contain expressions that, in some cases, should also be translated. As these expressions contain sensitive information about the customers, which should not be exposed, they have to be translated by the MT system. A solution to this problem could be a MT system dedicated exclusively to the translation of NEs. Some work has already been made in that direction as the translation of NEs is also useful for improving the performance of the MT and automatic alignment of parallel corpora. Among the most recent works on NE translation is the Optima News Translation System (Turchi *et al.* [57]), which uses a parallel NE database to suggest translations of NEs from 11 languages into English.

Finally, the anonymized texts delivered to the translators contain entity placeholders that are re-identified after the translation. Such placeholders may cause confusion during the MT and human translation as the context and MSFs of the entities are not directly revealed by the placeholders. Such confusion may result in semantic drifts of wrong grammatical agreement with the context. The following example of a de-identification, MT translation and re-identification of a text portraits a semantic drift in the translation of "*reunir-se*" (to meet), a wrong context before "*Islamic Republic of Afghanistan*" and an incorrect splitting of the same NE:

```
O Ministro da Defesa deverá reunir-se com o Ministro da Defesa afegão , Raheem
Wardag , e com o Presidente da República Islâmica do Afeganistão , Hamid Karzai .

O ***** deverá reunir-se com o ***** , ***** , e com o ***** da ***** , ***** .

The ***** should get together with the *****, *****, and the ***** the *****,
*****.

The Ministry of Defense should get together with the Afghan Defense Minister ,
Raheem Wardag , and the President the Islamic Republic of Afghanistan , Hamid
Karzai .
```

A partial solution to this problem could be to use an anonymization method whose placeholders provide clues for the context and MSFs of the entity, such as the random substitution and generalization methods. However, both anonymization methods showed to be limited at representing both original NE context and MSFs at the same time (Section 5.5).

The flowchart of the complete integration of the anonymization system in the Unbabel pipeline is presented in Figure 6.5.

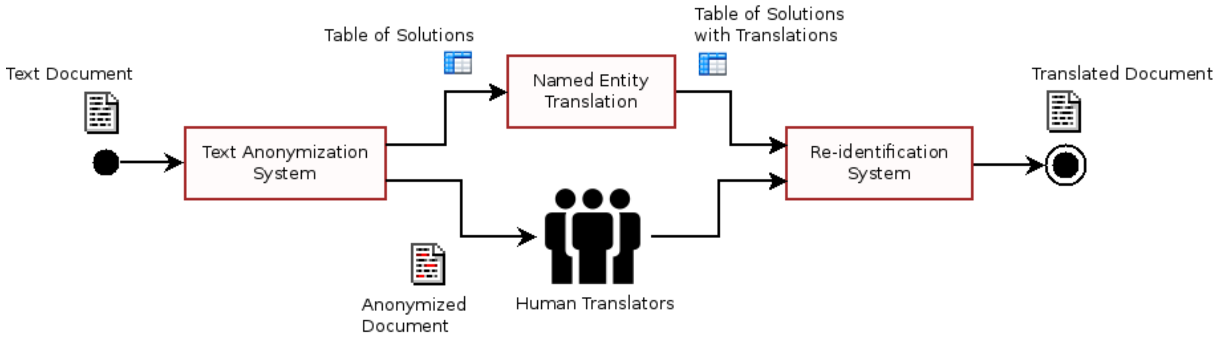


Figure 6.5: Flowchart of the complete anonymization pipeline

6.4 Summary

Our text anonymization system can be integrated in other systems in order to provide an anonymization service. This chapter provided three instances of integration of our system: a simple text re-identification system, the integration into the NLP STRING chain and in the crowdsourcing translation pipeline at Unbabel.

7 Conclusions

"There is no real ending. It's just the place where you stop the story."

Frank Herbert (1920-1986)

IN THIS CHAPTER, we conclude this dissertation by reviewing the contributions of our work and its main challenges, and by outlining future work. Throughout this document, we have presented an original, multilingual, text anonymization system. This work is quite extensive and complex, and it intersects several fields of computer science, while providing the possibility for further development.

7.1 Review

We have presented an implementation of a multilingual, automated anonymization system for text documents. In order to apply a text anonymization service to a multilingual context, we implemented our system in a modular structure, whose modules can be exchanged in the future to support different languages, implement new anonymization methods and change the configuration of the system in order to improve its performance.

The detection of sensitive information is performed by a NER module that consists of a main NER classifier, a pattern-matching module and a set of auxiliary parallel NER classifiers. The classification of all the components is merged into a voting module that determines the classification of each entity. We implemented a module that performs a second-pass detection in order to correct and improve the performance of the NER task.

The evaluation showed that, in different languages and text domains, the use of a single NER tool provides the best precision, but not the best performance. The use of a second-pass detection showed to improve significantly the performance of the detection. The use of parallel NER classifiers trained with corpora from the same text domain as the documents also resulted (as expected) in a significant improvement in the performance of the detection. We evaluated our detection module with the i2b2 de-identification shared task datasets, which resulted in an above-average performance in the text domain of medical reports.

Four text anonymization methods were implemented and evaluated: suppression, tagging, random substitution, and generalization. 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 more natural to a human reader, but most of the times it results in semantic drifts because the entities are chosen randomly from a predetermined list. The generalization method presents a more acceptable solution to text anonymization while keeping the natural fluency of the text, which is appealing 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.

Our system was implemented in the STRING chain as an anonymization system for this NLP chain at INESC-ID. We also provided an implementation of our system at Unbabel's translation pipeline.

7.2 Contributions

With this work we contribute with:

- An implementation of a multilingual text anonymization system, which can be extended in the future.
- A study of the performance of our text anonymization system in different languages, text domains and scenarios.
- A web-based annotation platform.
- A golden-standard for NER corpora composed of DCEP reports.
- The integration of this text anonymization system in the STRING chain.
- A proposal of integration of our system in the Unbabel's translation pipeline.
- As results from this research, a paper was accepted in the WCCI 2016 International Conference, describing our research on automated anonymization for Portuguese texts (Dias *et al.* [12]).
- An article describing our research on text anonymization to be submitted to an international journal.

7.3 Future work

We point the following directions of future work in order to improve this system and continue the research in this area:

- The improvement of the access to a dictionary and a Knowledge Base (KB) in order to boost their time performance and recall. Both substitution methods would benefit from this improvement.

- The improvement of the headword extraction algorithm would improve the performance of both substitution methods.
- The annotation of more documents, in order to improve the size of the evaluation golden corpora, could result in more significant results in the performance of the system especially in the German and Spanish datasets.
- The relevance of the substitutions in the anonymization method can be improved by using methods of Named Entity Linking, taking advantage of the context of the NE, instead of using only the NE surface to look up in a KB.
- The implementation of an intelligent generalization method that would provide a generalization of an entity with the same morphosyntactic features.
- The testing of an information content approach, such as in Sanchez *et al.* [47], and the comparison of its detection performance with the NER module is also envisaged.
- The use of the result of the dependency parsing provided by the STRING would enable us to create and test an anonymization method that would suppress not only the entities but also any indirect identifiers and actions that involve the entities.

References

[1] ARAKAMI, EIJI, IMAI, TAKESHI, MIYO, KENGO, & OHE, KAZUHIKO. 2006. Automatic Deidentification by using Sentence Features and Label Consistency. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 10–11.

[2] BAGGA, AMIT, & BALDWIN, BRECK. 1998. Algorithms for Scoring Coreference Chains. *Pages 563–566 of: In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.

[3] BECKWITH, BRUCE A, MAHAADEVAN, RAJESHWARRI, BALIS, ULYSSES J, & KUO, FRANK. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Med. Inform. Decis. Mak.*, Mar, 12–22.

[4] BERMAN, JULES J. 2003. Concept-Match medical data scrubbing. How pathology text can be used in research. *Arch. Pathol. Lab. Med.*, **127**(6), 680–686.

[5] BONTCHEVA, KALINA, DIMITROV, MARIN, MAYNARD, DIANA, TABLAN, VALENTIN, & CUNNINGHAM, HAMISH. 2002 (24–27 June). Shallow Methods for Named Entity Coreference Resolution. *In: Chaînes de références et résolveurs d 'anaphores, workshop TALN 2002*.

[6] CARVALHO, PAULA, OLIVEIRA, HUGO G, MOTA, CRISTINA, SANTOS, DIANA, & FREITAS, CLÁUDIA. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. 1 edn. Vol. 1. Linguateca. *Online:* http://www.linguateca.pt/LivroSegundoHAREM/. Chap. 1, pages 11–31.

[7] CHINCHOR, NANCY. 1992. The statistical significance of the MUC-4 results. *Pages 30–50 of: Proceedings of the 4th conference on Message understanding*. Association for Computational Linguistics.

[8] COHEN, JACOB. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

[9] CROCKFORD, DOUGLAS. 2006 (July). *RFC4627—The application/json Media Type for JavaScript Object Notation (JSON)*. *Online:* http://www.ietf.org/rfc/rfc4627.txt?number=4627. last accessed - 15/04/2016.

[10] DARPA. 2010 (Sept). *Request for Information (RFI), DARPA-SN-10-73, New Technologies to Support Declassification*. Online http://fas.org/sgp/news/2010/09/darpa-declass.pdf. last accessed - 15/04/2016.

[11] Dehghan, Azad, Kovacevic, Aleksandar, Karystianis, George, Keane, John A., & Nenadic, Goran. 2015. Combining knowledge- and data-driven methods for de-identification of clinical narratives. *Journal of Biomedical Informatics*, **58, Supplement**, S53–S59. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

[12] Dias, Francisco, Mamede, Nuno, & Baptista, Jorge. 2016 (July). Automated Anonymization of Text Documents. *In: Proceedings of the 2016 World Congress conference on Advances in Computational Intelligence*.

[13] Douglass, Margaret M, Clifford, Gari D, Reisner, Andrew, Moody, George B, & Mark, Roger G. 2004 (Sept). Computer-Assisted De-identification of Free Text in the MIMIC II Database. *Pages 341–344 of: Computers in Cardiology*.

[14] Ehrmann, Maud, Turchi, Marco, & Steinberger, Ralf. 2011. Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection. *Pages 118–124 of: Recent Advances in Natural Language Processing*.

[15] European Commission. 2012 (January). *Proposal for a General Data Protection Regulation. Online:* http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012PC0011&from=EN. last accessed - 15/04/2016.

[16] Faruqui, Manaal, & Padó, Sebastian. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. *In: Proceedings of KONVENS 2010*.

[17] Ferrández, Oscar, South, Brett R, Shen, Shuying, Friedlin, F Jeffrey, Samore, Matthew H, & Meystre, Stéphane M. 2013. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *J. Am. Med. Inform. Assoc.*, **20**(1), 77–83.

[18] Finkel, Jenny Rose, Grenager, Trond, & Manning, Christopher. 2005. Incorporating Non-local Information into Information. Extraction Systems by Gibbs Sampling. *Pages 363–370 of: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

[19] Fleiss, Joseph L. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, **76**(5), 378–382.

[20] Freelon, Deen G. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, **5**(1), 20–33.

[21] GARDNER, JAMES, & XIONG, LI. 2008. HIDE: An Integrated System for Health Information DE-identification. *Pages 254–259 of: Computer-Based Medical Systems*. IEEE Computer Society.

[22] GOLDBERGER, ARY L., AMARAL, LUIS A. N., GLASS, LEON, HAUSDORFF, JEFFREY M., IVANOV, PLAMEN CH., MARK, ROGER G., MIETUS, JOSEPH E., MOODY, GEORGE B., PENG, CHUNG-KANG, & STANLEY, H. EUGENE. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, **101**(23), 215–220.

[23] HAJLAOUI, NAJEH, KOLOVRATNIK, DAVID, VÄYRYNEN, JAAKKO, STEINBERGER, RALF, & VARGA, DANIEL. 2014. DCEP - Digital Corpus of the European Parliament. *Pages 3164–3171 of: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

[24] HARA, KAZUO. 2006. Applying a SVM based chunker and a text classifier to the Deid Challenge. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 10–11.

