Automatic Annotation of Unstructured Fields in Medical Databases

Margarida Andreia Rosa Correia

Thesis to obtain the Master of Science Degree in
Information Systems and Computer Engineering

Supervisors: Prof. Arlindo Manuel Limede de Oliveira
Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur

Examination Committee

Chairperson: Prof. José Carlos Alves Pereira Monteiro
Supervisor: Prof. Arlindo Manuel Limede de Oliveira
Members of the Committee: Prof. Bruno Emanuel Da Graça Martins

October 2018
Acknowledgments

In first place, I would like to thank my thesis advisors Professor Maria Luisa Coheur and Professor Arlindo Oliveira for all the meetings and emails exchanged and for always correcting me without ever discouraging me. Also, I would like to thank Professor Nuno Mamede for his active participation in this thesis, helping me with all the questions with the STRING chain.

I also express my gratitude to the Eng. Fernando Martins and the Dr. Maria José Santos for making available the data present in the Reuma.pt database, in partnership with Sociedade Portuguesa de Reumatologia. Without the data, all this work would be impossible.

A thesis is a long journey, which includes a trajectory composed of innumerable challenges, sorrows, uncertainties, joys and many mishaps along the way.

A journey may be made by plane and every plane needs two wings to fly and to support it and I want to thank to mine. To my mother and father for always being present during this long walk and for always being my support throughout my years of study and through the process of researching and writing this thesis.

There are more crucial pieces on a plane, and I’m grateful to them for letting me fly too. To my maternal and paternal grandparents, my uncles and my cousins, my dearest thanks.

Every plane also needs a pilot and I would also like to thank my boyfriend for taking me to the destination, without ever letting me have an accident. Thank you for your patience and dedication, I’m a lucky girl.

Finally, every plane needs passengers and I want to thank my friends for being present on this trip and for giving me the privilege of taking them there. This thesis could not come to a safe port without the precious support of several people. To all of you, thank you.
Abstract

The increased use of systems based on Electronic Health Records caused an enormous increment of information available electronically, which can be processed by Data Mining techniques, leading to relevant findings. The expected result was that this information becomes easy to access, analyze and share. However, the text present in the clinical notes is written in natural language, and is, thus, unstructured, and difficult to automatically process. These clinical notes might contain pertinent data for the health of the patient. In this thesis, with the help of Natural Language Processing and Information Extraction techniques, we present a system that, given a clinical note, extracts relevant named entities from it, such as names of diseases, symptoms, treatments, diagnosis and drugs, generating structured information from unstructured free text. In addition, in order to avoid privacy issues and considering that these clinical notes might contain references to names of patients, doctors or another health professionals, we also present an anonymization step. Finally, we add a module that automatically corrects typos from these medical notes. Final results show that the system, in general, is able to recognize and interpret medical entities.

Keywords

Electronic Health Record, Information Extraction, Natural Language Processing, unstructured data, structured data.
Resumo

O aumento do uso de sistemas baseados em Registos de Saúde Eletrónicos causou um enorme crescimento da informação disponível eletronicamente, que pode ser processada por técnicas de Data Mining, levando a resultados relevantes. O resultado esperado era que essa informação se tornasse fácil de aceder, analisar e partilhar. No entanto, o texto presente nas notas clínicas está escrito em língua natural e, portanto, não estruturado e difícil de processar automaticamente. Essas notas clínicas podem conter dados pertinentes para a saúde do paciente. Nesta tese, com a ajuda de técnicas de Processamento de Linguagem Natural e de Extração de Informação, apresentamos um sistema que, dada uma nota clínica, extrai entidades relevantes desta, tais como nomes de doenças, sintomas, tratamentos, diagnósticos e fármacos, gerando informação estruturada do texto não estruturado. Para além disso, de maneira a evitar problemas de privacidade e considerando que estes campos também podem conter referências a nomes de pacientes, médicos ou outros profissionais de saúde, também apresentamos um passo de anonimização. Por fim, adicionamos um módulo que automaticamente corrige erros tipográficos destas notas clínicas. Os resultados finais mostram que o sistema, em geral, está apto para reconhecer e interpretar entidades médicas.

Palavras Chave

Registos de Saúde Eletrónicos, Extração de Informação, Processamento de Linguagem Natural, dados não estruturados, dados estruturados.
Contents

1 Introduction 1
  1.1 Thesis Proposal and Contributions 3
  1.2 Outline for the Dissertation 4

2 Related Work 5
  2.1 Anonymization 6
    2.1.1 Anonymization with STRING Chain 6
    2.1.2 ARX Tool 7
    2.1.3 Best-of-Breed System 8
  2.2 Pre-Processing 9
    2.2.1 Spell Checking 9
    2.2.2 Word Sense Disambiguation 10
  2.3 Information Extraction 11
    2.3.1 AMBIT Framework 12
    2.3.2 Natural Language Text Processor for Clinical Radiology 13
    2.3.3 cTAKES system 14
  2.4 Overview 15

3 Data Analysis and Solution Proposal 17
  3.1 The Medical Records 18
    3.1.1 The Database 21
    3.1.2 Descriptive Analysis of the Field of Clinical Notes 22
    3.1.3 Description of Clinical Notes Topics 23
    3.1.4 Medical Entities to Recognize 24
  3.2 Methodology and Proposed Solution 25
    3.2.1 The STRING Chain 26
    3.2.2 System Overview 27
List of Figures

1.1 Frame that we want to obtain .................................................. 2
1.1 Figure taken from [1] that shows how the rheumatologic diseases affect the quality of life of the patients .................................................. 3

2.1 Figure taken from [2] which shows the architecture of the anonymizer from the STRING chain .................................................. 6
2.2 Figure taken from [2] which shows the architecture of the BoB system .............. 8
2.3 Figure taken from [3] which shows the architecture of the spell checker ............. 10
2.4 Architecture of the AMBIT Framework ........................................ 12
2.5 Figure taken from [4] which shows the architecture of the NLP for clinical radiology .... 13
2.1 Example taken from [5] that shows all the modules of the cTAKES system acting individually 14

3.1 Circular chart that shows the percentage of each rheumatologic disease on our data .... 18
3.2 Chart that shows the percentage of rheumatological diseases on male and female .... 19
3.3 Chart that shows the ages of the patients with rheumatological diseases ............. 20
3.4 Chart that shows the percentage of biologicals taken by rheumatological disease .... 20
3.5 The database tables organization ............................................ 22
3.1 Portion of clinical text that refers to a name of a doctor ................................ 22
3.2 Portion of clinical text with lack of accents and punctuation and typos .............. 23
3.3 Portion of clinical text with acronyms and abbreviations .......................... 23
3.2 STRING chain architecture .................................................. 26
3.3 System Overview and Solutions ............................................ 27

4.1 Clinical note already anonymized ............................................ 31
4.2 Syntax of a rule on the STRING chain ....................................... 36
4.3 Rule that chooses only one label to a word .................................... 36
4.4 Rule about the human body .................................................. 37
4.5 Rule about a situation of diagnosing ........................................ 37
List of Tables

4.1 Table to help in anonymization evaluation ................................. 32
4.2 Precision results of the initial evaluation of lists .......................... 36
4.3 Number of best suggestions to the word change .......................... 39
4.4 Number of entities on the preliminary evaluation ....................... 47
4.5 Precision and Recall metrics applied to the structure on the preliminary evaluation ...................................................... 47
4.6 Auxiliary table to help in Precision and Recall results .................. 48

5.1 Number of entities on the evaluation corpora ............................. 55
5.2 Precision and Recall metrics applied to the structure ................. 59
5.3 Auxiliary table to help in Precision and Recall results ............... 59

6.1 Number of entities that we add to the STRING .......................... 62
<table>
<thead>
<tr>
<th>Acronyms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEF</td>
<td>Clinical e-Science Framework</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional Random Field</td>
</tr>
<tr>
<td>CRR</td>
<td>Coreference Resolution</td>
</tr>
<tr>
<td>CWL</td>
<td>Corpus Words List</td>
</tr>
<tr>
<td>EHRs</td>
<td>Electronic Health Records</td>
</tr>
<tr>
<td>f-CWL</td>
<td>filtered Corpus Words List</td>
</tr>
<tr>
<td>HFL</td>
<td>High Frequency List</td>
</tr>
<tr>
<td>IE</td>
<td>Information Extraction</td>
</tr>
<tr>
<td>INESC-ID</td>
<td>Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento</td>
</tr>
<tr>
<td>KWL</td>
<td>Known Words List</td>
</tr>
<tr>
<td>LFL</td>
<td>Low Frequency List</td>
</tr>
<tr>
<td>MED</td>
<td>Medical Entities Dictionary</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>NE</td>
<td>Named Entity</td>
</tr>
<tr>
<td>NER</td>
<td>Named Entity Recognition</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>SPR</td>
<td>Sociedade Portuguesa de Reumatologia</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>UMLS</td>
<td>Unified Medical Language System</td>
</tr>
<tr>
<td>WSD</td>
<td>Word Sense Disambiguation</td>
</tr>
</tbody>
</table>
1 Introduction

Contents

1.1 Thesis Proposal and Contributions ........................................... 3
1.2 Outline for the Dissertation ...................................................... 4
On a daily basis, in hospitals and another health units, a large number of medical reports is written, and much of the information they contain is in a textual format [6, 7]. Electronic Health Records (EHRs) combine information from diverse sources like notes taken in consultations between doctors and patients, radiologic or laboratory results, data obtained directly from patients, among others. The record, thus, becomes an electronic merged collection of such data, organized in chronological order.

The purpose of creating EHRs is to gather essential information for each citizen to “improve the accuracy, efficiency, quality of health care and data recorded in a health record” [8]. As has already been mentioned, this is done by an electronic collection, for each citizen, of the data produced by healthcare providers, such as doctors, nurses, pediatricians and psychologists. To achieve this goal is necessary to register the patients in a system, collect the data and share this information among the health entities, ensuring the confidentiality of the people involved.

However, there are still fields that are not structured (free text clinical narratives), which are difficult to analyze and most often contain valuable information for the patient’s health status. These fields are shared rarely because they also include confidential patient information, such as name, address and citizen’s card number. This is commonly the case with clinical notes because of remains easier to the health professionals to express there the symptoms and patients complains, as well as, to document clinical events [9] in a field dedicated to unstructured free text.

With the help of Anonymization, Natural Language Processing (NLP) (which “is a subfield of computer science concerned with intelligent processing of human language” [10]) and Information Extraction (IE) (“which refers to the automatic extraction of concepts, entities, and events, as well as their relations and associated attributes from free text” [9]) areas, it is possible to extract data from unstructured free text, anonymize them (ensuring that there is no loss of information, only omission) and structure them into their relevant entities, as a way to acquire structured knowledge from the unstructured clinical text. We assume that structuring a text consists of creating a frame, that is a list of attribute-value pairs, regarding relevant information extracted from text. The attributes of the frame will be the entities and the value will be the value of these entities. As entities we can have, for instance: symptoms, drugs, treatments and diseases names. In Example 1.1 we show what we desire to obtain as frame.

**Example 1.1:** Frame that we want to obtain

Data: “O paciente tem febre e dor de cabeça quando acorda. Está a tomar Brufen.”
(The patient has flu and headache when waking up. Is taking Brufen.)

Extraction:

- **SINTOMA:** febre
- **SINTOMA:** dor de cabeça
- **FÁRMACO:** Brufen
Having the text structured, besides making it easier to analyze, summarize and share, makes it easier to extract knowledge of all this data, relating them to each other using inference techniques.

1.1 Thesis Proposal and Contributions

For this work, we proposed to analyze EHRs about rheumatologic patients. We obtained the information from “Sociedade Portuguesa de Reumatologia (SPR)”, which gave us data present in a database named “Registo Nacional de Doentes Reumáticos” (Reuma.pt).

Rheumatism is an age-related degenerative disease, so its frequency has been increasing with the increase of the average life expectancy of populations. This disease affects the quality of life of the society because it involves bones, muscles and joints. On Figure 1.1, taken from Pereira et al. (2016) article, the authors show that, currently, rheumatologic diseases, in comparison with other pathologies as allergies, gastrointestinal, mental, pulmonary and cardiologic diseases, are the most frequent (red circle on figure that shows the bigger diameter) and the ones which give the worse quality of life for the patients (the further down the curve, the worse the quality of life).

![Figure 1.1: Figure taken from [1] that shows how the rheumatologic diseases affect the quality of life of the patients](http://reuma.pt/)

The challenge of this work is to develop automatic annotation techniques in clinical notes using the methods from the NLP and IE areas. These clinical texts contain relevant information about the health of the patient and might help in clinical decision support. There is a bunch of work into applying Data Mining techniques on texts relatively to health. However, it is easier to infer knowledge with structured text. The focus of our work is to struct those texts and, with our results, will be possible to use them as data to future work on Data Mining.
To obtain our final results, we need to reach some goals:

- Pre-Process the texts due to the several uses of acronyms and spelling errors;
- Anonymize the texts to did not compromise the privacy of the people involved;
- Structure the texts through the relevant entities that constitute them. It will be reached using NLP, whose aim is to analyze unstructured free text applying Computational Techniques [11], and, more precisely, using IE to extract relevant information of data.