[25] KOEHN, PHILIPP, HOANG, HIEU, BIRCH, ALEXANDRA, CALLISON-BURCH, CHRIS, FEDERICO, MARCELLO, BERTOLDI, NICOLA, COWAN, BROOKE, SHEN, WADE, MORAN, CHRISTINE, ZENS, RICHARD, DYER, CHRIS, BOJAR, ONDŘEJ, CONSTANTIN, ALEXANDRA, & HERBST, EVAN. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Pages 177–180 of: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics.

[26] LAFFERTY, JOHN D., MCCALLUM, ANDREW, & PEREIRA, FERNANDO C. N. 2001 (June). Conditional random fields: probabilistic models for segmenting and sequence data. *Pages 282–289 of:* DANYLUK, AP. (ed), *Proceedings of International Conference on Machine Learning*.

[27] LANDIS, J. RICHARD, & KOCH, GARY G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

[28] LIU, ZENGJIAN, CHEN, YANGXIN, TANG, BUZHOU, WANG, XIAOLONG, CHEN, QINGCAI, LI, HAODI, WANG, JINGFENG, DENG, QIWEN, & ZHU, SUISONG. 2015. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of Biomedical Informatics*, **58, Supplement**, S47–S52. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

[29] MAMEDE, NUNO, BAPTISTA, JORGE, DINIZ, CLÁUDIO, & CABARRÃO, VERA. 2012. STRING: A hybrid statistical and rule-based natural language processing chain for Portuguese. *Online:* `http://www.inesc-id.pt/pt/indicadores/Ficheiros/8578.pdf`. last accessed - 15/04/2016.

[30] MARQUES, JOÃO SILVESTRE. 2013 (November). *Anaphora Resolution in Portuguese: an hybrid approach*. M.Phil. thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[31] Metathesaurus, UMLS. *Online:* https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/. last accessed - 15/04/2016.

[32] Meystre, Stephane M, Friedlin, F Jeffrey, South, Brett R, Shen, Shuying, & Samore, Matthew H. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, **10**(1), 70–86.

[33] Meystre, Stéphane M., Óscar Ferrández, Friedlin, F. Jeffrey, South, Brett R., Shen, Shuying, & Samore, Matthew H. 2014. Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*, **50**, 142–150.

[34] Miller, George, & Fellbaum, Christiane. 1998 (May). *Wordnet: An electronic lexical database*. Cambridge, MA, USA, MIT Press.

[35] Naber, Daniel. *JWordSplitter. Online:* http://www.danielnaber.de/jwordsplitter/. last accessed - 15/04/2016.

[36] Neamatullah, Ishna, Douglass, Margaret M, wei H Lehman, Li, Reisner, Andrew, Villarroel, Mauricio, Long, William J, Szolovits, Peter, Moody, George B, Mark, Roger G, & Clifford, Gari D. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, **8**(1), 1–17.

[37] Noreen, Eric W. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York, NY, USA: John Wiley & Sons.

[38] Nothman, Joel, Ringland, Nicky, Radford, Will, Murphy, Tara, & Curran, James R. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, **194**(Jan.), 151–175.

[39] Object Management Group (OMG). *Unified Modelling Language. Online:* http://www.uml.org/. last accessed - 15/04/2016.

[40] Oliveira, Diogo Correia de. 2010 (Nov). *Extraction and Classification of Named Entities*. M.Phil. thesis, Instituto Superior Técnico.

[41] Porter, Martin. *Snowball German Stemming Algorithm. Online:* http://snowballstem.org/. last accessed - 15/04/2016.

[42] PostgreSQL. *Online:* http://www.postgresql.org/. last accessed - 15/04/2016.

[43] Python. *Online:* http://www.python.org/. last accessed - 15/04/2016.

[44] Quinlan, John R. 2014. *C4.5: programs for machine learning*. Elsevier.

[45] Ruch, Patrick, Baud, Robert H, Rassinoux, Anne-Marie, Bouillon, Pierrette, & Robert, Gilbert. 2000. Medical document anonymization with a semantic lexicon. *Proc. AMIA Symp.*, 729–733.

[46] Samarati, Pierangela, & Sweeney, Latanya. 1998. *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression*. Tech. rept. Computer Science Laboratory, SRI International.

[47] Sánchez, David, Batet, Montserrat, & Viejo, Alexandre. 2013. Automatic General-Purpose Sanitization of Textual Documents. *IEEE Transactions on Information Forensics and Security*, **8**(6), 853–862.

[48] Sarmento, Luís, Pinto, Ana Sofia, & Cabral, Luís. 2006. *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg. Chap. REPENTINO – A Wide-Scope Gazetteer for Entity Recognition in Portuguese, pages 31–40.

[49] Schapire, Robert E. 1990. The strength of weak learnability. *Machine learning*, **5**(2), 197–227.

[50] Stubbs, Amber, Kotfila, Christopher, & Özlem Uzuner. 2015. Automated Systems for the De-identification of Longitudinal Clinical Narratives. *J. of Biomedical Informatics*, **58**(S), S11–S19.

[51] Sweeney, Latanya. 1996. Replacing Personally-Identifying Information in Medical Records, the Scrub System. *Journal of the American Medical Informatics Association*, 333–337.

[52] Szarvas, György, Farkasb, Richárd, & Busa-Feketeb, Róbert. 2007. State-of-the-Art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *Pages 574–580 of:* JAMIA (ed), *Journal of the American Medical Informatics Association*, vol. 14.

[53] Taulé, Mariona, Martí, Maria Antònia, & Recasens, Marta. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). *Online:* http://www.lrec-conf.org/proceedings/lrec2008/.

[54] Tiedemann, Jörg. 2012 (May). Parallel Data, Tools and Interfaces in OPUS. *Pages 2214–2218 of: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

[55] Tjong Kim Sang, Erik F. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *Pages 155–158 of: Proceedings of CoNLL-2002*. Taipei, Taiwan.

[56] Tjong Kim Sang, Erik F., & De Meulder, Fien. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Pages 142–147 of: Proceedings of CoNLL-2003*, vol. 4. Edmonton, Canada: Association for Computational Linguistics. Stroudsburg, PA, USA.

[57] Turchi, Marco, Atkinson, Martin, Wilcox, Alastair, Crawley, Brett, Bucci, Stefano, Steinberger, Ralf, & van der Goot, Erik. 2012. ONTS: "Optima" News Translation System. *Pages 25–30 of: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

[58] U.S. Department of Health & Human Services. 2008. 45 C.F.R. § 46 Protection of Human Subjects. October, 2008. *Online:* http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html. last accessed - 15/04/2016.

[59] Uzuner, Özlem, Luo, Yuan, & Szolovits, Peter. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association : JAMIA*, **14**(5), 550–563.

[60] Uzuner, Özlem, Sibanda, Tawanda C, Luo, Yuan, & Szolovits, Peter. 2008. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine*, **42**(1), 13–35.

[61] Vrandečić, Denny. 2012. Wikidata: A New Platform for Collaborative Data Collection. *Pages 1063–1064 of: Proceedings of the 21st International Conference Companion on World Wide Web*.

[62] Wellner, Ben, Huyck, Matt, Mardis, Scott, Aberdeen, John, Morgan, Alex, Peshkin, Leonid, Yeh, Alex, Hitzeman, Janet, & Hirschman, Lynette. 2007. Rapidly Retargetable Approaches to De-identification in Medical Records. *Journal of the American Medical Informatics Association*, **14**(5), 564–573.

[63] Wiktionary. *Online:* https://www.wiktionary.org. last accessed - 15/04/2016.

[64] Yang, Hui, & Garibaldi, Jonathan M. 2015. Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, **58, Supplement**, S30–S38. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

[65] Zwicky, Arnold M. 1985. Heads. *Journal of linguistics*, **21**(01), 1–29.

A UML Diagrams

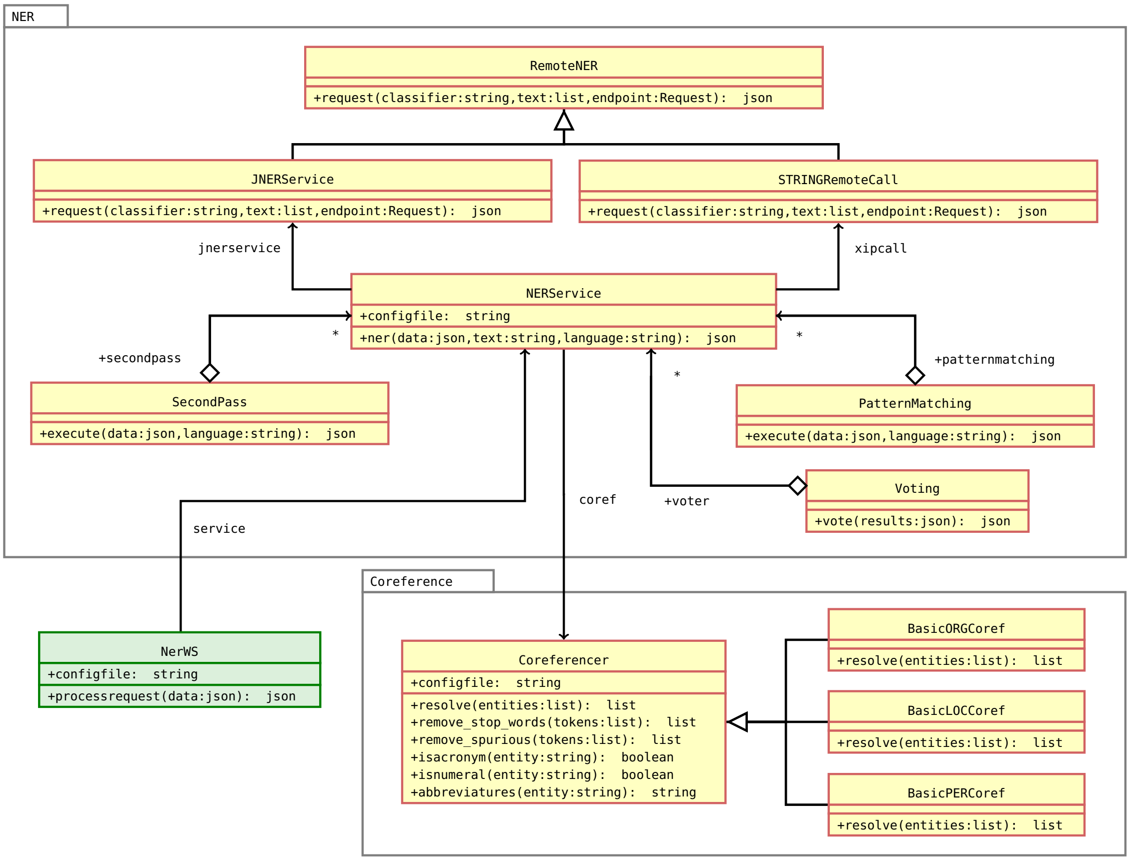

NER
RemoteNER
+request(classifier:string,text:list,endpoint:Request): json
JNERService
+request(classifier:string,text:list,endpoint:Request): json
STRINGRemoteCall
+request(classifier:string,text:list,endpoint:Request): json
jnerservice
xipcall
NERService
+configfile: string
+ner(data:json,text:string,language:string): json
*
*
*
+secondpass
+patternmatching
SecondPass
+execute(data:json,language:string): json
PatternMatching
+execute(data:json,language:string): json
Voting
+vote(results:json): json
coref
+voter
service
Coreference
NerWS
+configfile: string
+processrequest(data:json): json
Coreferencer
+configfile: string
+resolve(entities:list): list
+remove_stop_words(tokens:list): list
+remove_spurious(tokens:list): list
+isacronym(entity:string): boolean
+isnumeral(entity:string): boolean
+abbreviatures(entity:string): string
BasicORGCoref
+resolve(entities:list): list
BasicLOCCoref
+resolve(entities:list): list
BasicPERCoref
+resolve(entities:list): list

Figure A.1: UML diagram of the NER service.