### 1.2 Outline for the Dissertation

The organization of the contents of this dissertation is as follows: Chapter 2 introduces related work of three important tasks of this work: Anonymization, Pre-Processing tasks and IE. We present systems that solve some of the problems that we found in our data, as well as medical datasets and competitions related to this work goals. Chapter 3 presents an analysis of our medical records and specifies our methodology and solution, including a description of the tool that we will use. In Chapter 4, we describe in detail the five significant steps until we reach our goal: Dealing with Acronyms, Anonymization, Named Entity Recognition (NER), Spell Checking and Structuring the Clinical Notes. We explain the challenges, our approach and we also present a preliminary evaluation. In Chapter 5, we exhibit our evaluation corpora and experimental evaluation, applying the evaluation metrics to the corpora chosen and comparing the results with the preliminary evaluation made before. Finally, Chapter 6 summarizes the main contributions of this thesis and presents ideas for future work, as well as work limitations.
2 Related Work

Contents

2.1 Anonymization .......................................................... 6
2.2 Pre-Processing .......................................................... 9
2.3 Information Extraction .................................................. 11
2.4 Overview ............................................................... 15
The general architecture of an IE process usually has two modules: Pre-Processing and the IE task, following the Bird et al. (2009) book. In this work, in addition, to follow this architecture, we still need to implement, first of pre-processing, the anonymization task because we have in hands confidential information that we can not compromise. In this chapter, we present systems, competitions, datasets and papers for these three principal tasks. In the final section, we show how they influenced our solution.

2.1 Anonymization

Medical records are delicate and, before undergoing transformations or extractions, must be anonymized in a way that does not compromise patient’s privacy and data’s integrity – this has made it arduous to obtain a lot of data and to share that between entities [6]. When we consider NER tools, data anonymization consists of finding named entities that we identify as pertinent to our domain and applying some techniques to eliminate/change this information in another expression. These anonymization systems based on NER can be either rule-based (dictionaries or pattern-matching), model-based (Machine Learning (ML) models) or hybrid (a combination of these two techniques) [13].

2.1.1 Anonymization with STRING Chain

Most of the systems worked with records written in English. For Portuguese texts, there was no specific system until Mamede et al. (2016) created one. This hybrid system uses the STRING chain [14] (later explained in this document) and is divided into four different steps: Pre-Processing, NER, Coreference Resolution (CRR) and Anonymization, as we can see in Figure 2.1.

Figure 2.1: Figure taken from [2] which shows the architecture of the anonymizer from the STRING chain

In the Pre-Processing step, is used the STRING to normalize texts, separating it into sentences and tokens. In the NER’s module, the STRING is also used to obtain the list of named entities. This system is able to anonymize names, localizations and organizations. Next, CRR is used to verify if two different named entities refer to the same object in a way to not be replaced by different terms. Considering titles and abbreviations as “Maria Silva” and “Sra. Silva”. We know that these two different writing ways, in the same context, refer to the same person, and the system knows that too.

In the last module (Anonymization), we can choose between suppression, tagging, random substitution and generalization methods, depending on the purpose of anonymization. Suppression as the name implies, consists in the omission of the Named Entity (NE), using a symbol or character that
replaces the original text, for instance, we could change "Maria" by "XXXXX". According to Mamede et al. (2016), this is the most common method used by anonymization systems. **Tagging** is similar to Suppression, but instead of changing the original text by a symbol or character, it substitutes it for a label that indicates its class. Continuing with the Maria example, the anonymization returning from tagging might result in "[*"PersonName01"]". In the **Random Substitution** method, the system replaces a NE with another NE, respecting the same genre. The last method, **Generalization**, by default, can not be applied to entities that represent person names, but consists of “replacing an entity by another that mentions an item of the same type but more general. e.g. University of Lisbon becomes University” [2].

### 2.1.2 ARX Tool

This tool is specific for English texts about biomedical data (genetic) and is applied only to structured data. It implements a wide variety of privacy methods as k-Anonymity, l-Diversity, t-Closeness and δ-Presence, is well documented, contains a great programming interface, among other characteristics [15].

In the **k-anonymity** method, the probability of re-identifying data is 1/k so, the higher the k, the lower is the risk of re-identification. “Each record is similar to at least another k-1 other records on the potentially identifying variables” [16]. This method is applied using techniques as Suppression (replacing attributes using a specific character) or Generalization (replacing some values by an interval in which they are included) [16]. The **l-Diversity** method is an expansion of the k-anonymity as a way to combat limitations of that, ensuring “data privacy even without identifying the enemy’s background knowledge to avoid attribute disclosure” [17]. The **t-Closeness** method is an improvement of the l-Diversity method, which, in its turn, is yet another improvement to the k-anonymity. This method decreases the correlation between the quasi-identifiers (identifiers that do not directly identify the person but give an idea of who he/she is, or its characteristics) and the identifiers, using a distance named **Earth Mover’s Distance (EMD)** [18]. The last way, δ-Presence, is used when the others referred methods manifest themselves, unsatisfactory metrics. This method "clearly links the quality of anonymization to the risk posed by inadequate anonymization", because it needs a human-understandable policy, asking “knowledge of all entities, not just those in the dataset to be anonymized" [19].

In the Prasser et al. (2014) article, they frequently refer that is very important that the anonymization tools remain balanced between maintaining a good quality of the data and ensuring the protection and privacy of them. On the background, the way that the tool acts is the same, independently of the privacy methods. Firstly, it removes the identifiers that do not have any relevance in the analysis of the data, that is, directly identifiers as names or number of citizen cards. Secondly, it generalizes or suppresses the quasi-identifiers. For instance, if the person has 17 years, the tool changes her/his age to ≤ 19 (an interval in which the value is involved), or if in the data, the sex of the person is explicit, it suppresses
this information with a specific character.

This tool also allows importing data from diverse sources (but it is impossible to change them when they are in this tool) and analysis the transformed data (already anonymized), comparing it with the original input.

### 2.1.3 Best-of-Breed System

It is a de-identification system for clinical documents written in the English language. De-identification differs from anonymization because anonymized data makes it virtually impossible to recover the person’s identification, while with de-identified data, although difficult, it is possible to retrieve the data intervenients.

Nowadays, most of the anonymization systems exhibit a hybrid approach, like the Best-of-Breed (BoB) system developed by Ferrández et al. (2013), which “uses rules and dictionaries to score a higher recall, and it also uses model-based classifiers in order to score a higher precision” [2].

In this system, there are two focus components. First, the authors focus on prioritizing patient confidentiality, so the recall measure is the one that stands out (because it is essential that the system not miss a name), rather than the precision measure (because even if it anonymizes unnecessary data, that is not so serious). To this end, they implemented a “high-sensitivity extraction component” that allows detecting all the candidates that can put at risk the privacy of the patient, using rules, dictionaries (it “also performs dictionary lookups using Lucene\(^1\) to detect proper names and clinical terms” [2]) and the Conditional Random Field (CRF) classifier from the Stanford NLP\(^2\). The other focus component is the “false positive filtering component” to reduce the false positives that come from the previous focus, to try to increase the performance of the system. In this case, is used a ML classifier - Support Vector Machine (SVM) - to know how to differentiate whether the data compromise the privacy of the patients or not, separating them. In Figure 2.2, we can see the architecture of the BoB system.

Before applying these two components, the system processes the clinical texts through a NLP, using OpenNLP tools\(^3\), resorting to Chunking, Tokenization, POS tagging, among other techniques.

![Figure 2.2](https://example.com/figure2.2.png)

Figure 2.2: Figure taken from [2] which shows the architecture of the BoB system

---

2. [https://nlp.stanford.edu/](https://nlp.stanford.edu/)
3. [https://opennlp.apache.org/](https://opennlp.apache.org/)
2.2 Pre-Processing

The tasks of pre-processing must be done carefully to avoid losing important information and, if successfully done, the IE task will become easier because it significantly reduces the search space and time, facilitating the data analysis in this next step. In this section, we present two Pre-Processing methods helpful for our work.

2.2.1 Spell Checking

The health professionals write the reports with misspelled words, requiring some type of correction either manual (which will be time-consuming), automatic or semi-automatic.

This step is not common in a general Pre-Processing task, but for medical records it is essential because “the incidence of misspellings in medical records is around 10%, which is significantly higher than the misspelling incidence for other types of texts” [21]. Normally, the existence of misspellings occurs due to typographical errors that are “a mistake made in the typing process. Most typographical errors consist of substitution, transposition, duplication or omission of a small number of characters” [22]. Today all technologies, such as smartphones or laptops, when processing, automatically correct our typos but, for less common terms (medical terms), there are corrections that are not appropriated.

Any spell-checking tool will depend on a good similarity function for words to find that most probable word to correct the misspelled one. There are approaches such the edit distance [23] or the length of the longest common subsequence of strings [22] that can be applied in this case.

A tool developed by Carvalho and Curto (2014) adopts a semi-automatic detection and correction of misspelled words. The tool has a corpus of common words in the domain – Known Words List (KWL) – and a corpus of words that appear in the report ordered alphabetically – Corpus Words List (CWL). The tool finds similarity between the words of KWL and CWL, using the Jaro distance and ordering these words by similarity in a filtered Corpus Words List (f-CWL) to filter small typos. Next, the authors need to separate f-CWL into High Frequency List (HFL) and Low Frequency List (LFL) based on thresholds. In HFL they assume that the words that occur many times do not contain many errors because they filter these words in the previous step. They will use this list as the KWL in the correction final step. In the LFL they assume that the words that occur a small number of times have errors or are abbreviations or technical terms. They will correct this list with HFL. The unique disadvantage of this tool is the work to create the KWL and the CWL.

There is another tool that uses semi-automatic detection, developed by Ratnasari et al. (2017). This tool is composed of four components: Pre-Processing, Error Detection, Error Correction and User Feedback, as we can see in Figure 2.3.

---

[21] Normal citation

[22] Normal citation

[23] Normal citation

In the Pre-Processing phase, the text is prepared to the next component, using some pre-processing techniques as conversion to lower case and tokenization. Next, in the Error Detection phase, the process is similar to the previously presented tool. The only difference in this tool is that it uses a corpus to comparison retrieved in texts written by doctors. They compare this corpus with every token of the original text. If they are not equal, it means that the word of our original text is misspelled. On the Error Correction phase, the tool uses the Levenshtein distance \[23\] between the misspelled and all the other words of the retrieved medical corpus. The lower the distance between the words, the lower the cost of the transformation of the misspelled word into the right word, since that means that they are the most similar. The last phase is the User Feedback phase that consists, in the end, in the user providing feedback, indicating if the transformed word suffered a good or a bad transformation.

### 2.2.2 Word Sense Disambiguation

When some words after tokenization appear alone they become ambiguous. It turns the recognition harder because they can have the same orthographic form but be very distinct in meaning. Word Sense Disambiguation (WSD) is the task of choosing the true meaning of a word \[24\].

There are four typical approaches to apply WSD. The simplest way to disambiguate words that
emerge alone is using **Hand-crafted Rules**, in which we construct, basing on the context that the words appear, a set of rules to help us to disambiguate the words and to find the true meaning of them [7]. The principal disadvantage of this method is the time that it takes to make all these disambiguation rules. The other three approaches are basing on ML **techniques**. The first one is the **Supervised Machine Learning** in which we need to acquire disambiguation of corpora based on tagged words, using classifiers [7]. The disadvantage of this method is that manual annotation is an expensive task. The next technique is the **Semi-Supervised Learning** and, according to Pakhomov et al. (2005) it is composed of four steps: Sense Inventory (“we used the empirical sense inventory derived from the manually annotated samples” – this step must be done by the domain experts in a way to attribute the true meaning to acronyms), Data Collection (for each sense of an acronym we try to find the circumstances in which it appears), Data Merging (if there are acronyms that do not have contexts, we create additional sets in which we merge them) and Context Vectors Generation (representation of training and test samples). The last technique is the use of **Unsupervised Learning** which use the entire data unlabeled with the guess that identical purposes occur in similar contexts and if we do a cluster of words per meaning when we have a new occurrence of one word we can classify it into the closest induced cluster [26].

In the medical domain it is frequent the use of acronyms and abbreviations and the disambiguation of these terms is a sub-task of **WSD** [25, 27]. For instance, **MCF** can refer to “**metacarpofalangeana**” or to “**Moto Club Faro**” (as we are in a medical domain, the most probable meaning of this acronym will be the first one). This example situation is easy to solve due to the context in which we are and because we only sugest two options of extension of the acronym. But it is very common in the medical context that an acronym can come to have several meanings, all of them within the context of medicine, according to Liu et al. (2002) which demonstrated that 81% of acronyms found in MEDLINE⁵ abstracts are ambiguous and have on average 16 senses, which makes difficult the work of disambiguation. A solution to that is considering a “**global context**” [25], that is, within the medical context, see in what context the acronym arises. For example, if we are reading a medical report that is talking about a heart problem, it is more normal for “**RA**” to mean “**right atrial**” than “**rheumatoid arthritis**”. It requires a rule-based system to choose the correct acronym’s extension and a database of acronyms with all their true extensions.

### 2.3 Information Extraction

After anonymizing and pre-processing the text records, we are ready to apply **IE** to our data. **IE** consists of performing some “**techniques for extracting limited kinds of semantic content from text**” [28]. It turns the unstructured and/or semi-structured text into machine-readable documents [11]. Next, we present three systems to do the **IE** task and we also show the datasets used and competitions in this area.

---

2.3.1 AMBIT Framework

They develop the framework for Acquiring Medical and Biological Information from Text (AMBIT) in the context of two competitions: the Clinical e-Science Framework (CLEF) and the myGrid.

The first competition aims to provide a structured and well-organized repository of clinical information that can then be queried and summarized. The data of this competition are cancer data from various sources such as “case notes, large reports, discharge summaries, etc” [29]. It uses as entities to identify and extract information about the patients the following: drugs, symptoms, diseases, investigations and interventions.

The goal of the second competition is to present a single workbench to be applied in bioinformatics services and also on text mining techniques.

The AMBIT framework is composed of three stages, according to Harkema et al. (2005) article: lexical and terminological processing, syntactic and semantic processing and discourse processing, as we can see in Figure 2.4.

![Figure 2.4: Architecture of the AMBIT Framework](image)

The first step is to find and classify the relevant entities of the text, using a finite state term recognizer (to recognize the entities as drugs and diseases) and a term parser (to transform shorter terms, identified by the term recognizer, into longer terms). In the second step, the framework performs a partial syntactic and semantic analysis to sentence in the text. Finally, the third step integrates the analysis of the previous stage into a discourse model that represents the semantic content of the text. In the end, the information is read and placed in templates that can be imported into the CLEF repository.

This framework is also composed of a terminology engine and a query engine. The terminology engine used is the Termino [30] and has a database formed of terms collected in the Unified Medical Language System (UMLS)⁶, HUGO gene Nomenclature Database⁷ and Gene Ontology (GO)⁸. It helps to recognize and identify multiple words, bypassing the problem of not having a single nomenclature, that is, when there are several terms that refer to the same situation, as far as the medicinal context is concerned. For instance, in the Portuguese language, “dor de cabeça” (headache) and “cefaleia” (headache) both refer to the same problem. The query engine allows performing a search to access the information.

---

⁶https://www.nlm.nih.gov/research/umls/
⁷https://www.genenames.org/
⁸http://www.geneontology.org/
2.3.2 Natural Language Text Processor for Clinical Radiology

This tool of NLP was made with the goal of identifying clinical information in radiologic reports to, after that, structuring this information. It is composed, according to Friedman et al. (1994), for three processing phases: **Parser, Phrase Regularization** and **Encoder**, as we can see in Figure 2.5.

![Diagram of NLP architecture](http://med.dmi.columbia.edu/)

**Figure 2.5:** Figure taken from [4] which shows the architecture of the NLP for clinical radiology

The **Parser** is used to determine the structure of the text, following the rules of a semantic grammar. In order to determine the structure, it generates a structured preliminary output. This phase is complicated due to a problem already mentioned in the previous framework and that is quite common in natural language texts: there is no "single nomenclature". In this process, all variations of the same term have to give rise to the same results, but "not all variations are reduced to one form by this stage of processing, and the structured forms do not yet correspond to unique controlled vocabulary concepts" [4] thus, in the **Phrase Regularization** phase a Mapping composed of multi-word phrases is used in order to reduce these linguistic variations of the natural language to a default form defined for the phrases.

In the last phase, **Encoding**, a list of synonyms, whose terms were taken from the Medical Entities Dictionary (MED)\(^9\), was used in order to map all possible variations of a concept to a single general concept. Turning to the example of the previous framework, all variations used in the Portuguese language for a headache, such as "cafaleia" and "cefalgia" are reduced to "dor de cabeça", thus eliminating the problem of variations of the same term.

\(^9\)http://med.dmi.columbia.edu/
2.3.3 cTAKES system

The Clinical Text Analysis and Knowledge Extraction System (cTAKES) is an NLP system, developed on 2013 by Mayo Clinic, to extract events and clinical concepts from the text. This open source system (at http://ctakes.apache.org/index.html) is based on the Unstructured Information Management Architecture (UIMA) framework and works with EHRs unstructured clinical texts. This is the most frequently used tool for IE (with 26 papers), according to Wang et al. (2018).

This system combines rule-based and machine learning techniques (hybrid system) and is composed of six modules: Sentence Detection, Tokenization, Normalization, Part-of-Speech Tagging, Shallow Parser and a NER annotator, according to Savova et al. (2010). In Figure 2.1 we can see all these modules acting individually. First, the system detects the sentence, next it divides it into its constituent tokens (11). After that it normalizes all the tokens, that is, transforms every word into its basic form (lemma). Next, it identifies the constituent parts of sentences as nouns, verbs and determinants, and, after that, links them into nodes related syntactically as noun groups and verb groups. Finally, cTAKES recognizes the entities and find if they are negated or not.

Example 2.1: Example taken from [5] that shows all the modules of the cTAKES system acting individually

To detect sentences, the system finds where the punctuation that implies the end of a sentence (full stop mark, question or exclamation marks) is. Relatively to the tokenizer it has two components: one which splits the sentence into spaces and punctuation and another which merges the tokens that do not make sense to be split (as in dates and hours, it does not make sense to split the expression “12:40h”
into “12”: “40” “h”). To normalize the words, the authors resorted to a dictionary of the SPECIALIST lexical tools\textsuperscript{11}, the Norm dictionary\textsuperscript{12}, which gives the lemma of the words. Next, the Part-of-Speech Tagging and the Shallow Parser are “wrappers around OpenNLP’s modules for these tasks” \textsuperscript{5}. Finally, the NER module uses dictionaries with concepts from the SNOMED CT\textsuperscript{13} and RxNORM\textsuperscript{14}.

The entities to be recognized are the following: disorders/diseases, symptoms/signs, procedures, anatomy and drugs. In relation to the negation attributes, cTAKES uses the NegEx algorithm \textsuperscript{31}, which finds if there is any negative term near to the entities that could negate them.

The cTAKES can adapt itself for various tasks as smoking status extraction and detection of medication discrepancies.

### 2.4 Overview

In this section, we explain why we choose these before mentioned systems and tools and how did they help us.

Relatively to the anonymization step, we concluded that:

- We will use the STRING chain because it is prepared to texts written in the Portuguese language and we have access to it;
- With the ARX tool, we showed that not the all texts need to come in an unstructured format to be possible to anonymize them. It also allows us to present usual privacy methods;
- The BoB system helps us to understand that, what is desired in this task is to obtain a good recall because of, higher the recall, more names are anonymized. The precision metric is not so relevant because we prefer to anonymize all the names even though some more terms come anonymized, instead of having a small number of wrong anonymizations but miss some names.

Regarding the Pre-Processing tasks:

- In the Spell Checking, we presented with two tools that the idea besides correcting words is the same;
- With the task of WSD, we inferred that is important to disambiguate the acronyms and abbreviates to obtain better results. We also concluded that it needs to be done basing on the context and, if they are ambiguous, using rules.

Although we have demonstrated some tools that apply IE in the clinical context, we are conscious that one of the biggest difficulties of this work is due to the text is written in the Portuguese language and the current state-of-art is bigger for English texts. However, we retrieved some ideas of these tools:

- The **AMBIT** and **Natural Language Text Processor for Clinical Radiology** tools show us that using a database that is formed by various medical terms it is possible to contour the problem of not having a “single nomenclature”;

- The **cTAKES** tool is the one in which we based more. It applies rules and has modules related to the ones that we want for our work. It has a way of acting similar to that of the STRING chain, helping us to understand what types of entities are essential to structure, as also to create a special attention for the case of the negation attributes.
# Data Analysis and Solution Proposal

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 The Medical Records</td>
<td>18</td>
</tr>
<tr>
<td>3.2 Methodology and Proposed Solution</td>
<td>25</td>
</tr>
</tbody>
</table>
The purpose of this chapter is to understand what the medical records are about and how the database is organized in a way to know and acquire knowledge about the topics that we want to use as entities. After this notion, we describe our method to obtain a structure from these texts.

### 3.1 The Medical Records

The study objects of this thesis are the EHRs about rheumatologic patients. SPR made available 8673 patients from Portuguese centers, from a total of 73221 doctor’s appointments.

These appointments are, always, about one of the following thirteen rheumatologic diseases (retrieved from the Reuma.pt demo database): “Artrite Idiopática Juvenil (AIJ)”, “Artrite Psoriática (AP)”, “Artrite Reumatóide (AR)”, “Artrites Iniciais”, “Esclerodermia”, “Espondilartrites”, “Lúpus Eritematoso Sistémico (LES)”, “Osteoartrose”, “Outros Diagnósticos (adultos)”, “Outros Diagnósticos (juvenis)”, “Síndrome de Sjögren (SS)”, “Síndromes Auto-Inflamatórios” and “Vasculites”. In Figure 3.1, we could see the percentage of these diseases on the data that SPR made available to us.

![Circular chart showing percentage of diseases](image)

**Example 3.1:** Circular chart that shows the percentage of each rheumatologic disease on our data

Comparing the results that we obtained on the chart (based on the available data from 2017), with the table from Canhão et al. (2012) that we presented on the Appendix A (from 2012, with some characteristics of the patients on that year), we concluded that the AR continues in the vast majority on the patients, following by the “Espondilartrites” (that are the same as “Espondilite Anquilosante”), but, now, the AP exists with a higher percentage than AIJ, on the Reuma.pt database, contrarily to the 2012 year. The LES also exists in a number of patients similar to the AIJ.
Below, we continued our analysis on the given structured data from the SPR. This analysis continues being based on the table of Appendix A. Thereupon, we analyzed what is the incidence of each of these diseases on the male (blue) and female (pink), using as an auxiliary way the chart from the Figure 3.2.

Example 3.2: Chart that shows the percentage of rheumatological diseases on male and female

In this case, our data fit right with the data from 2012. The AR continues to have more incidence on the women than the AIJ. It remains to be more incident that the AP, which, in its turn, have more females than the EA. The percentages values are also very similar on both years. Based on the chart we can still conclude that the SS occurs, practically, only in women and we still have above the AR, the Esclerodermia and the LES, with a higher percentage of female patients. The disease with the lower rate of females are the “Espondilartrite” (just under the 50%), but we can assume that all of these rheumatological diseases affect, undoubtedly, much more the females.

Relatively to the ages of the patients who have these diseases, we used as support the chart shown in the Figure 3.3.

Comparatively with the data from the Appendix A, we concluded that all of these diseases starting later (except the AIJ, in which we have one person with only 2 years old that have this disease). However, the diseases time interval, increases. This shows that is possible to have an advanced age with these diseases, despite decreasing, and much, the quality of life of the patients.

Lastly, the last analysis was made to the number of biologicals taken by rheumatologic disease. Again we use the attached table as a comparison and Figure 3.4 as support to the data from 2017.

We note that, although there has been a general increase on the percentages “com biológico” (with biological, the blue color on the figure), the order keeps the same, that is, it continues to be the EA the
Example 3.3: Chart that shows the ages of the patients with rheumatological diseases

Example 3.4: Chart that shows the percentage of biologicals taken by rheumatological disease

disease with more biologicals taken (with 55.2% in 2012 and 63% in 2017), following by the AP (with 50.89% in 2012 and 61% in 2017). Third, there is the AR (with 28.2% of biologicals taken in 2012 and 50% in 2017) and, lastly, the AIJ (with 22.28% in 2012 and 38% in 2017).
The EA is, currently, the disease that has the greater percentage of biologicals taken. “Artrites Inicias”, “Esclerodermia” and SS have the percentages practically nil, concerning to the biologicals taken.

Relatively to the clinical notes, in general, they are written in a different format of traditional text. Sentences are small, contains abbreviations and acronyms, some misspelled words, many of them appear in the negative form and do not have punctuation.