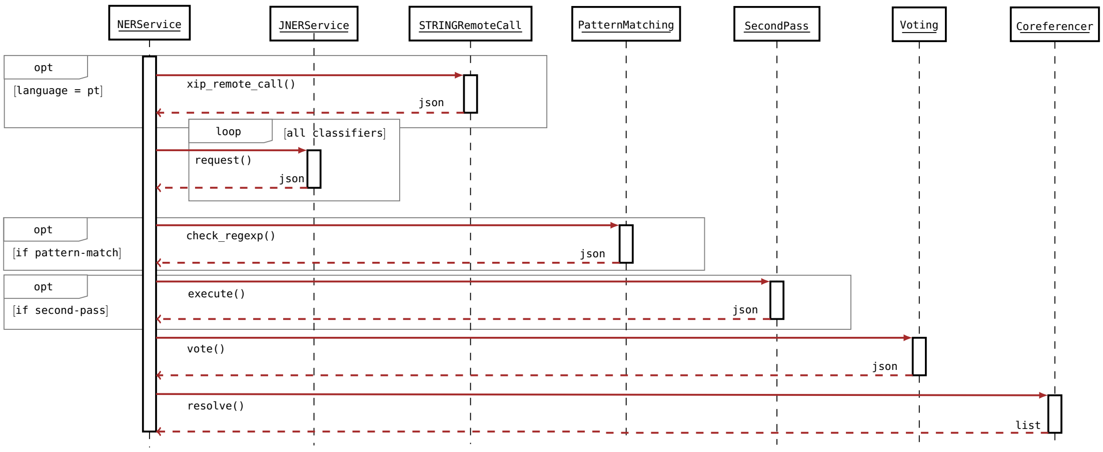


Figure A.2: Sequence diagram of the NER service.

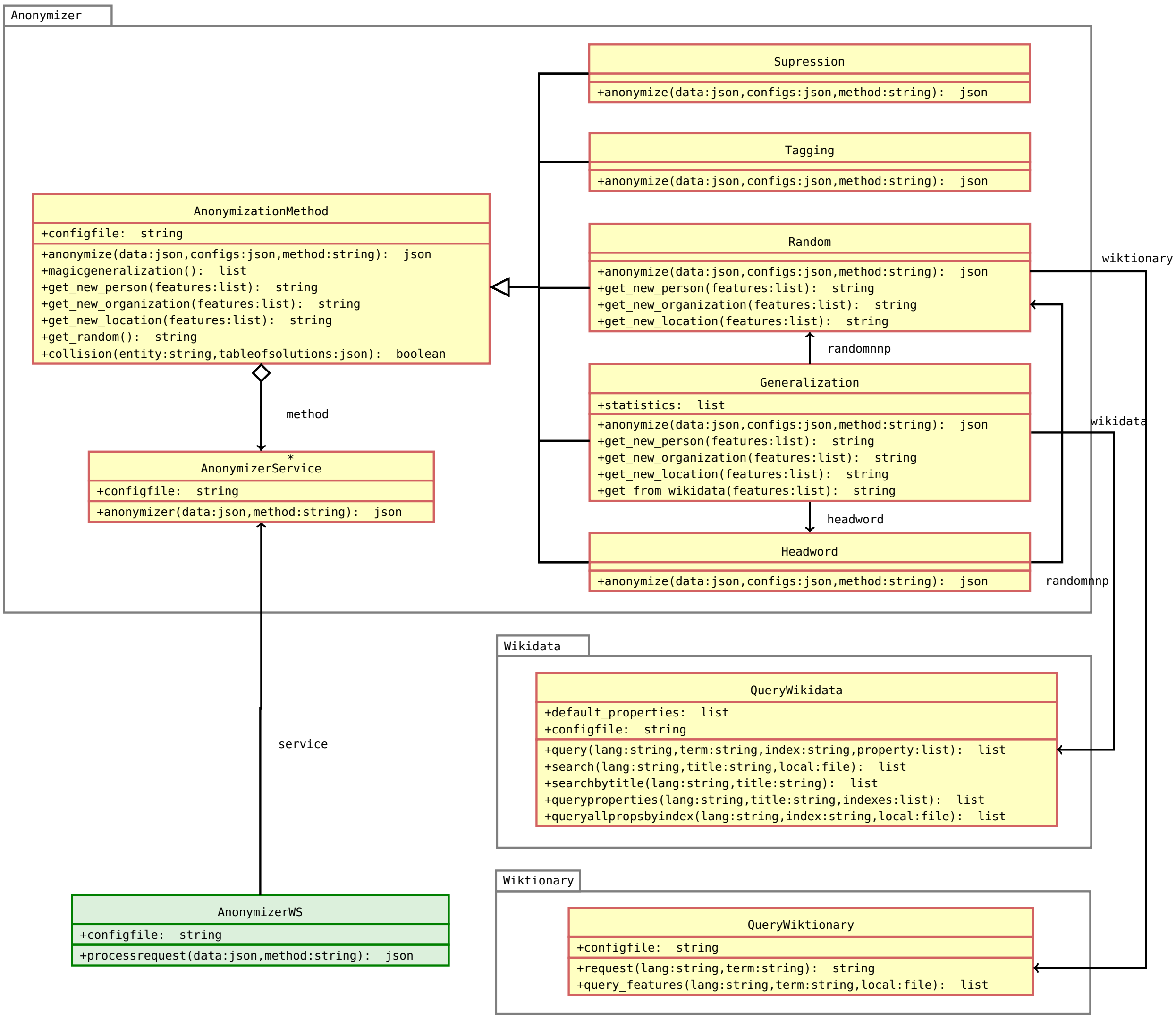

Anonymizer
Supression
+anonymize(data:json,configs:json,method:string): json
Tagging
+anonymize(data:json,configs:json,method:string): json
AnonymizationMethod
+configfile: string
+anonymize(data:json,configs:json,method:string): json
+magicgeneralization(): list
+get_new_person(features:list): string
+get_new_organization(features:list): string
+get_new_location(features:list): string
+get_random(): string
+collision(entity:string,tableofsolutions:json): boolean
Random
+anonymize(data:json,configs:json,method:string): json
+get_new_person(features:list): string
+get_new_organization(features:list): string
+get_new_location(features:list): string
wiktionary
randomnnp
Generalization
+statistics: list
+anonymize(data:json,configs:json,method:string): json
+get_new_person(features:list): string
+get_new_organization(features:list): string
+get_new_location(features:list): string
+get_from_wikidata(features:list): string
wikidata
method
*
AnonymizerService
+configfile: string
+anonymizer(data:json,method:string): json
headword
Headword
+anonymize(data:json,configs:json,method:string): json
randomnnp
Wikidata
QueryWikidata
+default_properties: list
+configfile: string
+query(lang:string,term:string,index:string,property:list): list
+search(lang:string,title:string,local:file): list
+searchbytitle(lang:string,title:string): list
+queryproperties(lang:string,title:string,indexes:list): list
+queryallpropsbyindex(lang:string,index:string,local:file): list
service
Wiktionary
AnonymizerWS
+configfile: string
+processrequest(data:json,method:string): json
QueryWiktionary
+configfile: string
+request(lang:string,term:string): string
+query_features(lang:string,term:string,local:file): list

Figure A.3: UML diagram of the Anonymizer service.

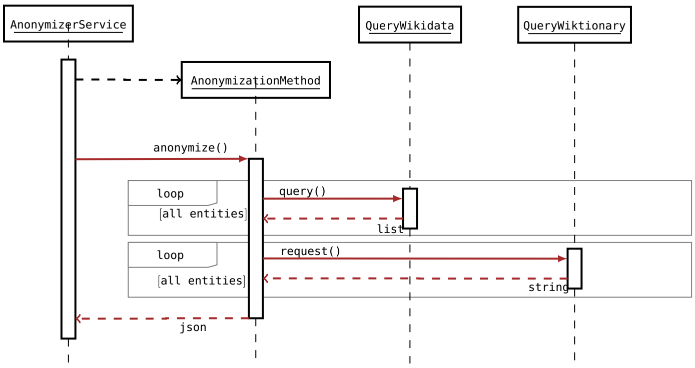


Figure A.4: Sequence diagram of the anonymizer service.

B JSON Schemas

B.1 Configuration File

```
{
  [LANGUAGE]: {
     "ner": {
        [MAIN NER CLASSIFIER]
        [PATTERN MATCHING]
        [PARALLEL NER CLASSIFIERS]
        [SECOND PASS]
        [COREFERENCE]
        [VOTING]
    },
     "anon": {
        [METHODS]
        [LABELS]
        [DELIMITERS]
        [OUTPUT]
    }
  }
}
```

Figure B.1: Overall schema of the configuration file.

- `[LANGUAGE]` - Unique code to identify a language in the system. We use the ISO 639-1 codes for that purpose: *“de”, “en”, “es”, “pt”*.
- `[MAIN NER CLASSIFIER]` - Definition of the main classifier of the NER module. The following JSON structure defines the classifier:

  ```
  "nerservice": [TYPE]
  "localner": { "model": [PATH MODEL], "classifier": [PATH CLASSIFIER] }
  "remotener": { "url": [URL CLASSIFIER], "classifier": [CLASSIFIER NAME] }
  ```

  The type of classifier (TYPE) is an attribute that defines if the main NER classifier. If the attribute is “*local*”, then the structure “*localner*” must define the paths to a CRF model and a classifier. If the attribute is “*remote*”, then the structure “*remotener*” must define the url to the NER webservice and the name of the classifier.
- `[PATTERN MATCHING]` - Lists of dictionaries and regular expression files that define entities.

  ```
  "dictionaries": [path1, path2, ...]
  "regexp": [path1, path2, ...]
  ```

  Dictionaries are defined in the attribute “*dictionaries*” as a list of absolute paths to the files. Regular expression files are defined in the attribute “*regexp*” as a list of absolute paths to the files.
- `[PARALLEL NER CLASSIFIERS]` - Definition of parallel NER classifiers in the key “*parallel*”.

```
"parallel": {
  "state": 1,
  "paralleltags": ["ORGZ", "PERZ", "LOCZ"],
  "classifiers": [ { "url": "http://localhost:1234/ner", "classifier": "esp" } ]
}
```

The attribute *"state"* switches on and off the use of parallel classifiers. The attribute *"paralleltags"* is the tagset of the parallel classifiers. The attribute *"classifiers"* is a list of JSON structures containing the *"url"* of a classifier and the *"classifier"* name.

- `[SECOND PASS]` - The second-pass dection module is activated by the boolean flag in the key *"secondpass"*.

  ```
  "secondpass": 0
  ```

- `[COREFERENCE]` - The key *"coreference"* contains a JSON structure with the attribution of the classes to process the coreference.

  ```
  "coreference": {
    "PERZ": "default_person.DefaultPersonCoreferencer",
    "ORGZ": "default_organization.DefaultOrganizationCoreferencer",
    "LOCZ": "simple_coref.SimpleCoreferencer"
  }
  ```

  This structure contains a JSON dictionary whose keys are NE class tags and the entries are CRR implementations.

- `[VOTING]` - The key *"voting"* assigns a class that implements the voting module. This class must extend the Voting abstract class.
- `[METHODS]` - The key *"methods"* contains a JSON dictionary whose keys are NE class tags and the entries are anonymization methods. This strucutre assigns the anonymization method to be applied to each NE class.
- `[LABELS]` - Contains a JSON dictionary that assigns convertion rules from the tagset of the NER modules as the tagset from our system.

  ```
  "converttags": {
    "oldtag1": "LOCZ",
    "oldtag2": "PERZ",
    "oldtag3": "ORGZ",
    "UNK": "O"
  }
  ```

  A tag *"UNK"* is catchall tag for converting all unknown tags that were not assigned in this dictionary.