### 3.1.1 The Database

The data is described in an Excel document with 5 tables named “Doentes”, “Consultas”, “Terapêuticas”, “Problemas Clínicos” and “Eventos Adversos Graves”, showed on the Figure 3.5. The table “Doentes” lists all the patients with, at least, one consult with clinical notes. It contains general information about the patient as id (identifier of the patient on the medical database), which rheumatologic disease he/she has, sex, date of birth and the quantity of biologicals he/she consumes. Table “Consultas” contains the information collected during a medical appointment, on the field named “Observação”, in textual format. It contains information obtained by exams or by the health professionals perception of the patient disease. The focus of our work is to extract information from this field, using a methodology described, then, in Section 3.2. We can also see, in this table, the medical appointments dates and if the patient is taking biologicals (biological drugs). In “Terapêuticas”, we have access to all the therapy that the patient had and is still having, with the respective dosages and the end date (if finished), indicating the end reason. The table “Problemas Clínicos” has all the pathologies recorded for these patients, including serious adverse events. As in “Terapêuticas”, this table also has the active pathologies and the completed ones, indicating the ended date (if finished) and if it is a serious adverse event. The last table is “Eventos Adversos Graves”: It has the details of the registered serious adverse events for these patients, indicating some possible causes and treatments.

Note that ID_Doente is present in all tables, but some of them only create a relation between them if a specific field of the table has the diamond symbol as True, as in “Terapêuticas”. One patient can have one or more therapies if we have information in “Doentes” table that this patient has therapeutic data (“Dados Terapêuticos”=1). Only with this information exists a relation between these two tables. The same occurs with “Problemas Clínicos” table when we have information if the patient has pathologies (“Comorbilidades”=1) in “Doentes” table and with “Eventos Adversos Graves” table when we have the information that the patient has a serious adverse event (“Ev_AdversoGrave”=1) in “Problemas Clínicos” table.
3.1.2 Descriptive Analysis of the Field of Clinical Notes

The clinical notes will be our principal focus, so we have explored these texts to understand which are the themes that they address and to obtain an idea of the challenges of extracting information from this field, considering a NLP point of view.

After having these texts available, we conclude that, first of all, we need to anonymize them to occult the identity of the people involved. Usually, in the doctor’s appointments, they refer the colleague’s names, as we can see in Example 3.1, to know which doctor of a specific specialty follows the patient. The nurses names, or, less commonly, patient names, are also mentioned in these clinical notes.

Example 3.1: Portion of clinical text that refers to a name of a doctor

After that, we need to pre-process the texts because there is lack of accents and punctuation and writing errors (most of them are typographical – typos) as we could see in Example 3.2. For instance, “Erosoes” instead of “Erosões” and “radiocarpica” instead of “rádio-cárpica”. In addition, before “Sem intercorrências” and “Iniciou”, there should be a full stop mark. An example of a typo is “sianl” where the letters “a” and “n” are exchanged.

In this same example, there is a situation with which we will not worry about in this work. “A Enf. **Maria Silva** fez ecografia aos punhos” (the nurse Maria Silva did an ultrasound to the fists). In common sense, we realize that the nurse did the ultrasound to the fists of the patient, but the system does not understand that and will assume that the ultrasound to the fists was made to the nurse.

We also have acronyms and abbreviations in these texts that need to be deciphered to know to what terms doctors refer to. An example can be seen in Example 3.3. “VGM” is an acronym to “Volume Globular Médio” and “act doença” is an abbreviation to “actual doença”.

3.1.3 Description of Clinical Notes Topics

Relatively to what themes these clinical notes refer to, after some manual examination, we concluded that the most discussed topics are the following ones:

- **Symptoms**: Doctors refer to a symptom when they realize that there is any change of a person’s usual perception of his/her own body and sensations. Example: “Despertares noturnos por dores nos ombros” (shoulder pain);

- **Personal Life**: Doctors refer personal life situations that result in a modification of patient health. Example: “Marido caiu duma oliveira há 15 dias, a doente muito stressada” (Husband fell from an olive tree 15 days ago, patient very stressed);

- **Medical Exams and Results**: The texts have many references to diagnostic exams and their results. Example: “Ecografia joelho e ombro conf rota” (Ultrasound knee and shoulder conf break);
- **Medication Plan**: Doctors always have a portion of text related to the current drugs of the patient and the new changes in the medication plan. Example: “Plano: Aumentar a dose de folcicil 2cp/dia excepto no dia do MTX” (Increase the dose of folicil 2cp per day except in the day of MTX);

- **Results of blood tests**: The texts have references to relevant and important results of blood tests. Example: “Hg 11,8 VGM 86,9 Leucograma N Plaquetas 278000 VS 4 PCR 0,63 PFH N PFR N Urina II”;

- **Clinical Problems**: When the patient has serious adverse events, doctors recur to the clinical notes to refer to that. Example: “Tem outro nódulo sólido na mama esq que se desenvolveu em menos de um ano” (It has another solid nodule on the left breast that has developed in less than a year);

- **Daily Routine**: Doctors point out the daily routine of the patient that could or not improve his/her health. Example: “A fazer ginástio diariamente, muito motivada!” (To do gym daily, very motivated!);

- **Treatments**: The texts have the previous and current treatments of the patient. Example: “Sem outras queixas além de periartrites tratadas, com recurso a infiltrações” (No other complaints than treated periarthritis, with infiltrations).

The identification of these topics helps us to realize what the clinical notes are about and which entities we need to recognize in our texts. This will be explained in the next subsection.

### 3.1.4 Medical Entities to Recognize

The previous task about the identification of the topics discussed in the clinical notes made it easier to us to choose which entities we want to recognize, in a way to transform the unstructured text into a structured one, based on these entities. The entities chosen were the following:

- **Rheumatic Diseases**: It is important to separate the rheumatic diseases from the other clinical problems because the clinical notes are about rheumatic medical appointments, so it is essential to know which rheumatic disease the patient has. Therefore, we need to identify rheumatic diseases as “Artrite Reumatóide” (rheumatoid arthritis) and “Osteoporose” (osteoporosis).

- **Drugs**: The drugs are the most common topic among the clinical notes, so it is relevant to have a notion of what type of drugs are prescribed. The drugs mentioned are, for instance, “Actifed” and “Naprosyn”.
• **Drugs Brands:** Drugs brands are regularly referred to in these texts. We need to make a distinction between them, so we separate the drugs entities from the drugs brands entities. As drugs brands, we need to identify names as “Alpharma” and “Mylan”.

• **Clinical Problems:** As explained before, clinical problems refer to situations when the patient has serious adverse events. We need to recognize that. As clinical problems, we can accept, for instance, “lúpus” (lupus) or “taquicardia” (tachycardia).

• **Bacteria:** In these texts is also common to find bacteria names. We need to be able to identify names as “Bacillus” and “Pasteurella”.

• **Active Substances:** It is not only essential to distinguish between drugs and drugs brands, but also between the active substances present in drugs. Usually, the doctors write what are the active substance necessary to combat some clinical problem. Due to that, we want to identify names as “Aciclovir” and “Metyprednisolona”.

• **Human Body:** Frequently, parts of human body are referred in these texts. The more usual are bones, muscles and articulations as “fémur” (femur), “trapézio” (trapeze) and “metacarpo-falangeana” (metacarpophalangeal).

• **Hormones:** As the case of bacteria, hormones names also occur often. Then, we need to recognize names as “calcitonina” and “insulina”.

• **Diagnose:** When a patient complains of a given clinical problem or rheumatic disease, the doctor needs to do some diagnosis exams or tests, and we need to know what kind of diagnostics exist. Therefore, we need to recognize exams as “hemograma” (blood count) and “eletrocardiograma” (electrocardiogram).

• **Hospitals/Centers:** It is relevant to know in which medical center or hospital the patient has medical appointments and the doctors register this. So, we need to recognize hospitals as “Hospital Santa Maria” and “Hospital Egas Moniz” and centers as “Centro Clínico Champalimaud” and “Centro de Diagnóstico Pneumológico”.

• **Treatments:** Usually, at the end of the clinical note, the doctor registers what would be the treatments and possible surgeries that will try to solve the clinical problems of the patients. We need to recognize names as “cateterismo” (catheterization) and “quimioterapia” (chemotherapy).

### 3.2 Methodology and Proposed Solution

In this section we will describe the STRING chain tool and explain how we will deal with the problems and the challenges that we have identified, using an hybrid approach.
3.2.1 The STRING Chain

This is an in-house tool of Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID), the local where I developed my thesis. STRING was developed by Mamede et al. (2012) and is an hybrid, statistical and rule-based NLP chain for the Portuguese language. It adopts an architecture based on modules, as we can see in Figure 3.2 – taken from the STRING site\(^1\).

![STRING chain architecture](image)

**Figure 3.2**: STRING chain architecture

The first module receives the input text and separates it into their respective fragments (tokens). These tokens are the input from the second module that applies the LexMan\(^2\) morphological tagger to assign to every token its morphological category as name, adjective and verb.

The third module receives as input the segments grouped into sentences and consists of applying the RuDriCo\(^3\) to makes disambiguation and segmentation rules. The disambiguation is used when a word can be from two or more morphological types. The disambiguator needs to choose the correct category, considering the surrounding text. The segmentation rules can be subdivided into expansion

\(^1\)https://string.i2i.inesc-id.pt
\(^2\)https://string.i2i.inesc-id.pt/w/index.php/LexMan
\(^3\)https://string.i2i.inesc-id.pt/w/index.php/RuDriCo2
and contraction rules. Expansion rules consist of converting a segment into two or more fragments as transform “no” into “em” and “a”. Contraction rules are the opposite, they consist of transform two or more segments into only one as “artrite” and “reumatóide” into “artrite reumatóide”.

The last module applies the XIP4 parser for the syntactic analysis. This parser allows us to locally apply our disambiguation rules and grammar and also add lexical, syntactic and semantic information.

Hereupon, we can validate that STRING was very useful for this work because it performs NER and anonymization tasks and, as previously written, it enables us to add more semantic information, helping us to process our medical text. Another positive point is the fact that one possible output of this processing is in XML format which will be useful for the principal goal of this work: obtain a frame with structured text.

3.2.2 System Overview

We divide our solution into five steps, shown in Figure 3.3: Dealing with Acronyms, Anonymization, NER, Spell Checking and Structuring the Clinical Notes.

Primarily, we do a list of pairs acronym-extensive full form in a way to substitute the acronyms by their extensive full form in the original text. Secondly, we solve the anonymization step using and applying rules on the STRING chain. After that, we do lists of terms that compose all the entities mentioned before and insert it on the STRING chain. It helps us to easily recognize the entities. In the fourth step, to correct the writing errors in the clinical notes, we adopt an idea of Carvalho and Curto (2014), using a measure of similarity that compares our text with a medical dataset. Lastly, after having our problems solved and the clinical notes processed, we will use the output of STRING to structure these texts, obtaining the desired frames.

---

4https://string.l2f.inesc-id.pt/w/index.php/XIP
4

Processing the Clinical Notes

Contents

4.1 Dealing with Acronyms ......................................................... 29
4.2 Anonymization ................................................................. 30
4.3 Named Entity Recognition .................................................. 34
4.4 Spell Checking ............................................................... 37
4.5 Structuring the Clinical Notes ............................................. 41
In this chapter, we will explain in detail the five steps that we take until we reach our final goal, that is, structure the text of the clinical notes. Whenever feasible, we will follow the same structure to illustrate these steps: first, we explain the challenges of each step, next we describe our approach to solve these challenges and, finally, we perform a preliminary evaluation in some steps.

4.1 Dealing with Acronyms

In this first step of our work we made a list of acronyms and their respective expansion in full form. Although in STRING it is possible to introduce acronyms, it would be damaging for the performance of the tool to introduce some medical acronyms. There are acronyms already inserted on the STRING that will create ambiguity with our medical acronyms, although they have different meanings. For instance, “AV” is inserted on the STRING, representing an abbreviation to “avenida” (avenue) and, in a clinical context, we want that it represents an acronym to “atrio ventricular” (ventricular atrium).

The best solution that we found to overcome this problem was to make this list of medical acronyms and replacing them in the text with their full form, before moving on to the next step. We based on Liu et al. (2002) idea of expansion of acronyms and abbreviations to do that.

4.1.1 Challenges

In addition to the problem that medical acronyms create ambiguity with “normal” acronyms on the STRING, medical acronyms have ambiguities between them too, as “AU” that can represent “altura uterina” (uterine height) or “ácido úrico” (uric acid). We need to know in which textual situation we use one or another. In this case, for instance, if in the medical record the doctor evaluated a situation of a pregnant patient and wrote about the pregnancy, it is likely that “AU” will mean “altura uterina”. However, if the doctor wrote about urine tests, we know that “AU”, in this context, will mean “ácido úrico”.

4.1.2 Approach

Firstly, as already mentioned, we gathered a list of acronyms with their respective expansion in full form. This list was made at the same time that we were studying the clinical notes to know what the medical records are about (Subsection 3.1.3). The final list is composed of about 450 acronyms. To understand their meaning to expand them in full form, we used the context in which they appeared. Some of them are quite intuitive, as “Dirigi-se ao HEM para a consulta com médico de família” (Went to the HEM for consultation with a family doctor). In this case, it was intuitive for us that “HEM” refers to “Hospital Egas Moniz”.