- `[DELIMITERS]` - The key *"delimiters"* defines a pair of placeholders to surround an anonymized entity.
- `[OUTPUT]` - The key *"output"* defines the type of output from the anonymization module. It can be *"json"* (default value) or *"plaintext"*.

B.2 JSON Job

```
{
  "text":     [TEXT],
  "language": [LANGUAGE],
  "method": [ANON METHOD],
  "ner": {
     "ner_time":  [TIMESTAMP],
     "text": [PROCESSED TEXT],
     "entries": [
         {
            "entity_class": [CLASS],
            "begin": [START],
            "end": [END],
            "uid": [UNIQUE ID],
            "surface": [ENTITY SURFACE]
         },
         (...)
     ]
  },
  "coref": [MENTIONS],
  "anonymizer": {
     "anon_time": [TIMESTAMP],
     "output": [ANONYMIZED TEXT],
     "tableofsolutions": {
         [UNIQUE ID] : {
            "entity_class": [CLASS],
            "coref": [MENTION ID],
            "original": [ORIGINAL SURFACE],
            "replace": [REPLACEMENT]
         },
         (...)
     }
  }
}
```

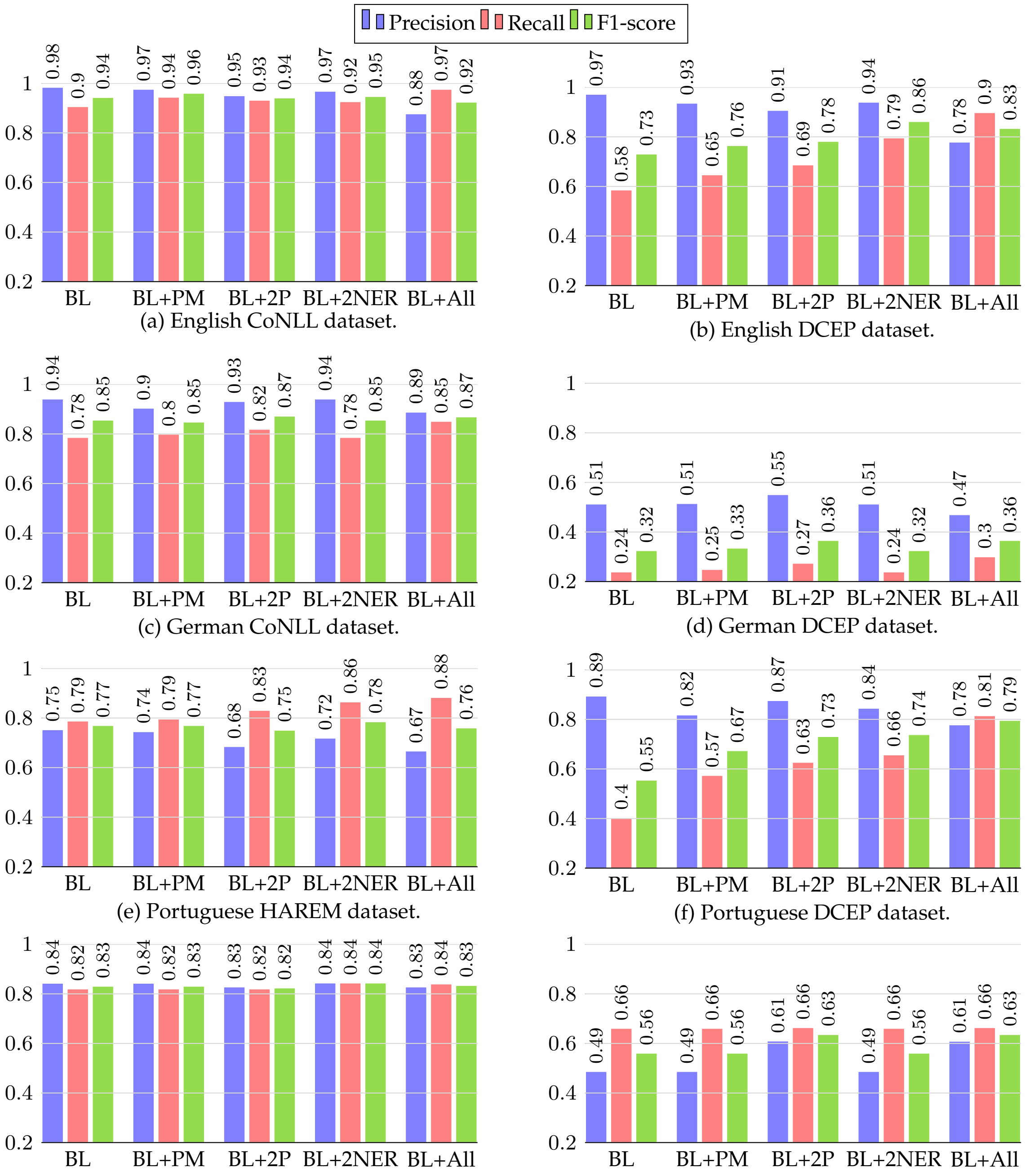
Figure B.2: Overall schema of the JSON job.

- `[TEXT]` - Normalized text to be anonymized.
- `[LANGUAGE]` - The language of the text. This language determines the configuration and modules that are used during the detection and anonymization.
- `[ANON METHOD]` - Method of anonymization to be applied to the text. This tag is registered in the configuration file and, by default, it is one of the following tags: *"DEL"*, *"TAG"*, *"RAN"* and *"GEN"*.
- `[COREF]` - List of mentions. Each mention is a list of *unique ids* of entities from the text.
  Example: `[[0, 1] [2, 3], [4]]`
- `[TIMESTAMP]` - Time (ms) taken to perform an action.
- `[PROCESSED TEXT]` - Text after being processed by the NER module (in the case that the NER module applies pre-processing to the text, otherwise is similar to `[TEXT]`).
- `[CLASS]` - Class of an entity. By default, it is one of the following tags: *"LOCZ"*, *"ORGZ"*, *"PERZ"* or *"MIZC"*.
- `[START]` - Offset to the beginning of the entity.
- `[END]` - Offset to the end of the entity.

- `[UNIQUE ID]` - Unique number that identifies an entity after the detection.
- `[ENTITY SURFACE]` - Surface of the entity.
- `[ANONYMIZED TEXT]` - Plain text after being applied the anonymization method.
- `[MENTION ID]` - Index of the mention inside `[COREF]` where this entity belongs.
- `[ORIGINAL SURFACE]` - Surface of the entity.
- `[REPLACEMENT]` - Surface of the replacement entity.

C Results

C.1 NER Binary Performance



(a) English CoNLL dataset.

(b) English DCEP dataset.

(c) German CoNLL dataset.

(d) German DCEP dataset.

(e) Portuguese HAREM dataset.

(f) Portuguese DCEP dataset.

(g) Spanish CoNLL dataset.

(h) Spanish DCEP dataset.

Figure C.1: NER binary performance.

C.2 NER Performance in English texts

			CoNLL			DCEP		
			Pr	Re	F1	Pr	Re	F1
BINARY		NER (baseline)	0.982	0.904	0.941	0.970	0.584	0.729
		NER + pattern-matching	0.974	0.942	0.958	0.934	0.645	0.763
		NER + second-pass	0.948	0.930	0.939	0.905	0.685	0.780
		NER + 2nd NER in parallel	0.966	0.924	0.945	0.938	0.794	0.860
		NER All	0.875	0.974	0.922	0.777	0.896	0.832
MACRO AVERAGED		NER (baseline)	0.936	0.885	0.905	0.936	0.600	0.715
		NER + pattern-matching	0.924	0.924	0.922	0.898	0.659	0.749
		NER + second-pass	0.886	0.918	0.894	0.869	0.688	0.758
		NER + 2nd NER in parallel	0.910	0.906	0.903	0.911	0.777	0.832
		NER All	0.867	0.936	0.894	0.852	0.847	0.845
TOKEN LEVEL	OVERALL	NER (baseline)	0.943	0.900	0.921	0.943	0.576	0.715
		NER + pattern-matching	0.932	0.940	0.936	0.904	0.638	0.748
		NER + second-pass	0.899	0.926	0.912	0.872	0.677	0.762
		NER + 2nd NER in parallel	0.919	0.919	0.919	0.912	0.789	0.846
		NER All	0.885	0.944	0.913	0.851	0.867	0.859
	PERSON	NER (baseline)	0.965	0.945	0.955	1.000	0.765	0.867
		NER + pattern-matching	0.963	0.965	0.964	1.000	0.782	0.878
		NER + second-pass	0.944	0.953	0.948	1.000	0.800	0.889
		NER + 2nd NER in parallel	0.926	0.949	0.938	0.980	0.814	0.889
		NER All	0.911	0.959	0.935	0.970	0.852	0.907
	LOCATION	NER (baseline)	0.918	0.936	0.927	0.851	0.811	0.830
		NER + pattern-matching	0.900	0.956	0.927	0.840	0.862	0.851
		NER + second-pass	0.862	0.945	0.902	0.813	0.868	0.840
		NER + 2nd NER in parallel	0.908	0.958	0.932	0.858	0.931	0.893
		NER All	0.870	0.970	0.917	0.800	0.957	0.872
	ORGANIZ.	NER (baseline)	0.944	0.798	0.865	0.955	0.485	0.643
		NER + pattern-matching	0.929	0.887	0.907	0.896	0.560	0.689
		NER + second-pass	0.881	0.866	0.874	0.855	0.611	0.712
		NER + 2nd NER in parallel	0.924	0.833	0.876	0.909	0.755	0.825
		NER All	0.866	0.891	0.878	0.838	0.853	0.846
ENTITY LEVEL	OVERALL	NER (baseline)	0.941	0.880	0.910	0.823	0.595	0.691
		NER + pattern-matching	0.926	0.910	0.918	0.769	0.684	0.724
		NER + second-pass	0.897	0.893	0.895	0.746	0.690	0.717
		NER + 2nd NER in parallel	0.922	0.889	0.905	0.832	0.708	0.765
		NER All	0.886	0.903	0.894	0.753	0.794	0.773
	PERSON	NER (baseline)	0.971	0.928	0.949	0.847	0.686	0.758
		NER + pattern-matching	0.961	0.938	0.950	0.837	0.686	0.754
		NER + second-pass	0.972	0.938	0.955	0.820	0.695	0.753
		NER + 2nd NER in parallel	0.944	0.931	0.937	0.864	0.724	0.788
		NER All	0.941	0.945	0.943	0.837	0.733	0.782
	LOCATION	NER (baseline)	0.947	0.936	0.942	0.866	0.846	0.856
		NER + pattern-matching	0.935	0.950	0.942	0.853	0.846	0.849
		NER + second-pass	0.902	0.939	0.920	0.836	0.862	0.848
		NER + 2nd NER in parallel	0.934	0.944	0.939	0.930	0.923	0.927
		NER All	0.898	0.950	0.923	0.878	0.938	0.907
	ORGANIZ.	NER (baseline)	0.894	0.749	0.815	0.796	0.498	0.612
		NER + pattern-matching	0.873	0.822	0.847	0.724	0.635	0.677
		NER + second-pass	0.807	0.776	0.791	0.697	0.637	0.666
		NER + 2nd NER in parallel	0.876	0.764	0.816	0.787	0.639	0.705
		NER All	0.804	0.792	0.798	0.699	0.765	0.730

Table C.1: NER performance in English datasets.

C.3 NER Performance in German texts

			CoNLL			**DCEP**		
			Pr	Re	F1	Pr	Re	F1
BINARY		NER (baseline)	0.939	0.784	0.854	0.511	0.237	0.323
		NER + pattern-matching	0.902	0.797	0.846	0.513	0.247	0.333
		NER + second-pass	0.929	0.817	0.870	0.549	0.272	0.364
		NER + 2nd NER in parallel	0.939	0.784	0.854	0.511	0.237	0.323
		NER All	0.886	0.849	0.867	0.468	0.298	0.364
MACRO AVERAGED		NER (baseline)	0.789	0.740	0.745	0.288	0.152	0.198
		NER + pattern-matching	0.756	0.759	0.738	0.306	0.161	0.209
		NER + second-pass	0.770	0.767	0.746	0.299	0.171	0.216
		NER + 2nd NER in parallel	0.789	0.740	0.745	0.288	0.152	0.198
		NER All	0.719	0.805	0.743	0.321	0.203	0.246
TOKEN LEVEL	**OVERALL**	NER (baseline)	0.816	0.759	0.787	0.500	0.233	0.317
		NER + pattern-matching	0.786	0.774	0.780	0.503	0.243	0.328
		NER + second-pass	0.795	0.793	0.794	0.544	0.271	0.361
		NER + 2nd NER in parallel	0.799	0.791	0.795	0.432	0.254	0.320
		NER All	0.753	0.827	0.788	0.460	0.295	0.359
	PERSON	NER (baseline)	0.779	0.632	0.698	0.889	0.923	0.906
		NER + pattern-matching	0.652	0.632	0.642	0.857	0.923	0.889
		NER + second-pass	0.775	0.674	0.721	0.857	0.923	0.889
		NER + 2nd NER in parallel	0.736	0.674	0.703	0.857	0.923	0.889
		NER All	0.625	0.707	0.663	0.828	0.923	0.873
	LOCATION	NER (baseline)	0.762	0.816	0.788	0.907	0.722	0.804
		NER + pattern-matching	0.764	0.857	0.808	0.911	0.759	0.828
		NER + second-pass	0.748	0.833	0.788	0.943	0.909	0.926
		NER + 2nd NER in parallel	0.748	0.847	0.794	0.915	0.811	0.860
		NER All	0.744	0.906	0.817	0.946	0.981	0.964
	ORGANIZ.	NER (baseline)	0.860	0.790	0.824	0.245	0.088	0.129
		NER + pattern-matching	0.862	0.800	0.830	0.254	0.094	0.138
		NER + second-pass	0.827	0.827	0.827	0.277	0.101	0.148
		NER + 2nd NER in parallel	0.857	0.820	0.838	0.204	0.101	0.135
		NER All	0.823	0.843	0.833	0.227	0.121	0.157
ENTITY LEVEL	**OVERALL**	NER (baseline)	0.804	0.654	0.722	0.360	0.188	0.247
		NER + pattern-matching	0.760	0.674	0.715	0.361	0.195	0.254
		NER + second-pass	0.778	0.667	0.718	0.414	0.226	0.292
		NER + 2nd NER in parallel	0.804	0.654	0.722	0.360	0.188	0.247
		NER All	0.728	0.701	0.715	0.354	0.240	0.286
	PERSON	NER (baseline)	0.800	0.526	0.635	0.917	0.846	0.880
		NER + pattern-matching	0.615	0.526	0.567	0.846	0.846	0.846
		NER + second-pass	0.764	0.553	0.641	0.846	0.846	0.846
		NER + 2nd NER in parallel	0.800	0.526	0.635	0.917	0.846	0.880
		NER All	0.616	0.592	0.604	0.846	0.846	0.846
	LOCATION	NER (baseline)	0.866	0.755	0.807	0.714	0.641	0.676
		NER + pattern-matching	0.862	0.798	0.829	0.722	0.667	0.693
		NER + second-pass	0.845	0.755	0.798	0.944	0.872	0.907
		NER + 2nd NER in parallel	0.866	0.755	0.807	0.714	0.641	0.676
		NER All	0.815	0.824	0.820	0.900	0.923	0.911
	ORGANIZ.	NER (baseline)	0.761	0.656	0.705	0.152	0.065	0.092
		NER + pattern-matching	0.765	0.672	0.715	0.158	0.070	0.097
		NER + second-pass	0.737	0.669	0.702	0.156	0.070	0.097
		NER + 2nd NER in parallel	0.761	0.656	0.705	0.152	0.065	0.092
		NER All	0.730	0.679	0.704	0.128	0.076	0.096

Table C.2: NER performance in German datasets.

C.4 NER Performance in Portuguese texts

			CoNLL			**DCEP**		
			Pr	Re	F1	Pr	Re	F1
BINARY		NER (baseline)	0.751	0.786	0.768	0.892	0.401	0.553
		NER + pattern-matching	0.743	0.794	0.768	0.816	0.572	0.672
		NER + second-pass	0.683	0.829	0.749	0.874	0.625	0.729
		NER + 2nd NER in parallel	0.717	0.863	0.783	0.843	0.655	0.737
		NER All	0.665	0.881	0.758	0.776	0.813	0.794
MACRO AVERAGED		NER (baseline)	0.630	0.758	0.678	0.755	0.342	0.457
		NER + pattern-matching	0.624	0.765	0.678	0.739	0.520	0.600
		NER + second-pass	0.569	0.806	0.654	0.773	0.564	0.634
		NER + 2nd NER in parallel	0.586	0.839	0.678	0.793	0.579	0.648
		NER All	0.547	0.860	0.655	0.726	0.768	0.733
TOKEN LEVEL	**OVERALL**	NER (baseline)	0.620	0.752	0.680	0.810	0.378	0.516
		NER + pattern-matching	0.614	0.761	0.679	0.757	0.553	0.639
		NER + second-pass	0.557	0.798	0.656	0.808	0.607	0.693
		NER + 2nd NER in parallel	0.580	0.832	0.683	0.764	0.632	0.692
		NER All	0.534	0.856	0.658	0.709	0.799	0.752
	PERSON	NER (baseline)	0.734	0.827	0.778	0.756	0.506	0.606
		NER + pattern-matching	0.733	0.832	0.779	0.756	0.517	0.614
		NER + second-pass	0.659	0.854	0.744	0.637	0.522	0.574
		NER + 2nd NER in parallel	0.696	0.875	0.775	0.756	0.604	0.672
		NER All	0.646	0.892	0.749	0.727	0.660	0.691
	LOCATION	NER (baseline)	0.559	0.771	0.648	0.654	0.989	0.788
		NER + pattern-matching	0.541	0.779	0.638	0.422	0.989	0.591
		NER + second-pass	0.506	0.828	0.628	0.614	0.989	0.757
		NER + 2nd NER in parallel	0.535	0.865	0.661	0.654	1.000	0.791
		NER All	0.478	0.882	0.620	0.396	1.000	0.567
	ORGANIZ.	NER (baseline)	0.387	0.474	0.426	0.899	0.296	0.445
		NER + pattern-matching	0.392	0.494	0.437	0.886	0.518	0.654
		NER + second-pass	0.349	0.548	0.426	0.896	0.585	0.708
		NER + 2nd NER in parallel	0.365	0.637	0.464	0.787	0.602	0.682
		NER All	0.352	0.690	0.466	0.781	0.801	0.791
ENTITY LEVEL	**OVERALL**	NER (baseline)	0.597	0.621	0.609	0.648	0.191	0.296
		NER + pattern-matching	0.582	0.626	0.603	0.601	0.401	0.481
		NER + second-pass	0.536	0.649	0.587	0.755	0.444	0.559
		NER + 2nd NER in parallel	0.553	0.668	0.605	0.665	0.355	0.463
		NER All	0.484	0.680	0.565	0.618	0.648	0.633
	PERSON	NER (baseline)	0.649	0.700	0.674	0.395	0.254	0.309
		NER + pattern-matching	0.642	0.700	0.670	0.395	0.254	0.309
		NER + second-pass	0.587	0.728	0.650	0.375	0.305	0.336
		NER + 2nd NER in parallel	0.612	0.724	0.664	0.366	0.254	0.300
		NER All	0.551	0.734	0.629	0.360	0.305	0.330
	LOCATION	NER (baseline)	0.631	0.685	0.657	0.804	0.953	0.872
		NER + pattern-matching	0.599	0.692	0.642	0.333	0.953	0.494
		NER + second-pass	0.576	0.716	0.638	0.731	0.950	0.826
		NER + 2nd NER in parallel	0.583	0.741	0.652	0.788	0.953	0.863
		NER All	0.487	0.748	0.590	0.318	0.953	0.477
	ORGANIZ.	NER (baseline)	0.384	0.340	0.361	0.667	0.119	0.202
		NER + pattern-matching	0.382	0.352	0.366	0.770	0.371	0.501
		NER + second-pass	0.332	0.361	0.346	0.833	0.420	0.558
		NER + 2nd NER in parallel	0.363	0.422	0.390	0.690	0.315	0.433
		NER All	0.327	0.448	0.378	0.731	0.663	0.695

Table C.3: NER performance in Portuguese datasets.

C.5 NER Performance in Spanish texts

			CoNLL			DCEP		
			Pr	Re	F1	Pr	Re	F1
BINARY		NER (baseline)	0.841	0.818	0.829	0.485	0.659	0.559
		NER + pattern-matching	0.841	0.818	0.829	0.485	0.659	0.559
		NER + second-pass	0.826	0.818	0.822	0.608	0.662	0.634
		NER + 2nd NER in parallel	0.842	0.842	0.842	0.485	0.659	0.559
		NER All	0.826	0.838	0.832	0.607	0.662	0.634
MACRO AVERAGED		NER (baseline)	0.737	0.794	0.755	0.450	0.665	0.527
		NER + pattern-matching	0.737	0.794	0.755	0.450	0.665	0.527
		NER + second-pass	0.726	0.796	0.750	0.558	0.667	0.597
		NER + 2nd NER in parallel	0.739	0.828	0.771	0.450	0.665	0.527
		NER All	0.726	0.827	0.762	0.558	0.667	0.597
TOKEN LEVEL	OVERALL	NER (baseline)	0.712	0.791	0.750	0.445	0.639	0.525
		NER + pattern-matching	0.712	0.791	0.750	0.445	0.639	0.525
		NER + second-pass	0.704	0.793	0.745	0.558	0.643	0.597
		NER + 2nd NER in parallel	0.712	0.817	0.761	0.445	0.639	0.525
		NER All	0.702	0.815	0.754	0.557	0.643	0.597
	PERSON	NER (baseline)	0.907	0.925	0.916	0.444	1.000	0.615
		NER + pattern-matching	0.907	0.925	0.916	0.444	1.000	0.615
		NER + second-pass	0.892	0.936	0.913	0.444	1.000	0.615
		NER + 2nd NER in parallel	0.897	0.932	0.914	0.444	1.000	0.615
		NER All	0.882	0.935	0.908	0.444	1.000	0.615
	LOCATION	NER (baseline)	0.692	0.874	0.773	0.769	0.222	0.345
		NER + pattern-matching	0.692	0.874	0.773	0.769	0.222	0.345
		NER + second-pass	0.669	0.850	0.749	0.667	0.222	0.333
		NER + 2nd NER in parallel	0.687	0.874	0.769	0.769	0.222	0.345
		NER All	0.664	0.850	0.746	0.667	0.222	0.333
	ORGANIZ.	NER (baseline)	0.588	0.669	0.625	0.442	0.662	0.530
		NER + pattern-matching	0.588	0.669	0.625	0.442	0.662	0.530
		NER + second-pass	0.586	0.669	0.625	0.557	0.665	0.607
		NER + 2nd NER in parallel	0.600	0.715	0.652	0.441	0.662	0.529
		NER All	0.597	0.715	0.650	0.557	0.665	0.606
ENTITY LEVEL	OVERALL	NER (baseline)	0.650	0.674	0.662	0.417	0.701	0.523
		NER + pattern-matching	0.650	0.674	0.662	0.417	0.701	0.523
		NER + second-pass	0.624	0.680	0.651	0.487	0.708	0.577
		NER + 2nd NER in parallel	0.657	0.705	0.680	0.417	0.701	0.523
		NER All	0.627	0.706	0.664	0.486	0.708	0.577
	PERSON	NER (baseline)	0.852	0.852	0.852	0.083	0.105	0.093
		NER + pattern-matching	0.852	0.852	0.852	0.083	0.105	0.093
		NER + second-pass	0.867	0.872	0.870	0.083	0.105	0.093
		NER + 2nd NER in parallel	0.852	0.858	0.855	0.083	0.105	0.093
		NER All	0.853	0.865	0.859	0.083	0.105	0.093
	LOCATION	NER (baseline)	0.667	0.641	0.653	0.385	0.156	0.222
		NER + pattern-matching	0.667	0.641	0.653	0.385	0.156	0.222
		NER + second-pass	0.659	0.648	0.654	0.333	0.156	0.213
		NER + 2nd NER in parallel	0.667	0.641	0.653	0.385	0.156	0.222
		NER All	0.659	0.648	0.654	0.333	0.156	0.213
	ORGANIZ.	NER (baseline)	0.527	0.581	0.552	0.430	0.778	0.554
		NER + pattern-matching	0.527	0.581	0.552	0.430	0.778	0.554
		NER + second-pass	0.481	0.576	0.524	0.508	0.784	0.616
		NER + 2nd NER in parallel	0.547	0.643	0.591	0.429	0.778	0.553
		NER All	0.500	0.638	0.561	0.507	0.784	0.616

Table C.4: NER performance in Spanish datasets.