Another way of knowing the extensive full form of acronyms was searching in the web, continuing
to take into account the context in which they appear. For instance, if the doctor referred to a drug's list taken by the patient and the acronym “SZP” appears, we know that we are dealing with the drug “Salazopirina”.

However, there were acronyms that we could not even understand their meaning because either they were very specific to that doctor’s writing style or were very specific to the field of rheumatology, which made them difficult to decode through the web. Hereupon, we sent a list, with about 100 acronyms, to the doctors who provided us the data of Reuma.pt, asking for help in interpreting what those acronyms mean, in extensive full form. They returned the list providing an answer to what we ask, but it was also difficult for them to decipher what their proper co-workers refer to when using certain acronyms.

We still have the case, although it only occurs about two or three times, in which the acronyms themselves contain a typographical error as “ENEA”. Actually what they wanted to write was “ANEa” – “Associação Nacional de Espondilite Anquilosante” (National Association of Ankylosing Spondylitis) – this example was part of the list that was sent by us.

Regarding the acronyms used when the blood and urine tests are being diagnosed, they are not changed to their extensive full form, they are being placed on a specific label (“SEM-analises”), which will be shown later in this document.

Relatively to the situation of the two ambiguity cases of medical acronyms that we previously explained, we solve that automatically replacing all the not ambiguous acronyms (fortunately they are the majority). After that, we manually replace the ones left over, case by case.

4.2 Anonymization

Anonymization is the step that carries with it more responsibility because we are in the presence of real data, which contain personal and confidential information. To avoid compromising the privacy of the people involved, we need to anonymize this data. However, for this step to be completed successfully we must take into account two aspects: in the end, there can be no proper names in the data, but, at the same time, we must ensure that there is no loss of information.

4.2.1 Anonymization using STRING

As written in the previous Subsection 2.1.1 this system receives Portuguese texts and is prepared to anonymize names, localizations and organizations. In our case, we only want it to anonymize names of people because it is the only sensitive information. So, we made small changes in the original code to reach our goal. Although the precision for Portuguese texts is 53%, the recall is 86% and the recall is the most valuable measure in this situation because we want to anonymize all the names, even if other terms are also anonymized.
There are four ways to anonymize text using STRING, which have already been described in the Related Work (chapter 2). In our work, we choose to randomize the name of the person, because, in the next steps, the system would have to know that this word was a person's name, having no loss of information.

As input, we pass the clinical note without acronyms (or with these acronyms in their extensive full form), and, as output, we obtain the text without any reference to someone. This step removes any confidential information or identification of the people involved, as we can see in Example 4.1.

Example 4.1: Clinical note already anonymized

"(...) Fez análises hoje. Tensão Arterial: 156/93mmHg. Frequência cardíaca: 93bc/minutos. Temperatura: 36,0°C. Seguido em consulta com a ["Maria"]). Próxima avaliação agendada para o dia 1/10/2014"

4.2.2 Challenges

Various problems of bad anonymizations were detected and were due to:

1. **Medical writing style:** Several uses of acronyms and abbreviations that the system confuses with people names. Example: “ALT Frenal” is not introduced in the system, so it assumes that, as it starts with a capital letter, it is a people name;

2. **Lack of punctuation:** This is the less frequent problem. A new sentence starts with a capital word, but the previous sentence does not end with a full stop mark, so the system assumes that the capitalized word is a name of a person. Example: “Menos reação local (sob antihistaminico continuo) Transaminase Normal” becomes “Menos reação local (sob antihistaminico continuo) ["Mariana"] Normal” because there is a missing of a full stop mark before “Transaminase” word;

3. **Hospital and drugs names:** The system often considers as person’s names, hospital names like “Santa Maria” and “S. Francisco Xavier” and drug names as “Humira”;

4. **Portuguese specificities:** In some cases, the system confuses a Portuguese verb with a person name. An example is “Marco consulta” (Schedule a consultation). It appears several times and the system assumes that “Marco” is a male name (which is, indeed).

4.2.3 Approach

The first problem mentioned above is the more common and could be solved evaluating manually one iteration, registering the exact way that an error occurs, saving this in a file to, in the next iteration, use that as a filter to not commit the same mistake again. We have cases like “MT IFX” that was anonymized
and after this filter, it is no longer. But, we had no idea of the exhaustive amount of different acronyms used and the results were not good. We realized that, after a preliminary evaluation, the more important reason for a bad performance of our anonymizer is the fact that it anonymizes most of the acronyms that appear in the text, so we change the order of our schedule for the one shown on the Subsection 3.2.2.

Initially, we use as a schedule of our work the same schedule, but with the box of “Dealing with Acronyms” changed with the box “Named Entity Recognition”. With this schedule, we adopt an iterative method. However, it is time-consuming and, as we before described, obtains bad results. Despite the use of the mentioned file as a filter, the doctors have a writing style very different, and when we start to analyze texts of another doctor, the writing patterns are not present in the filter list. Since this process is very laborious and time-consuming, we only anonymize 15000 texts of 73221. To evaluate and facilitate this process, we did some iterations with a small number of clinical notes, to help in the construction of our filter list. In parallel, we created Table 4.1 that helped us in the evaluation task too.

As we described before, the main reason for this bad performance of anonymization is the use of many different acronyms in clinical notes. After understanding this, we change our schedule to obtain better results in the next iterations.

<table>
<thead>
<tr>
<th>Number of Clinical Notes</th>
<th>Anonymized</th>
<th>Total Anonymized</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>40</td>
<td>552</td>
</tr>
<tr>
<td>700</td>
<td>54</td>
<td>874</td>
</tr>
<tr>
<td>500</td>
<td>28</td>
<td>1124</td>
</tr>
<tr>
<td>1200</td>
<td>58</td>
<td>1408</td>
</tr>
<tr>
<td>700</td>
<td>52</td>
<td>437</td>
</tr>
<tr>
<td>1000</td>
<td>67</td>
<td>628</td>
</tr>
<tr>
<td>700</td>
<td>29</td>
<td>361</td>
</tr>
<tr>
<td>1900</td>
<td>99</td>
<td>1958</td>
</tr>
<tr>
<td>1300</td>
<td>150</td>
<td>1737</td>
</tr>
<tr>
<td>1100</td>
<td>87</td>
<td>739</td>
</tr>
<tr>
<td>1000</td>
<td>59</td>
<td>563</td>
</tr>
<tr>
<td>1000</td>
<td>84</td>
<td>686</td>
</tr>
<tr>
<td>1000</td>
<td>89</td>
<td>518</td>
</tr>
<tr>
<td>800</td>
<td>60</td>
<td>384</td>
</tr>
<tr>
<td>1200</td>
<td>105</td>
<td>780</td>
</tr>
</tbody>
</table>

Table 4.1: Table to help in anonymization evaluation

The first line of the table has a white space because these 300 initial clinical notes was used to manually annotate the text and start the creation of the filter list.

The second problem occurs a few times, so, although that there is no solution for it, it is not a big issue. The third problem was faster to solve because we use the lists already done to filter drugs and hospital names. The last problem also has no solution, because “Marco” could appear in many contexts and we cannot create a rule for that.
4.2.4 Preliminary Evaluation

To evaluate the performance of this task we based on Dias et al. (2016) article and use precision, recall and f-score metrics. This is the three evaluation metrics that are common in works like the ones that we present in the Related Work (chapter 2). We will use these metrics in the sections of preliminary evaluation throughout the work.

**Precision:** Number of predicted entity name spans that line up exactly with spans in the gold standard evaluation data (e.g. when “polimialgia”: disease and “reumática”: disease is predicted but “polimialgia reumática” was required, precision for the predicted disease is equal to zero). Precision is then averaged over all predicted entity names [28].

\[
\text{Precision} = \frac{TP}{(TP + FP)}
\]

**TP:** Terms classified by the system as positives, that actually are positives;

**FP:** Terms classified by the system as positives, that actually are negatives.

**Recall:** The recall is similarly the number of names in the gold standard that appear at the same location in the predictions (fraction of documents that are relevant to the query) [28].

\[
\text{Recall} = \frac{TP}{(TP + FN)}
\]

**FN:** Terms classified by the system as negatives, that actually are positives (the system should have detected, but did not detect).

**F-measure:** F-Measure score is the harmonic mean of these two measures.

\[
F\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}
\]

In the same article, they made a distinction on the evaluation of performance at the token and instance level. The token level is the most popular and the one that we will use to evaluate. The tokens are the base of evaluation (overall performance on the detection). In another approach, instance level evaluates the correct detection of each instance (underrates entities with partial errors in the classification).

We apply the precision formula at the token level to a gold-collection made manually, or in other words, to a collection of clinical notes annotated by us. This collection has 200 clinical notes and when we run that with the system and our actual schedule (without acronyms), we obtain 13 correct anonymizations, in a total of 39 anonymizations given by the system. We know that FP, in that situation, will be 26 (number of bad anonymizations) and FN is equal to 0 because never occurs the case that the
system fails to anonymize a person name. So, applying this information to the formulas it becomes:

\[
\text{Precision} = \frac{13}{13 + 26} = 0.3
\]

\[
\text{Recall} = \frac{13}{13 + 0} = 1
\]

\[
F - \text{measure} = 2 \times \frac{0.3 \times 1}{(0.3 + 1)} = 0.46
\]

Analyzing the results of performance that we obtain, makes sense that recall is equal to 1 because the system never misses a name. Out of curiosity, the results of the first approach would have been the following:

\[
\text{Precision} = \frac{13}{13 + 76} = 0.15
\]

\[
\text{Recall} = \frac{13}{13 + 0} = 1
\]

\[
F - \text{measure} = 2 \times \frac{0.15 \times 1}{(0.15 + 1)} = 0.26
\]

The actual precision doubled, although it remains small. By these results, we could prove that it was essential to change the approach.

4.3 Named Entity Recognition

In Chapter 3 we analyzed what kind of topics are discussed in the clinical notes. The solution that we found to recognize these topics and entities is through lists of terms. These lists were done based on external information (present on the web) and internal information (present on the database of Reuma.pt).

4.3.1 Challenges

When we started to gather medical named entities, we faced four challenges:

1. Medical terms are, usually, extensive and complex, so it is required a high capacity of terms recognition;

2. The task of medical terms recognition is also complicated because many terms can refer to the same situation, that is, they do not have a single nomenclature. Example: “Gota” and “artrite gotosa” (gouty arthritis) both refers to the same rheumatologic disease, but there are these two manners to say the same thing;
3. Acronyms are ambiguous as we explain in Section 4.1;

4. It is not easy to decide to which category a term belongs, because it can belong to more than one. In other words, there are ambiguous terms that we need to disambiguate, recording their categories. For instance, the term “Estradiol” can be a hormone or an active substance, and regarding the context in which this term appears, we need to assign it to the correct label.

4.3.2 Approach

Several lists were created, based on information present in the structured data of Reuma.pt. We generated lists of: rheumatic diseases (19 entries) with terms as “Gota” and “Síndrome de Sjögren”, clinical problems (1803 entries) with terms as “abcesso abdominal” (abdominal abscess) and “pancreatite” (pancreatitis), human body (174 entries) with terms related to bones and articulations as “joelho” (knee) and “metacarpofalangeana” (metacarpophalangeal) and diagnose exams and tests (104 entries) with terms as “electrocardiograma” (electrocardiogram) and “mammografia” (mammography).

After that, we gathered lists of: treatments (143 entries) with terms as “cateterismo” (catheterization) and “laparotomia” (laparotomy) and hospitals/centers (83 entries) with terms as “Hospital da Luz” and “Hospital de Braga”. These two lists were made based on both internal and external information.

The lists based on external information were created based on Wikipedia¹, on the general web and sites like the SPR² and Atlas da Saúde – Listas de Medicamentos do Infarmed³ (to retrieve the medicines names). The gathered lists based on only external information are: drugs (1314 entries) with unigrams as “Clavamox” and “Ibuprofeno”, drugs brands (160 entries) with unigrams as “Omezolan” and “Ratiopharm”, bacteria (178 entries) with unigrams as “Enterococcus” and “Yersinia”, active substances on medicines (717 entries) as “Naloxona” and “Tramadol” and hormones (83 entries) as “Prostaciclinina”.

Lastly, we end this point with an exhaustive list of medical acronyms and abbreviations present in these texts (492 entries) with their respective extensive full form. This list contains pairs as “ADM – Amplitude de Movimento” (Range of Motion) and “SZP – Salazopirina”. These two lists were made before we changed our architecture, but the time spent to do that was availed because the process made it easier for us to replace the acronyms/abbreviations by their full form.

We had to do all these lists based on medical terms because the STRING tool was not ready for texts written in a clinical context. All these new terms needed to be inserted before we started to analyze anything and before we perform NER with STRING.

¹https://www.wikipedia.org/
²http://www.spreumatoologia.pt/
³http://www.atlasdasauade.pt/lista-de-medicamentos-infarmed
4.3.3 Preliminary Evaluation

To evaluate these lists, we manually created lists with exhaustive terms for any one of these entities, changing their genres and numbers in a way to guarantee that the plurals, singulars, feminines and masculines (if they exist) are also recognized. These lists also ensure that the entity is attributed to the category it belongs to. For instance, “flu” is not a rheumatic disease and the system needs to know that and assigns it to a symptom.

After a first evaluation using our lists as input on the STRING, we obtained the following results:

<table>
<thead>
<tr>
<th>Lists</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterias</td>
<td>99</td>
</tr>
<tr>
<td>Human Body</td>
<td>75</td>
</tr>
<tr>
<td>Diagnose</td>
<td>98</td>
</tr>
<tr>
<td>Rheumatic Diseases</td>
<td>85</td>
</tr>
<tr>
<td>Drugs</td>
<td>99</td>
</tr>
<tr>
<td>Hormones</td>
<td>78</td>
</tr>
<tr>
<td>Hospitals</td>
<td>100</td>
</tr>
<tr>
<td>Drugs Brands</td>
<td>92</td>
</tr>
<tr>
<td>Active Substances</td>
<td>97</td>
</tr>
<tr>
<td>Clinical Problems</td>
<td>98</td>
</tr>
<tr>
<td>Treatments</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.2: Precision results of the initial evaluation of lists

With this first evaluation, we concluded that we need to add more rules and dependencies in specific situations, using the STRING chain. The syntax of the rules on STRING is shown in Figure 4.2. In this figure, we can see that the rules have a number (each one has a unique number). We specify what tag we want for the sequence on the “Tag that we want for the sequence” field and, next to the “=” symbol we have the sequence of the tags that we want to attribute to just one tag.

Example 4.2: Syntax of a rule on the STRING chain

Relatively to dependencies, we present one in Example 4.3. We created it because on STRING “proton” refers to “protão”, but in a medical sense we want it to refer to a drug name, so we specify that if we find a common name “proton” we want that it becomes equal to a name, namely a drug name.

Example 4.3: Rule that chooses only one label to a word

We created a total of 75 rules due to diverse situations. For instance, as we are doing text analysis and not text generation, we assume that the doctors do not write “abnormal” things as “nariz direito”
(right nose) or “fígado esquerdo” (left liver). So, we wrote the rule in Example 4.4, that says that if we found any term that has the tag “SEM-corpoHumano” (Human Body) and it has “direito” (right) forward, we continue in a presence of a human body part. We do the same rule but with the “esquerdo” (left) term too, considering that any part of the human body could have a right and a left side.

Example 4.4: Rule about the human body

Another important rule can be seen in Example 4.5 that says that if we read the lemma “dor” (pain) followed by a preposition “em” or “de” with an optional definite article and ending with a human body term, we are in the presence of a diagnose situation because we obtain sequences as “dor no peito” (chest pain) or “dor de cabeça” (headache).

Example 4.5: Rule about a situation of diagnosing

The last rule that we show is presented in Example 4.6. Firstly, we attribute to all the acronyms relatively to blood tests the tag “SEM-analises”. Acronyms as “GGT”, “AST” or “LDL” are very common in blood tests and in our texts they appear often. With this rule, when we find an acronym with the tag “SEM-analises”, followed by a digit, we have a good chance that this sequence will be an analysis to the blood tests of the patient. An example of this sequence could be “LDH 247” or “GGT 44”.

Example 4.6: Rule about diagnose with blood tests

4.4 Spell Checking

In the medical texts, typographical errors (as substitution of one letter for another, transposition and omission) are usual. In the Ruch et al. (2003) article they conclude that the incidence of misspellings in clinical notes is around 10%, which is significantly higher than the incidence of misspellings in other types of texts. With that information, we have certainties that this task will be necessary to reach our goal.

4.4.1 Challenges

To correct these typos, we adopted the idea of Carvalho and Curto (2014), explained on the Related Work (Chapter 2). They created a semi-automatic error detection based on a bag-of-words model. For
that, they used two lists as resources. The first list is the Corpus Words List (CWL) and is composed of all the different existent unigrams on the reports. The other list is the Known Words List (KWL) that is a file of well-written common words of the domain (in this case, medical domain) and common words in Portuguese.

For that, firstly, we need to do a script that returns all the different existent unigrams that constitute the reports, creating the CWL. To create the KWL we download a corpus of Portuguese text from Linguateca⁴. Next, we run the script that returns the uniques unigrams from this corpus. We add to the unigrams list the terms from our drugs, brand drugs and active substances lists because these terms are very usual on the texts, and they are not present on the Linguateca corpus.

The final list of KWL has a total of almost 123M terms and the final list of CWL has a total of almost 56K terms. We only considered the unigrams with more than 3 letters and less than 15 letters and did not consider unigrams that contain numbers as in the Carvalho and Curto (2014) article.

**4.4.2 Approach**

First of all, we needed to know which unigrams are well written and which are not. For that, we compared all the unigrams of the CWL with all the unigrams of the KWL. If they are equal, it means that they are well written. After this comparison, we obtained a list of 16036 correct unigrams. However, it does not imply that all the others are wrongly written: they could only be out of our KWL list.

Then, we made a count of all the unigrams of the remaining CWL list (39793 unigrams) – f-CWL of the figure 4.1.

---

**Figure 4.1:** Schema of the writing errors correction

<table>
<thead>
<tr>
<th>KWL</th>
<th>Known Words List</th>
</tr>
</thead>
<tbody>
<tr>
<td>CWL</td>
<td>Corpus Words List</td>
</tr>
<tr>
<td>f-CWL</td>
<td>filtered Corpus Words List</td>
</tr>
<tr>
<td>HFL</td>
<td>High Frequency List</td>
</tr>
<tr>
<td>LFL</td>
<td>Low Frequency List</td>
</tr>
</tbody>
</table>

⁴https://www.linguateca.pt/
We concluded that the terms that usually appear (with the highest score) only have accentuation problems or did not appear in our Medical Dataset. Regarding again the Carvalho and Curto (2014) article, they made the following assumptions:

- The unigrams that appear 5 or fewer times (with the count less than or equal to 5) are wrongly written (LFL), so we choose to apply the Jaro distance between them and the Medical Dataset;
- The unigrams that arise more than 5 times may not have been caught by the algorithm because they do not appear on the Medical Dataset. These unigrams are on the HFL.

The unigrams that appear on the LFL will be replaced by the unigrams that have higher similarity according to the Jaro measure.

4.4.3 Preliminary Evaluation

The 73221 original clinical notes contain 55829 unigrams of which 40510 are possibly misspelled. Firstly, we applied this algorithm to the unigrams that appear less than 5 times (including). In this situation, there are a total of 35028 different unigrams that are possibly misspelled. With this, we split the dataset into files with 7000 lines and pass this data as input, separately.

Table 4.3 shows the lines group that we use as input in the algorithm and the number of best suggestions that we obtained. The best suggestions are the words, found on the Portuguese dataset, that have a similarity, higher than the Threshold T2, with our word.

<table>
<thead>
<tr>
<th>Lines Group</th>
<th>Best Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6999 (7000)</td>
<td>726</td>
</tr>
<tr>
<td>6999-13999 (7000)</td>
<td>5070</td>
</tr>
<tr>
<td>14000-20999 (7000)</td>
<td>4995</td>
</tr>
<tr>
<td>21000-27999 (7000)</td>
<td>5147</td>
</tr>
<tr>
<td>28000-35028 (7028)</td>
<td>5371</td>
</tr>
</tbody>
</table>

Table 4.3: Number of best suggestions to the word change

After that, we applied this algorithm to the 5482 remaining unigrams that occur more than 5 times. Here, we compare these unigrams not only with our dataset but also with our drugs, drugs brands and active substances lists. It gives us a total of 4621 best suggestions.

Of this total of 25930 best suggestions, we obtained 21240 right substitutions (the right substitution for a word is the maximum similarity that we found between the best suggestions for that word). For instance, in Example 4.6 we present the best suggestions to the “analises” word, according to our algorithm. It gave us 14 best suggestions (the ones that obtain a similarity greater than the defined threshold), but the right substitution to the word “analises” will be the best substitution with the greater value. In this example is “análises” with a similarity of among 0.92.
Example 4.6: Best suggestions to the “analises” word

We only executed this algorithm to our 15000 records of anonymized data. These data contain 24169 different unigrams. The 73221 original clinical notes, contain 21240 right substitutions, as we wrote. In common, the 15000 data and the 21240 right substitutions have 5376 unigrams, that is, there are 5376 different terms that are misspelled in these 15000 clinical notes.

For these 5376 unigrams, the ones that occur less or equal than 5 times originated a total of 5744 substitutions in the anonymized data. For the remaining unigrams (the ones that occur more than 5 times), we found that for the 4621 best suggestions originated by them, 3522 are the right substitution, according to the algorithm. For the 3522 right substitutions, we substitute in the original text (15000 data anonymized) 39559 times, because these misspelled words usually occur in the text.

Regarding the preliminary evaluation, what we did was an intrinsic evaluation in which we had to look at the data and verify if the suggested substitution is correct. For this, we randomly choose 300 substitutions of the 5376 suggested. Of these 300 suggested changes, 66 are not being correct (for instance “['dentario', ‘dentro’, 0.9166666666666666]”), which gives a precision of 0.78.

\[
\text{Precision} = \frac{234}{(234 + 66)} = 0.78
\]

There are terms that do not receive any best suggestion because:

• None of the words of the dataset have a similarity with them, higher than the threshold. It can occur because the well-written word may not exist in our Portuguese dataset;

• The fact that some substitutions are not correct may also be due to the same case. For example, “['invaginacao', ‘inalacao’, 0.8371212121212121]” (“invagination’, ‘inhalation’, 0.8371212121212121)). “invaginacao” is the already well-written word but, as the dataset does not contain this word, the algorithm suggests the most similar word to that, that is “inalacao”.

• The Jaro measure assigns more weight by making substitutions for accented letters than removing
letters. For instance, “[‘comunicacao’, ‘comunica’, 0.9090909090909092]”. The right substitution will be “comunicação” (comunication), but as taking 3 characters is heavier than adding 2 accented letters, the distance between these words would be smaller, which gives a greater similarity.

The recall is equal to one because the system is never wrong to classify a term as correct, that is, it never assumes that a misspelled concept is correct.

4.5 Structuring the Clinical Notes

In this section, we will describe the last task of this work. Structured data enables us to quickly access the relevant information that we want to analyze and perhaps it becomes crucial in a medical context. We present how we converted the free text of the clinical notes into a structured text, resorting to the output of the STRING chain.

4.5.1 Challenges

As we wrote on the System Overview Section (3.2.2), in this task we need the output of the STRING chain, obtained when we process the text with the Anonymization (Chapter 4.2) and Spell Checking (Chapter 4.4) tasks already done. The STRING chain has several types of outputs as syntactic trees, information of the syntactic nodes, showing the dependencies between the words and XML.

To our work and to obtain our final output of the structured text, the simplest and best to use from all the outputs that the STRING provides us is the XML output. The text in XML format is semi-structured (since that this is a semi-structured language – Example 4.7) and would be only necessary to develop a code that parses this XML file and transforms it into a structured text in the desired format. The XML text already coming with the associated tags, as we could see in Example 4.7 when processing the text “O Manuel tem artrite reumatóide” (Manuel has rheumatoid arthritis). In this example, we can see that STRING treats the words or the set of words as nodes. On this segment of the XML output, the node starts on the 13th index and end on the 30th. Inside this node, there are node features as the grammatical analysis of the word (noun). Still, inside this, there is a reference to this node as a noun and exactly in these features there is an attribute saying that this has a tag “SEM-disease”, indicating that the token who starts on the index number 13 and end on the 30th index, is a rheumatologic disease.

We had already developed all the Python code in a way to pass as input our text that is of the form of this XML output shown in Example 4.7, transforming it in a structured format. Unfortunately, there was a bug on the STRING when we process some text that uses the rules made by us (some of them are shown on the Chapter 4.3.3). This text did not come with the associated tags that we define, despite the rules triggering in these cases. In Example 4.8, we could see a wrong XML output.
Example 4.7: Portion of XML output when we process the text “O Manuel tem artrite reumatoide”

Example 4.8: Wrong XML output from the STRING chain

With this problem, we have two solutions. The first one would be, as we already have all the XML code developed, process the text and structure it, ignoring all the rules done. Another solution would be to see if the output would come well and to consider the rules, in other types of output.

The solution chosen by us was the second one, because, although we already have all the python
code to process the XML output done, we would only catch basic terms inserted on the STRING and, what interests us, is the combination of these terms with another terms, catching that using the rules.

When we process the text “O João tem um tumor maligno na perna” (João has a malignant tumor in his leg) we could see that the system considers only as node “tumor maligno” (malignant tumor) and we had made a rule (see Example 4.9) that says that if we found a clinical problem followed by a preposition and an article, ended with a human body part, we continue to have a clinical problem.