C.6 Coreference Resolution

	English		German		Portuguese		Spanish	
Dataset	CoNLL	DCEP	CoNLL	DCEP	HAREM	DCEP	CoNLL	DCEP
B^3-precision	0.99	0.97	1.00	0.99	1.00	0.99	1.00	0.97
B^3-recall	0.96	0.94	0.96	0.80	0.93	0.93	0.95	0.88

Table C.5: Performance of the CRR in terms of macro-averaged B^3-scores of the entity chains.

	English		German		Portuguese		Spanish	
Dataset	CoNLL	DCEP	CoNLL	DCEP	HAREM	DCEP	CoNLL	DCEP
B^3-precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
B^3-recall	0.63	0.60	0.80	0.37	0.64	0.37	0.61	0.37

Table C.6: Performance of the CRR in terms of macro-averaged B^3-scores assuming singleton groups.

	English		German		Portuguese		Spanish	
Scores	CoNLL	DCEP	CoNLL	DCEP	HAREM	DCEP	CoNLL	DCEP
B^3-precision	0.69	0.88	0.87	1.00	0.64	0.83	0.88	0.94
B^3-recall	0.20	0.11	0.41	0.06	0.16	0.07	0.13	0.06

Table C.7: The performance of CRR carried out manually by humans in a subset of anonymized texts.

C.7 Inter-rater Agreement for Relevance

Table C.8 presents the inter-rater agreement for the rating of the relevance. This rating was made by two human readers, who assigned a binary score (Equation 3.9 on page 23) to each replacement entry. This score is based on the proper fit of the entry into the context. We used a subset of 5% of all replacements made by the anonymization methods based on substitution a common dataset between two raters, in order to assess the inter-rater agreement of the relevance. No inter-rater agreement was performed for the German datasets.

According to the table proposed by Landis & Koch [27], we could consider an almost perfect agreement between raters in Portuguese and Spanish datasets as Cohen's kappa and Fleiss kappa present sufficiently high values. The English DCEP could be considered in the lower limit of an almost perfect (or a high substantial) agreement. According to Landis & Koch [27], the English CoNLL dataset showed a moderate agreement between raters as the values of kappa are lower than 0.60.

	English		German		Portuguese		Spanish	
Dataset	CoNLL	DCEP	CoNLL	DCEP	HAREM	DCEP	CoNLL	DCEP
Average agreement[1]	78.6%	90.5%	*	*	87.5%	89.1%	89.3%	90.0%
Average Cohen-κ	0.569	0.809	*	*	0.747	0.769	0.772	0.713
Fleiss-κ	0.569	0.809	*	*	0.743	0.769	0.771	0.713
Observed agreement	0.786	0.905	*	*	0.875	0.891	0.893	0.9
Expected agreement	0.503	0.501	*	*	0.514	0.528	0.531	0.651

Table C.8: Agreement between annotators in each annotation task by language.
Data for the German datasets are not available.

[1] Average pairwise percent agreement

D Outputs

This Appendix presents the outputs of the anonymization system for some texts using the baseline NER module to detect sensitive information. These texts, one for each language, feature some of the issues that may occur on automated anonymization. The "*original*" output shows the text before the anonymization process. All other outputs are the results of each anonymization method over the original text. In the original text, the Named Entities from the golden-standard have been highlighted in bold typeface. These outputs from the anonymization system are presented *verbatim et literatim*.

It is possible to notice that all information related to NEs is lost due to the entity suppression. However, some entities still can be recovered from the context such as in the German ("*Hofheim*" can be recovered from the context) and Spanish texts (almost all the entities can be recovered from the context).

The tagging method replaces the entities by an artificial entry surrounded by square brackets. Although it does not seem natural, it gives enough information to the reader in order to follow the context of the document.

The random substitution replaces the entities by a random entity from a list with the same morphosyntactic features (MSF)s. It was achieved successfully in the replacement of person names, such as "*Keith O'Neill*" (masculine) by "*Jacob*", "*Gustavo Kuerten*" (masculine) by "*Álvaro*" and "*Vanessa Schultz*" (feminine) by "*Daniela*", and locations, such as "*Amazonas*" (masculine, singular) by "*Porto*" and "*Hessen*" (neuter, singular) by "*Leipzig*".

The generalization method resulted in a more natural anonymized text. Some replacements successful replacements for each language were: "*Liechtenstein*" by "*country*", "*Hofheim*" by "*Stadt*", "*Vale do Javari*" by "*região*" and "*Playboy*" by "*revista*". Some replacements were incorrect by linking with the wrong entry from the KB, such as "*Veja*" by "*género musical*". In the Portuguese text, there are two examples of unavailable translation of the entity such as "*en:diocese of the Catholic Church*".

D.1 English

This English text was extracted from the CoNLL 2003 dataset.

(a) Original

SOCCER - IRISH ERASE PAINFUL MEMORIES WITH 5-0 WIN . ESCHEN , Liechtenstein 1996-08-31 The Republic of Ireland 's new-look side dispelled painful memories of their last visit to Liechtenstein by beating the Alpine part-timers 5-0 in a World Cup qualifier on Saturday . The Irish , under new manager Mick McCarthy , took a 4-0 lead within 20 minutes through captain Andy Townsend , 20-year-old Norwich striker Keith O'Neill , Sunderland forward Niall Quinn and teenager Ian Harte . Quinn added his second and Ireland 's fifth just after the hour to complete the rout and give the Irish their biggest-ever away win . The result helped erase memories of Ireland 's visit to the Eschen stadium 14 months ago , when Jack Charlton 's side were held to a frustrating 0-0 draw which ultimately cost them a place in the European championship finals .

(b) Suppression

SOCCER - IRISH ERASE PAINFUL MEMORIES WITH 5-0 WIN . —— , —— 1996-08-31 The Republic of —— 's new-look side dispelled painful memories of their last visit to —— by beating the Alpine part-timers 5-0 in a World Cup qualifier on Saturday . The Irish , under new manager —— , took a 4-0 lead within 20 minutes through captain —— , 20-year-old —— striker —— , —— forward —— and teenager —— . —— added his second and —— 's fifth just after the hour to complete the rout and give the Irish their biggest-ever away win . The result helped erase memories of —— 's visit to the —— stadium 14 months ago , when Jack Charlton 's side were held to a frustrating 0-0 draw which ultimately cost them a place in the European championship finals .

(c) Tagging

SOCCER - IRISH ERASE PAINFUL MEMORIES WITH 5-0 WIN . [#**LOCATION1**#] , [#**LOCATION2**#] 1996-08-31 The Republic of [#**LOCATION3**#] 's new-look side dispelled painful memories of their last visit to [#**LOCATION2**#] by beating the Alpine part-timers 5-0 in a World Cup qualifier on Saturday . The Irish , under new manager [#**PERSON1**#] , took a 4-0 lead within 20 minutes through captain [#**PERSON2**#] , 20-year-old [#**ORGANIZATION1**#] striker [#**PERSON3**#] , [#**ORGANIZATION2**#] forward [#**PERSON4**#] and teenager [#**PERSON5**#] . [#**PERSON4**#] added his second and [#**LOCATION3**#] 's fifth just after the hour to complete the rout and give the Irish their biggest-ever away win . The result helped erase memories of [#**LOCATION3**#] 's visit to the [#**LOCATION1**#] stadium 14 months ago , when Jack Charlton 's side were held to a frustrating 0-0 draw which ultimately cost them a place in the European championship finals .

(d) Random Substitution

SOCCER - IRISH ERASE PAINFUL MEMORIES WITH 5-0 WIN . **Manchester** , **Leeds** 1996-08-31 The Republic of **Edinburgh** 's new-look side dispelled painful memories of their last visit to **Leeds** by beating the Alpine part-timers 5-0 in a World Cup qualifier on Saturday . The Irish , under new manager **Charlie** , took a 4-0 lead within 20 minutes through captain **William** , 20-year-old **Foundation** striker **Jacob** , **Company** forward **William1** and teenager **Charlie1** . **William1** added his second and **Edinburgh** 's fifth just after the hour to complete the rout and give the Irish their biggest-ever away win . The result helped erase memories of **Edinburgh** 's visit to the **Manchester** stadium 14 months ago , when Jack Charlton 's side were held to a frustrating 0-0 draw which ultimately cost them a place in the European championship finals .

(e) Generalization

SOCCER - IRISH ERASE PAINFUL MEMORIES WITH 5-0 WIN . **municipality of Liechtenstein** , **country** 1996-08-31 The Republic of **country 2** 's new-look side dispelled painful memories of their last visit to **country** by beating the Alpine part-timers 5-0 in a World Cup qualifier on Saturday . The Irish , under new manager **Oliver** , took a 4-0 lead within 20 minutes through captain **George** , 20-year-old **city** striker **Oscar** , **city 2** forward **Thomas** and teenager **Oliver 2** . **Thomas** added his second and **country 2** 's fifth just after the hour to complete the rout and give the Irish their biggest-ever away win . The result helped erase memories of **country 2** 's visit to the **municipality of Liechtenstein** stadium 14 months ago , when Jack Charlton 's side were held to a frustrating 0-0 draw which ultimately cost them a place in the European championship finals .

D.2 German

This German text was extracted from the CoNLL 2003 dataset.

(a) **Original**

SPD-Forum Hofheim zur Umweltpolitik HOFHEIM . Energie , Abfall und Abwasser sind die Hauptthemen beim SPD-Forum 2010 zu kommunaler Umweltpolitik am Donnerstag , 27. August . Die Podiumsdiskussion im Casino der Stadthalle beginnt um 20 Uhr . Haimo Brackmann und Thomas Norgall vom Bund für Umwelt und Naturschutz (BUND) Hessen , Dr. Arno Grau von der Deutschen Abwasser Reinigungsgesellschaft (DAR) und Dr. Thomas Rautenberg vom Umlandverband Frankfurt wollen erklären , wie der Naturschutz in Hofheim aktiv unterstützt und verbessert werden kann . Aber das Quartett auf dem Podium will sich auch mit den Zuhörinnen und Zuhörern unterhalten und Fragen beantworten .

(b) **Suppression**

SPD-Forum —— zur Umweltpolitik —— . Energie , Abfall und Abwasser sind die Hauptthemen beim SPD-Forum 2010 zu kommunaler Umweltpolitik am Donnerstag , 27. August . Die Podiumsdiskussion im Casino der Stadthalle beginnt um 20 Uhr . —— und —— vom —— (——) —— , Dr. —— von der Deutschen Abwasser Reinigungsgesellschaft (DAR) und Dr. —— vom —— wollen erklären , wie der Naturschutz in —— aktiv unterstützt und verbessert werden kann . Aber das Quartett auf dem Podium will sich auch mit den Zuhörinnen und Zuhörern unterhalten und Fragen beantworten .