> 214 noun[sem-problema=+] = noun[sem-problema], prep, art, noun[sem-corpoHumano].

**Example 4.9:** Rule to a clinical problem

Thus, we had to parse the output in another format. Fortunately, using other outputs we can obtain the right results. In Example 4.10 we could see the output that we choose to parse, that shows the dependencies between the words. The red fragment is the difference between this output and the wrong XML output because it assigns to the “NE_SEM_PROBLEMA” tag all the clinical problem, including the human body part, respecting the rule above.

![XML output example](image)

**Example 4.10:** Output chosen, showing the dependencies between the words

### 4.5.2 Approach

Although the chosen output is not intuitive and simple to process and parse, this output considers all the rules made by us, correctly associating the tags that we specify in the rules.

We develop a code to parse this output, in the Python programming language, as a way to catch the
tags that we want to use to structure, as well as the associated text to this tags. We want to do frames by clinical notes. The final and desired frame to the text previous used as an example (“O João tem um tumor maligno na perna”) is shown in Example 4.11.

Example 4.11: Frame of the output chosen

After we have developed an initial version of our code to process the output of the dependencies and relations between the terms, we test, passing as input a small number of clinical notes and we conclude that we already have some terms to introduce on the STRING chain, that we do not notice before, as diagnose terms like “colonoscopia” (colonoscopy) and “endoscopia” (endoscopy), drugs names as “Fucithalmic” and “Pneumo 23”, among others.

Furthermore, we still need to improve our grammar, adding more rules, to get the desired output without losing relevant information for the health of the patient. In the total were added 244 new rules. The new rules added in this step cover specific situations. We will explain some of them, showing their importance.

First, if we do not take into account words that appear before or after of the drugs names recognized by the system, the final frame appears with many errors and failures. For instance, if it is written in a clinical note “Interrompe Enbrel” (Interrupt Enbrel) and we do not add any new rule, we will obtain as final output “Fármaco: Enbrel” (Drug: Enbrel). This information will come wrong to someone that analyze this structure because what actually is in the running text is that the patient stop taking the drug, so we need to do a rule to this case, showed on the Example 4.12.

Example 4.12: Rule to structure drugs when the patient stops taking them

Secondly, considering the dosages and quantities taken, it occurs in the clinical notes that the drug name is associated to its dosage. For instance, “Metotrexato 7.5mg” or yet the indication of a reduction of a dose or change of medication as “reduzir dose de Tocilizumab” (reduce dose of Tocilizumab) or “substituo omeprazol por esomeprazol” (substitute omeprazol by esomeprazol). All of these situations need to be considered and properly structured, so we made a set of rules to it (see Example 4.13).

So, to “Metotrexato 7.5mg” instead of arising only “Fármaco: Metotrexato”, already will arise “Fármaco: Metotrexato 7.5mg”. To the other referred situation will appear “Fármaco: reduzir dose de Tocilizumab” instead of only “Fármaco: Tocilizumab”. Other important rules were made related to drugs as the type of administration that they can have, like “intramuscular, comprimidos, endovenoso ou via subcutânea”
Example 4.13: Rules related to drugs

(intramuscularly, pills, intravenous or subcutaneous way).

In addition to rules related to drugs, rules related to symptoms and clinical problems were also added. For instance, with the sentence “crises de tensão arterial” (blood pressure crisis), without any rule added for this specific situation, the output will become “Diagnóstico: tensão arterial”, when, actually, what we want would be “Sintoma: crises de tensão arterial”. To this specific situation, we made the rule shown on Example 4.14.

Example 4.14: Rule to accept “crises de tensão arterial” as a symptom

Another usual situation which sometimes occur, facilitating our analysis is when the health professionals write “Sintomas: ” and forward of this, they list all the symptoms that the patient has. Some of the listed symptoms cannot be caught by STRING or by our rules, but it is possible to develop specific rules that, when in the text the word “Sintomas” (Symptoms) appear, followed by colon (“:”), it associates all the words that appearing forward of this until reach a full stop mark, to symptoms. This rule is shown on Example 4.15.

Example 4.15: Rule to accept all the symptoms that appear followed by “Sintomas:” until a full stop mark

Rules to structure diagnosis were also developed. Rules to catch excerpts as “Tensão Arterial: 156/93 mmHg” in which we specify that, as in the previous example, when in the text appear the sequence of words “Tensão Arterial” (blood pressure), followed by colon (“:”), it associates all that appearing forward of this until appear a full stop mark, to a patient diagnose (see Example 4.16).

Example 4.16: Rule to accept the words that appear followed by diagnoses and “:” until a full stop mark

Another example that shows that, with the help of the rules, the interpretation made can vary is shown on 4.17.
Example 4.17: Rule to help on interpretation of diagnostic tests

When the doctor writes “Pec¸o an´alises” (ask for blood tests), without the mentioned rule, the output is “Diagn´ostico: an ´alises” (Diagnose: blood tests), what insinuate that the doctor was using some blood tests already realized by the patient to diagnose him/her, when, actually, what the doctor made was to ask blood tests to be able to make a better diagnosis, in the future.

Lastly, we already made rules to another tag used by us that is the “sem-tratamento” tag (treatment). For instance, “cateterismo direito” (catheterization on the right side) with the rule shown on 4.18 is already being accepted.

Example 4.18: Rule to treatments

Another situation that has to be considered is related to the negative attributes. This is a very specific case that had to be treated very carefully because, as we know, the negative changes completely the sense of the expression. Sentences as “n˜ao cumpre a dose de Metotrexato” (do not take the Metotrexato dose), using the already done rules to drugs and structuring it becomes “F´armaco: cumpre a dose de Metotrexato”. However, the term “n˜ao” appearing before this expression, changing it completely. That is, the final output needs to be “F´armaco: n˜ao cumpre a dose de Metotrexato”. Thus, we made specific rules to the negative, shown on the Example 4.19, which covers the more casual negative expressions.

Example 4.19: Rules applied to the negative

4.5.3 Preliminary Evaluation

For the preliminary evaluation, we randomly choose a set of 50 clinical notes, from the 15000 that were previously anonymized and spell checked, as well as with all the acronyms replaced by their extensive full form. In order to be possible to evaluate if our final structure comes in a complete format, without losing relevant information for the patient health, a gold-collection was done by us. This is based on the entities that should be recognized.
This randomly chosen text to structure is composed by a total of 2319 unigrams. Structuring that, we obtained the following number of entities, shown on Table 4.4.

<table>
<thead>
<tr>
<th>Entities</th>
<th>Number of Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fármacos</td>
<td>51</td>
</tr>
<tr>
<td>Doenças Reumatológicas</td>
<td>10</td>
</tr>
<tr>
<td>Diagnóstico</td>
<td>82</td>
</tr>
<tr>
<td>Bactéria</td>
<td>1</td>
</tr>
<tr>
<td>Hormona</td>
<td>0</td>
</tr>
<tr>
<td>Hospital</td>
<td>4</td>
</tr>
<tr>
<td>Sintoma/Problema Clínico</td>
<td>121</td>
</tr>
<tr>
<td>Tratamento</td>
<td>17</td>
</tr>
<tr>
<td>Substância Ativa</td>
<td>40</td>
</tr>
<tr>
<td>Marca de Fármaco</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.4: Number of entities on the preliminary evaluation

On Table 4.5 we could see the entities given by the STRING chain on the column *Number of Entities given by the STRING chain*, evaluating it using the precision and recall metrics. As can be seen, we evaluate, separately, each entity. To calculate the Precision and Recall metrics to each of these entities, we based on Table 4.6 that shows the values of True Positive (TP), False Positive (FP) and False Negative (FN) values.

<table>
<thead>
<tr>
<th>Entities</th>
<th>Number of Entities on the Gold-Collection</th>
<th>Number of Entities given by the STRING Chain</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fármacos</td>
<td>51</td>
<td>36</td>
<td>100</td>
<td>70.1</td>
</tr>
<tr>
<td>Doenças Reumatológicas</td>
<td>10</td>
<td>9</td>
<td>100</td>
<td>90</td>
</tr>
<tr>
<td>Diagnóstico</td>
<td>82</td>
<td>85</td>
<td>89.4</td>
<td>92.7</td>
</tr>
<tr>
<td>Bactéria</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Hormona</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Hospital</td>
<td>4</td>
<td>4</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Sintoma/Problema Clínico</td>
<td>121</td>
<td>66</td>
<td>95.5</td>
<td>52.1</td>
</tr>
<tr>
<td>Tratamento</td>
<td>17</td>
<td>10</td>
<td>100</td>
<td>58.8</td>
</tr>
<tr>
<td>Substância Ativa</td>
<td>40</td>
<td>26</td>
<td>100</td>
<td>65</td>
</tr>
<tr>
<td>Marca de Fármaco</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.5: Precision and Recall metrics applied to the structure on the preliminary evaluation

Analyzing the results, only on “Diagnóstico” (diagnostic) and “Sintoma/Problema Clínico” (symptom/-clinical problem) labels the precision is different of 100%. In diagnostic it occurs because, sometimes, the system assumes as diagnostic more than what we supposed to and it divides what we assume to be just a diagnostic in two or more, or divides a symptom into a symptom and a diagnostic, as we can see in Examples 4.20 and 4.21. Comparing our desired output with the output obtained, we can observe that on Example 4.20 (our wanted output) we assume as a symptom “Tenossinovite dos extensores ao nível do punho direito” and, on Example 4.21 (the output given by the algorithm), the system gives as a symptom “Tenossinovite” and assumes, badly, the remaining as a diagnostic.
<table>
<thead>
<tr>
<th>Entities</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fármacos</td>
<td>36</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Doenças Reumatológicas</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Diagnóstico</td>
<td>76</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Bactéria</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hormona</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hospital</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sintoma/Problema Clínico</td>
<td>63</td>
<td>3</td>
<td>58</td>
</tr>
<tr>
<td>Tratamento</td>
<td>10</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Substância Ativa</td>
<td>26</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Marca de Fármaco</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.6: Auxiliary table to help in Precision and Recall results

Example 4.20: Manual frame of the clinical note 26

Example 4.21: Output to the clinical note 26

Recall is of 100% on the most basic entities, in which is practically impossible for the system to confound itself because are very specific terms and the doctors do not have many ways to vary the style of writing to refer to them.

In the drugs case, we obtained a recall of 70.1% because we were not prepared for the writing variance of the doctors. They have many ways to refer to the drugs and to their quantities and ways to consume. For instance, if the doctor writes “Metotrexato de 4 para 2 comprimidos por semana”, the STRING chain did not detect that until the rule 4.22 have been made.

Example 4.22: Rule to accept drugs in a specific format

Concerning rheumatologic diseases and diagnostics, they obtain a recall very close to 100%, so we do not exemplify the specific situation in which they fail.

With this preliminary evaluation, were added to the STRING chain:

- **Three rules to treatments.** For instance, “vacina da gripe” (flu’s vaccine) was not being identified until we made the rule shown in Example 4.23. We only had considered situations in which the doctors wrote “vacina gripe” and with the compound preposition “da” the STRING chain did not catch this expression;
Example 4.23: Rule to accept “vacina da gripe” as a treatment

• Five rules to active substances. For instance, we forgot the simple case in which the doctors could write “aumento de Prednisolona” (increase of Prednisolona). In order to not miss this case, we made the rule shown on Example 4.24, in which we specify that if we found the verb “aumentar” (increase), following by a preposition and, if at the end of this sequence, appears an active substance, we are in the presence of an active substance.

Example 4.24: Rule to accept “aumento de Prednisolona” as an active substance

• Nine rules to drugs. For instance, “Retoma lepicortinolo 5 mg 12/12 horas” led to “lepicortinolo 5 mg” and we want a more complete output. For that we made the rule shown on Example 4.25.

With that, we can show that some of these added rules are not only because the STRING does not catch all the terms, but also to complete the output given, in a way to not compromise the patient health.

Example 4.25: Rule to accept “Retoma lepicortinolo 5 mg 12/12 horas” as a drug

• Eight rules to diagnostic. For instance, as in the example of the rule added to treatments, we found on the clinical notes “raio-x da bacia” and was just being accepted “raio-x bacia”, until we made the rule shown on the Example 4.26.

Example 4.26: Rule to accept “raio-x da bacia” as a diagnose

• Twenty four rules to symptoms. This rules are very specific for some circumstances. In the example 4.27 we show one of them. If we found “aumento progressivo” (progressive increase), followed by a compound proposition, the word “nível”, a preposition, ended with a term relatively to blood tests, we are in a presence of a symptom/clinical problem. An example of a sentence that could be catched by this rule is “Aumento progressivo dos níveis de AST.” (Progressive increase of the AST levels).

The active substances have a recall lower than the drugs because there are drugs that can also be an active substance and a drug, so, in this cases, the STRING chain choose only the drug to give as
Example 4.27: Rule to accept a symptom/clinical problem

output and we, on the gold-collection, considered to return both as output.