(c) **Tagging**

SPD-Forum __[#**LAGE1**#]__ zur Umweltpolitik __[#**LAGE1**#]__ . Energie , Abfall und Abwasser sind die Hauptthemen beim SPD-Forum 2010 zu kommunaler Umweltpolitik am Donnerstag , 27. August . Die Podiumsdiskussion im Casino der Stadthalle beginnt um 20 Uhr . __[#**PERSON1**#]__ und __[#**PERSON2**#]__ vom __[#**ORGANISATION1**#]__ (__[#**ORGANISATION2**#]__) __[#**LAGE2**#]__ , Dr. __[#**PERSON3**#]__ von der Deutschen Abwasser Reinigungsgesellschaft (DAR) und Dr. __[#**PERSON4**#]__ vom __[#**ORGANISATION3**#]__ wollen erklären , wie der Naturschutz in __[#**LAGE1**#]__ aktiv unterstützt und verbessert werden kann . Aber das Quartett auf dem Podium will sich auch mit den Zuhörinnen und Zuhörern unterhalten und Fragen beantworten .

(d) **Random Substitution**

SPD-Forum **Dresden** zur Umweltpolitik **Dresden** . Energie , Abfall und Abwasser sind die Hauptthemen beim SPD-Forum 2010 zu kommunaler Umweltpolitik am Donnerstag , 27. August . Die Podiumsdiskussion im Casino der Stadthalle beginnt um 20 Uhr . **Elias** und **Luka** vom **Unternehmen** (**Institut**) **Leipzig** , Dr. **Paul** von der Deutschen Abwasser Reinigungsgesellschaft (DAR) und Dr. **Leon** vom **Unternehmen1** wollen erklären , wie der Naturschutz in **Dresden** aktiv unterstützt und verbessert werden kann . Aber das Quartett auf dem Podium will sich auch mit den Zuhörinnen und Zuhörern unterhalten und Fragen beantworten .

(e) **Generalization**

SPD-Forum **Stadt** zur Umweltpolitik **Stadt** . Energie , Abfall und Abwasser sind die Hauptthemen beim SPD-Forum 2010 zu kommunaler Umweltpolitik am Donnerstag , 27. August . Die Podiumsdiskussion im Casino der Stadthalle beginnt um 20 Uhr . **Elias** und **Louis** vom **eingetragener Verein** (**Verkehrsweg**) **Bundesland** , Dr. **Fynn** von der Deutschen AbwasserReinigungsgesellschaft (DAR) und Dr. **Jonas** vom **kreisfreie Stadt** wollen erklären , wie der Naturschutz in **Stadt** aktiv unterstützt und verbessert werden kann . Aber das Quartett auf dem Podium will sich auch mit den Zuhörinnen und Zuhörern unterhalten und Fragen beantworten .

D.3 Portuguese

This Portuguese text was extracted from the HAREM dataset.

(a) **Original**

Amazonas - Índios querem sobrenomes característicos da etnia HOME PAGE RONDONIA AO VIVO , 24.12.2007 Índios do Amazonas reivindicam o direito de registrar os filhos de forma organizada , respeitosa e com os sobrenomes que caracterizam suas etnias . Os pontos foram apontados pelos indígenas na reunião de encerramento da primeira fase do Projeto Registro Civil dos Povos Indígenas do Amazonas , que aconteceu na última semana em Manaus . De acordo com o projeto , iniciado em setembro de 2007 , 17 líderes indígenas estiveram em 44 comunidades no Alto Rio Negro , Alto e Médio Solimões , Vale do Javari , Purus , Juruá e Manaus para aplicar os questionários de avaliação sobre a atual situação dos registros civis da população e coletar dados para a elaboração de um relatório . O documento , será apresentado em março de 2008 e incluirá o resultado geral das avaliações e as sugestões dos povos indígenas . Ao todo , foram mais de 1,4 mil questionários aplicados em 325 comunidades , o equivalente a 43 etnias visitadas . Fonte : Clipping da 6ªCCR do MPF .

(b) **Suppression**

— - — querem sobrenomes característicos da etnia HOME PAGE RONDONIA AO VIVO , 24.12.2007 — do — reivindicam o direito de registrar os filhos de forma organizada , respeitosa e com os sobrenomes que caracterizam suas etnias . Os pontos foram apontados pelos indígenas na reunião de encerramento da primeira fase do Projeto Registro Civil dos Povos Indígenas do Amazonas , que aconteceu na última semana em — . De acordo com o projeto , iniciado em setembro de 2007 , 17 líderes indígenas estiveram em 44 comunidades no — , — , — , — , — e — para aplicar os questionários de avaliação sobre a atual situação dos registros civis da população e coletar dados para a elaboração de um relatório . O documento , será apresentado em março de 2008 e incluirá o resultado geral das avaliações e as sugestões dos povos indígenas . Ao todo , foram mais de 1,4 mil questionários aplicados em 325 comunidades , o equivalente a 43 etnias visitadas . Fonte : Clipping da — .

(c) **Tagging**

[#LOCAL1**#]** - **[#**PESSOA1**#]** querem sobrenomes característicos da etnia HOME PAGE RONDONIA AO VIVO , 24.12.2007 **[#**PESSOA1**#]** do **[#**LOCAL1**#]** reivindicam o direito de registrar os filhos de forma organizada , respeitosa e com os sobrenomes que caracterizam suas etnias . Os pontos foram apontados pelos indígenas na reunião de encerramento da primeira fase do Projeto Registro Civil dos Povos Indígenas do Amazonas , que aconteceu na última semana em **[#**LOCAL2**#]** . De acordo com o projeto , iniciado em setembro de 2007 , 17 líderes indígenas estiveram em 44 comunidades no **[#**LOCAL3**#]** , **[#**LOCAL4**#]** , **[#**LOCAL5**#]** , **[#**LOCAL6**#]** , **[#**LOCAL7**#]** e **[#**LOCAL2**#]** para aplicar os questionários de avaliação sobre a atual situação dos registros civis da população e coletar dados para a elaboração de um relatório . O documento , será apresentado em março de 2008 e incluirá o resultado geral das avaliações e as sugestões dos povos indígenas . Ao todo , foram mais de 1,4 mil questionários aplicados em 325 comunidades , o equivalente a 43 etnias visitadas . Fonte : Clipping da **[#**ORGANIZACAO1**#]** .

(d) **Random Substitution**

Porto - **Rodrigo** querem sobrenomes característicos da etnia HOME PAGE RONDONIA AO VIVO , 24.12.2007 **Rodrigo** do **Porto** reivindicam o direito de registrar os filhos de forma organizada , respeitosa e com os sobrenomes que caracterizam suas etnias . Os pontos foram apontados pelos indígenas na reunião de encerramento da primeira fase do Projeto Registro Civil dos Povos Indígenas do Amazonas , que aconteceu na última semana em **Faro** . De acordo com o projeto , iniciado em setembro de 2007 , 17 líderes indígenas estiveram em 44 comunidades no **Faro1** , **Vila Real** , **Viana do Castelo** , **Viana do Castelo1** , **Faro2** e **Faro** para aplicar os questionários de avaliação sobre a atual situação dos registros civis da população e coletar dados para a elaboração de um relatório . O documento , será apresentado em março de 2008 e incluirá o resultado geral das avaliações e as sugestões dos povos indígenas . Ao todo , foram mais de 1,4 mil questionários aplicados em 325 comunidades , o equivalente a 43 etnias visitadas . Fonte : Clipping da **Instituto** .

(e) **Generalization**

departamento da Colômbia - **Miguel** querem sobrenomes característicos da etnia HOME PAGE RONDONIA AO VIVO , 24.12.2007 **Miguel** do **departamento da Colômbia** reivindicam o direito de registrar os filhos de forma organizada , respeitosa e com os sobrenomes que caracterizam suas etnias . Os pontos foram apontados pelos indígenas na reunião de encerramento da primeira fase do Projeto Registro Civil dos Povos Indígenas do Amazonas , que aconteceu na última semana em **cidade** . De acordo com o projeto , iniciado em setembro de 2007 , 17 líderes indígenas estiveram em 44 comunidades no **en:diocese of the Catholic Church** , **en:town of the United States** , **região** , **rio** , **Local** e **cidade** para aplicar os questionários de avaliação sobre a atual situação dos registros civis da população e coletar dados para a elaboração de um relatório . O documento , será apresentado em março de 2008 e incluirá o resultado geral das avaliações e as sugestões dos povos indígenas . Ao todo , foram mais de 1,4 mil questionários aplicados em 325 comunidades , o equivalente a 43 etnias visitadas . Fonte : Clipping da **Organizacao** .

D.4 Spanish

This Spanish text was extracted from the CoNLL 2002 dataset.

(a) **Original**

El tenista brasileño Gustavo Kuerten , actual número dos del mundo , terminó su romance de varios meses con la modelo Vanessa Schultz porque ésta decidió posar desnuda para un ensayo de la edición nacional de la revista Playboy , informa el semanario Veja en su último número . Schultz , una escultural jovencita de ojos y cabellos castaños , será la portada de la edición de Playboy del mes de julio , lo que irritó al tenista , quien ya había manifestado varias veces su rechazo a otras ofertas hechas por la misma publicación a su novia para que mostrase sus atributos al público brasileño . " La revista ya me había invitado antes , pero él (Kuerten) nunca aceptó " , indicó Schultz , y agregó que tomó la decisión de salir desnuda en Playboy porque quiere " brillar por cuenta propia " y no vivir a la sombra de " Guga " , como se conoce popularmente en Brasil al tenista . La belleza de la modelo atrajo la atención de los medios de comunicación del mundo la mayoría de las veces que acompañó al tenista en los torneos internacionales que ha disputado . Sin entrar en detalles sobre los escenarios donde mostrará su belleza al natural , Schultz añadió que " después de ver las fotografías Guga nunca más va a hablar conmigo " . Santander , 23 may (EFE) .

(b) **Suppression**

El tenista brasileño —— , actual número dos del mundo , terminó su romance de varios meses con la modelo —— porque ésta decidió posar desnuda para un ensayo de la edición nacional de la revista —— , informa el semanario —— en su último número . —— , una escultural jovencita de ojos y cabellos castaños , será la portada de la edición de —— del mes de julio , lo que irritó al tenista , quien ya había manifestado varias veces su rechazo a otras ofertas hechas por la misma publicación a su novia para que mostrase sus atributos al público brasileño . " La revista ya me había invitado antes , pero él (——) nunca aceptó " , indicó —— , y agregó que tomó la decisión de salir desnuda en —— porque quiere " brillar por cuenta propia " y no vivir a la sombra de " —— " , como se conoce popularmente en —— al tenista . La belleza de la modelo atrajo la atención de los medios de comunicación del mundo la mayoría de las veces que acompañó al tenista en los torneos internacionales que ha disputado . Sin entrar en detalles sobre los escenarios donde mostrará su belleza al natural , —— añadió que " después de ver las fotografías —— nunca más va a hablar conmigo " . —— , 23 may (——) .

(c) **Tagging**

El tenista brasileño [#**PERSONA1**#] , actual número dos del mundo , terminó su romance de varios meses con la modelo [#**PERSONA2**#] porque ésta decidió posar desnuda para un ensayo de la edición nacional de la revista [#**ORGANIZACION1**#] , informa el semanario [#**ORGANIZACION2**#] en su último número . [#**PERSONA2**#] , una escultural jovencita de ojos y cabellos castaños , será la portada de la edición de [#**ORGANIZACION1**#] del mes de julio , lo que irritó al tenista , quien ya había manifestado varias veces su rechazo a otras ofertas hechas por la misma publicación a su novia para que mostrase sus atributos al público brasileño . " La revista ya me había invitado antes , pero él ([#**PERSONA1**#]) nunca aceptó " , indicó [#**PERSONA2**#] , y agregó que tomó la decisión de salir desnuda en [#**ORGANIZACION1**#] porque quiere " brillar por cuenta propia " y no vivir a la sombra de " [#**PERSONA3**#] " , como se conoce popularmente en [#**LOCALIZACION1**#] al tenista . La belleza de la modelo atrajo la atención de los medios de comunicación del mundo la mayoría de las veces que acompañó al tenista en los torneos internacionales que ha disputado . Sin entrar en detalles sobre los escenarios donde mostrará su belleza al natural , [#**PERSONA2**#] añadió que " después de ver las fotografías [#**PERSONA3**#] nunca más va a hablar conmigo " . [#**LOCALIZACION2**#] , 23 may ([#**ORGANIZACION3**#]) .

(d) **Random Substitution**

El tenista brasileño **Álvaro** , actual número dos del mundo , terminó su romance de varios meses con la modelo **Daniela** porque ésta decidió posar desnuda para un ensayo de la edición nacional de la revista **Instituto** , informa el semanario **Instituto1** en su último número . **Daniela** , una escultural jovencita de ojos y cabellos castaños , será la portada de la edición de **Instituto** del mes de julio , lo que irritó al tenista , quien ya había manifestado varias veces su rechazo a otras ofertas hechas por la misma publicación a su novia para que mostrase sus atributos al público brasileño . " La revista ya me había invitado antes , pero él (**Álvaro**) nunca aceptó " , indicó **Daniela** , y agregó que tomó la decisión de salir desnuda en **Instituto** porque quiere " brillar por cuenta propia " y no vivir a la sombra de " **Pablo** " , como se conoce popularmente en **Barcelona** al tenista . La belleza de la modelo atrajo la atención de los medios de comunicación del mundo la mayoría de las veces que acompañó al tenista en los torneos internacionales que ha disputado . Sin entrar en detalles sobre los escenarios donde mostrará su belleza al natural , **Daniela** añadió que " después de ver las fotografías **Pablo** nunca más va a hablar conmigo " . **Valencia** , 23 may (**Escuela**) .

(e) **Generalization**

El tenista brasileño **Javier** , actual número dos del mundo , terminó su romance de varios meses con la modelo **María** ésta decidió posar desnuda para un ensayo de la edición nacional de la revista **revista 1** , informa el semanario **género musical** su último número . **María** , una escultural jovencita de ojos y cabellos castaños , será la portada de la edición de **revista 1** mes de julio , lo que irritó al tenista , quien ya había manifestado varias veces su rechazo a otras ofertas hechas por la misma publicación a su novia para que mostrase sus atributos al público brasileño . " La revista ya me había invitado antes , pero él (**Javier**) nunca aceptó " , indicó **María** , y agregó que tomó la decisión de salir desnuda en **revista 1** quiere " brillar por cuenta propia " y no vivir a la sombra de " **Diego** " , como se conoce popularmente en **país** tenista . La belleza de la modelo atrajo la atención de los medios de comunicación del mundo la mayoría de las veces que acompañó al tenista en los torneos internacionales que ha disputado . Sin entrar en detalles sobre los escenarios donde mostrará su belleza al natural , **María** que " después de ver las fotografías **Diego** más va a hablar conmigo " . **provincia de España** , 23 may (**Organizacion**) .

E Annotation Guidelines

E.1 Annotation Guidelines of Coreferences in an Anonymized Text

Introduction

In this task, we consider *coreferences* as anaphoric relations between named entities in a text. In this annotation task, two named entities have a coreference relation if both refer the same extralinguistic object. In a text document, a human reader would identify such relation between two entities by their surface form.

Figure E.1 is a sample sentence[1] whose relations between entities are indicated with diferent colors:

➔ Mr. Trump made the comments in a town-hall-style forum with Chris Matthews of MSNBC, which was prerecorded for broadcast Wednesday night. Mr. Matthews pressed Mr. Trump on his support for a ban on abortion, asking him how he would enforce such a ban .

Figure E.1: Sample sentence to be anonymized.

However, when a text document is anonymized, the original surface forms of the entities are lost. In that case, a human reader can only identify the coreference relation between anonymized entities by the context of the mentions in the document.

The Figure E.2 presents the previous sample sentence after being anonymized

➔ [******] made the comments in a town-hall-style forum with [******] of [******], which was prerecorded for broadcast Wednesday night. [******] pressed [******] on his support for a ban on abortion, asking him how he would enforce such a ban.

Figure E.2: Sample sentence (E.1) after the anonymization process.

Scope

The objective of this task is to determine how much information a human reader can uncover from an anonymized text. In this task, we ask an annotator to discover coreference relations between entities whose form was hidden in a text, and mark their suggestions in our annotation tool.

[1] Sentence extracted from: http://www.nytimes.com/2016/03/31/us/politics/donald-trump-abortion.html

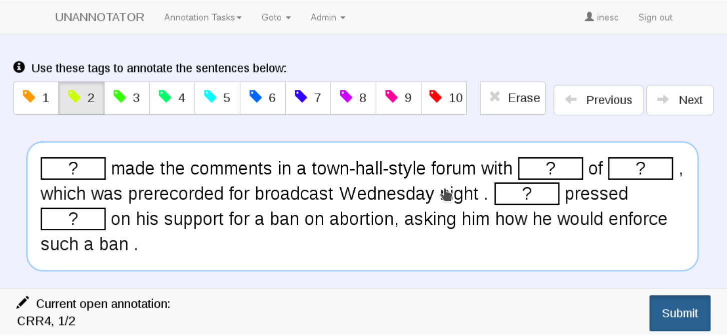
The objective of these guidelines is to maintain some annotation consistency when performed by different annotators. This guidelines clarify some questions that the annotator may encounter during the annotation task. We provide some examples and exclusions for each annotation markable. We use a schema of colors to characterize each markable, in a similar way it appears in the annotation tool.

Annotation Tool

For this annotation task we will use Unannotator[2] in the coreference mode. This annotation tool is accessed by the human annotator through an Internet browser. This tool also works in browsers from touch screen devices.

Markables

The markables in this task are entities that have been hidden from the document. These entities were hidden by a rectangle containing a interrogation point as shown in Figure E.3.



Use these tags to annotate the sentences below:

? made the comments in a town-hall-style forum with ? of ? , which was prerecorded for broadcast Wednesday night . ? pressed ? on his support for a ban on abortion, asking him how he would enforce such a ban .

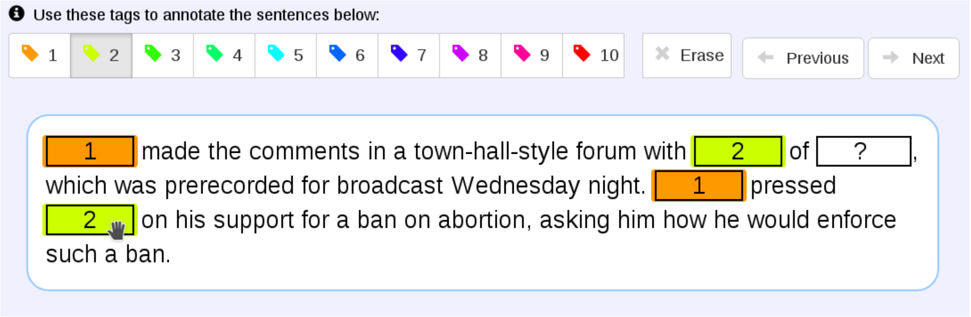
Current open annotation:
CRR4, 1/2

Figure E.3: Interface of the annotation tool showing the sample sentence.

To mark the coreference relation between two or more entities, the annotator must group all these entities into one slot. Slots are available as numbered buttons in the top left menu of Unannotator. In order to assign a slot to a hidden entity, the annotator must click over a numbered slot in the upper left menu of Unannotator and then click over the rectangles that hide the entities. The rectangles change in color and are labeled with the number of the slot. Clicking over an already assigned hidden entity will remove it from the slot, and therefore will remove the color and label from the rectangle. A sample is shown in Figure E.4.

After marking all sugestions, the annotator should click the blue "Submit" button in the lower right corner of the annotation tool window.

[2] URL: https://www.l2f.inesc-id.pt/~fdias/unannotator



1 made the comments in a town-hall-style forum with 2 of ?, which was prerecorded for broadcast Wednesday night. 1 pressed 2 on his support for a ban on abortion, asking him how he would enforce such a ban.

Figure E.4: Example of the annotation of two groups of hidden entities in the sample sentence.

Exclusions

- There is a maximum of 12 groups of entities in each text.
- There are entities that do not belong to any group (eg. "MSNBC" in the sample sentence). Do not mark a group with only one hidden entity. These entities should be kept unmarked, showing the interrogation point inside the rectangle.

E.2 Annotation Guidelines for Entity Relevance Task

Introduction

Relevance is the relation of something that is suitable with a given context. In this annotation task, we aim to determine of some expressions, herein after referred as entities, are relevant in a given context.

Figure E.5 is a sample sentence whose [3] entities are indicated inside colored boxes:

➔ Mr. Trump made the comments in a town-hall-style forum with Chris Matthews of MSNBC, which was prerecorded for broadcast Wednesday night. Mr. Matthews pressed Mr. Trump on his support for a ban on abortion, asking him how he would enforce such a ban.

Figure E.5: Sample sentence containing entities to be replaced.

However, when we replace the original entities from the text, the original meaning of the text may change. In that case, we say that the replacements are irrelevant for the original context. The Figure E.6 presents the previous sample sentence with replaced entities. Showing irrelevant substitution from Mr. Matthews and MSNBC.

➔ JOHN made the comments in a town-hall-style forum with PRESIDENT of UNIVERSITY, which was prerecorded for broadcast Wednesday night. PRESIDENT pressed JOHN on his support for a ban on abortion, asking him how he would enforce such a ban.

Figure E.6: Sample sentence (E.5) after replacement of the entities.

[3] Sentence extracted from: http://www.nytimes.com/2016/03/31/us/politics/donald-trump-abortion.html

Scope

The objective of this task is to determine the relevance of substitute entities when randomly replaced in a text. In this task, we ask an to rate these entities as 'relevant' or 'not relevant' given the context, using our annotation tool.

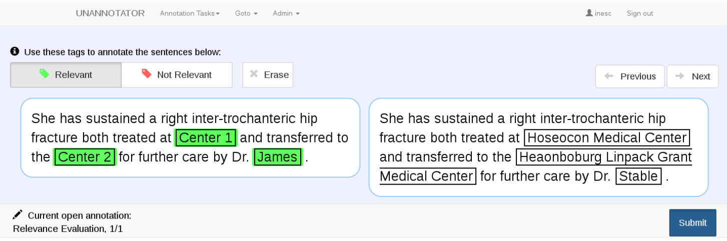
The objective of this guidelines is to maintain some annotation consistency when performed by different annotators. This guidelines clarify some questions that the annotator may encounter during the annotation task. We provide some examples and exclusions for each annotation markable. We use a schema of colors similar to the schema of the annotation tool to characterize each markable.

Annotation Tool

For this annotation task we will use Unannotator[4] in the evaluation mode. This annotation tool is accessed by the human annotator through an Internet browser. This tool also works in browsers from touch screen devices.

Markables

The markables in this task are entities that have been replaced in the document. These entities are emphatized by a rectangular border and bold typeface as shown in Figure E.7.



Use these tags to annotate the sentences below:

She has sustained a right inter-trochanteric hip fracture both treated at Center 1 and transferred to the Center 2 for further care by Dr. James .

She has sustained a right inter-trochanteric hip fracture both treated at Hoseocon Medical Center and transferred to the Heaonboburg Linpack Grant Medical Center for further care by Dr. Stable .

Current open annotation:
Relevance Evaluation, 1/1

Figure E.7: Interface of the annotation tool.

To evaluate the relevance of an entity, the annotator must click one of the buttons (relevante or not relevant) in the upper left menu of Unannotator and then click over the rectangles that contain the entities. The rectangles change in color. Clicking over an already marked entity will remove its, and therefore will remove the color from the rectangle. A sample is shown in Figure E.7.

After marking all the entities, the annotator should click blue "Submit" button in the lower right corner of the annotation tool.

Exclusions

- By default, we assume that all replacements of person names are always relevant.

[4] URL: `https://www.l2f.inesc-id.pt/~fdias/unannotator`