A general problem is the fact that, sometimes, the system does not catch certain terms because they returned misspelled from the previous step. For instance, “intercorrências infecciosas” are not corrected on the previous step so, the STRING chain will not assume it as a symptom, because it is misspelled (the correct way is “intercorrências infecciosas”). We consider it as a symptom on the gold-collection and it origins a big portion of lower recall.

Another problem that we found on this step was due to the order that the structure takes. It returns a different order structured than the real order of the running text. We can prove that with the following examples. In Example 4.28, we show the clinical note 10 in running text. In the Example 4.29, we show the desired output, with the order of structure according to the order in which the terms appear in the running text. In the Example 4.30 is shown the output given by our algorithm and we can see the differences in terms of the order, comparing with the previous example.

This problem can not be solved because it comes from the STRING manner to give the output. It only becomes a problem because the history of the patient on the clinical note could have an order and that order sometimes could be missed with this feature of the STRING.
5

Experimental Results

<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Evaluation Corpora</td>
<td>52</td>
</tr>
<tr>
<td>5.2 Results</td>
<td>55</td>
</tr>
</tbody>
</table>
In this chapter, we describe the evaluation corpora built and used by us to assess the performance of our last four steps of this work. For that we show a general analysis of it as the one made on Section 3.1, the number of unigrams that it contains, as well as all the transformations done manually until we reach the final structure that we want. We also show the final results of our system and compare that with the preliminary evaluation made before.

## 5.1 Evaluation Corpora

We used for evaluation a portion of text never seen before by us, from the data of Reuma.pt. 300 clinical notes were randomly chosen of the remaining non-annotated data (58221 clinical notes). We prepared this text as a gold collection, that is, we annotate, for each of these four steps, the desired output.

Starting from the general analysis of the evaluation corpora and following the order from Section 3.1, we have on Figure 5.1 the percentage of rheumatological diseases that exist on these test data.

Example 5.1: Chart that shows the percentage of rheumatological diseases on the evaluation corpora

Comparing this chart with the previous chart from Figure 3.1, we can note that the slices of each rheumatological disease maintain practically the same values. The “Artrite Reumatóide (AR)” continues to be the disease with a greater impact on the patients, being the “Espondilartrites”, the “Artrite Psoriática (AP)” and the “Lúpus Eritematoso Sistémico (LES)”, the following with more impact.

Figure 5.2 represents a chart that illustrates the analysis of the incidence of these diseases on both male and female, in a way to be possible to compare it with the chart from Figure 3.2.

On these evaluation corpora, we are faced with a different situation from the general data regarding...
Example 5.2: Chart that shows the percentage of rheumatological diseases on the evaluation corpora on male and female

the most affected sex by rheumatologic diseases. In our evaluation corpora the “Osteoartrose”, the “Síndromes Auto-Inflamatórios” and the “Vasculites” are just diseases that belong to men, while, on the general data, the “Osteoartrose” was one of the diseases with more percentage for women. The remaining percentages of our evaluation corpora maintain very similar to the ones of the Chart 3.2.

Turning now to the ages of those who have these rheumatological diseases and taking into account a chart previously made (Figure 3.3), we made the following Figure 5.3 basing on the evaluation corpora.

Comparing this chart with the one from Section 3.1, as expected, we are working with a smaller age range. This means that, overall, we have a lower average age. On our evaluation corpora, the “Artrites Iniciais” is the disease whose average age is greater. The “Artrite Reumatóide (AR)”, only appears as third.

We end, as on Section 3.1, with the percentage of biologicals taken by rheumatological disease. In figure 5.4 we show this only for our evaluation corpora.

Comparing this figure with the Chart 3.4, there is a big difference between both with regard to the “Síndromes Auto-Inflamatórios” and “Vasculites” diseases. On our evaluation corpora, 100% of the patients are taking biologicals, but only exists one patient per each one of this diseases, so these results do not have validity. The entry “Outros Diagnósticos (adultos)” also has a significant rise. The remaining percentual values maintain the same.

Moving on to an analysis related to the unigrams in terms of quantity. These corpora are composed of a total of 16619 words. From these was 12146 fulfill the requirements that we stipulate of ideal size (between 3 to 15 characters). Only 2817 of these unique words that meet the requirements are different
Example 5.3: Chart that shows the ages of the patients from the evaluation corpora

Example 5.4: Chart that shows the percentage of biologicals taken by rheumatological disease on the evaluation corpora

from each other.

Relatively to the transformations manually done, firstly, we replace the acronyms with their extensive
full form. We then use this changed text to manually annotate the names of people that the system is supposed to find and apply the anonymization. We made a total of 28 annotations of people names, obtaining the gold collection for the anonymization step.

Next, we moved to the spell checking. We applied our algorithm to the 2817 different unigrams that meet the size requirements. Manually, we identified 339 substitutions to misspelled words as “aidda” which the well-written word is “ainda” (still).

We end this preparation of the evaluation corpora making a gold collection to the last step of this work – Structuring the Clinical Notes. This step was time-consuming to be realized because we had to manually structure 300 clinical notes. As referred, what we want in the final is to obtain the tags associated with the medical terms.

When structuring these evaluation corpora, we obtained the following number of tags, shown on Table 5.1.

<table>
<thead>
<tr>
<th>Entities</th>
<th>Number of Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fármacos</td>
<td>344</td>
</tr>
<tr>
<td>Doenças Reumatológicas</td>
<td>43</td>
</tr>
<tr>
<td>Diagnóstico</td>
<td>635</td>
</tr>
<tr>
<td>Bactéria</td>
<td>6</td>
</tr>
<tr>
<td>Hormona</td>
<td>2</td>
</tr>
<tr>
<td>Hospital</td>
<td>68</td>
</tr>
<tr>
<td>Sintoma/Problema Clínico</td>
<td>867</td>
</tr>
<tr>
<td>Tratamento</td>
<td>89</td>
</tr>
<tr>
<td>Substância Ativa</td>
<td>221</td>
</tr>
<tr>
<td>Marca de Fármaco</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.1: Number of entities on the evaluation corpora

5.2 Results

We compared each of the outputs defined by us in the gold collections previously mentioned, with the output returned by the system. We also compared the results of the Preliminary Evaluation with our Final Evaluation results.

5.2.1 Anonymization Evaluation

As already mentioned, we manually annotate the 28 person names that we found in this corpora. Running this same text in the STRING chain, randomizing the names that this tool is encountering, we obtained a total of 91 anonymized names. Using this information and applying the previously explained formulas, we obtained the next results:
Precision = \frac{28}{(28 + 63)} = 0.31

Recall = \frac{28}{(28 + 0)} = 1

F - measure = 2 * \frac{0.31 * 1}{(0.31 + 1)} = 0.47

The erroneous anonymizations were due to:

• Misspelled words that the system confuses with a personal name. **Example:** The doctor wrote “Urían II” instead of “Urina II”. So the system did not captured the NE “Urina” and considers “Urían” as a proper noun because it starts with a capitalized letter;

• Medical tests and experiences with personal names. **Example:** There is a test diagnose whose name is “ELISA”, which it is also a Portuguese name and the system assumes the exam name as being a person;

• Acronyms that were not replaced for their expansive full form because neither we nor the doctors know the real expansive full form from them. **Example:** “DDCPC”;

• According with the information provided to the system, some acronyms were missing. **Example:** “UII” that means “Urina II”;

• Lack of punctuation that generates an error in the system. **Example:** “Azitromicina Imonugenicidade Junho 2016”, the system gives as output “[**Matilde**] Junho 2016”, because it does not recognize the name “Imonugenicidade”. If we run only the term “Azitromicina”, the system will know that it is a drug name, but, as we insert it on the STRING without punctuation, the system joins all as one sentence and it gives a bad result.

Comparatively with the Preliminary Evaluation, although we increased the number of clinical notes in 100, the results remain the same.

Although precision is not the highest in this step, we are interested in the recall metric. It is more important than the system anonymizes all the person names, but all of them becomes anonymized, instead of having less bad annotations, but fails in names. Fortunately, our system never fails a person name.
5.2.2 Spell Checking Evaluation

As we previously wrote, we only applied our algorithm to the 2817 different unigrams whose size is between 3 and 14 characters (including). After comparing these unigrams with all the unigrams of the Portuguese dataset created by us, we obtained as a result that 2200 unigrams are possibly well-written (appear in the dataset), so, only 617 are misspelled. To these 617 unigrams left over, a list of drugs was compared, that gave us as a result that, after all, is only 596 the misspelled unigrams.

When running the algorithm with a threshold of similarity bigger than 0.8 (that is, 80%), 953 changes of terms were suggested. Of these suggested changes we only choose the one with bigger similarity for each word (best suggestion), obtained a total of 238 suggestions of change by the algorithm. The Example 5.1 shows the output to the word “sistemicas”, giving six best suggestions to replace this word by another word. Of these six best suggestions, we choose the bigger one, in this case, the one with bigger similarity (“sístemicas”) – the best suggestion.

Example 5.1: Best suggestions to the word “sistemicas”

The remaining 358 terms of the misspelled 596 not have any suggestion possibly due to:

• The corrected words do not appear in our Portuguese dataset. Example: “artralgias” (arthralgias) and “gonalgias” are well-written terms, but do not occur as suggestions because they are not present in the Portuguese dataset done by us.

• Some terms are in English and we do not have any Portuguese terms with a similarity bigger than 80% with English terms. Example: “ankylosing” and “spondylitis” are English terms present on the clinical notes.

As we already referred, we identified manually the 339 misspelled terms, indicating the possible well-written word that should arise. Our algorithm run with the same corpora and gives as output 238 misspelled terms, indicating, also, the possible well-written word that should arise in that situation. Of these 238 suggested words by the system, only 105 are in agreement with our manual suggestions. Example: “bilatrais” to “bilaterais” are suggestions that appear in our manual suggestions and as output of the algorithm.
We list some situations that originate this number:

- The system still correctly identified 44 misspelled terms, but do not suggests well the well-written word that should appear (gives us a bad suggestion of correction, although it originates the bigger similarity with the misspelled word). **Example:** ["considerar", 'considerando', 0.8560606060606061], the well-written word should be “considerar” (consider);

- Some suggestions provided by the system are well-suggested and we do not notice that these terms are misspelled, so we do not put them in our manual suggestions. **Example:** ["terapêutica", 'terapêutica', 0.9393939393939394];

- In these 238 suggestions provided by the system, 120 are not totally correct but are cases as, for instance, the misspelled word be in the plural and the suggestion of correction is good, but appear in the singular. **Example:** “urinários” to “urinário”.

Using these data and applying the previously explained formulas and already used in the preliminary evaluation, we obtained the next results:

\[
Precision = \frac{105}{105 + 103} = 0.44
\]

Recall in this case is 1 because the system never classifies a term as correct without this being, that is, it always ranks correctly all the correct terms.

5.2.3 Entities and Structure Evaluation

To do the Structure Final Evaluation, we recurred to the Table 5.1 presented above, that shows us what is the supposed number of each entity that is identified by the system, basing on our gold-collection.

After running on the STRING chain our 300 clinical notes for evaluation and parsing the result to a structured format, was possible to compare the gold-collection with the STRING result, in order to do the following Table 5.2, using an auxiliary Table 5.3.

The Bar Chart 5.2 compares the Preliminary and Final evaluations of the entities, focused on the precision and recall metrics.

The precision always remained within the same values, which is good, because our precision was being quite reasonable.

Relatively to the recall metric, it has some notable descents, principally on the “Diagnóstico” and “Doenças Reumatológicas” entities. Many times this occur due to the lack of punctuation, causing the system to confuse itself. However, with the rules that we create, demonstrated on the Preliminary Evaluation, was still possible to increase the recall metric, relatively to “Sintomas/Problemas Clínicos”, “Tratamentos” and “Substâncias Ativas” entities.
Table 5.2: Precision and Recall metrics applied to the structure

<table>
<thead>
<tr>
<th>Entities</th>
<th>Number of Entities on the Gold-Collection</th>
<th>Number of Entities given by the STRING Chain</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fármacos</td>
<td>344</td>
<td>233</td>
<td>93.1</td>
<td>63.1</td>
</tr>
<tr>
<td>Doenças Reumatológicas</td>
<td>43</td>
<td>29</td>
<td>100</td>
<td>67.4</td>
</tr>
<tr>
<td>Diagnóstico</td>
<td>635</td>
<td>556</td>
<td>88.9</td>
<td>77.8</td>
</tr>
<tr>
<td>Bactéria</td>
<td>6</td>
<td>3</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Hormona</td>
<td>2</td>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Hospital</td>
<td>68</td>
<td>66</td>
<td>97</td>
<td>94.1</td>
</tr>
<tr>
<td>Sintoma/Problema Clínico</td>
<td>867</td>
<td>620</td>
<td>98</td>
<td>70</td>
</tr>
<tr>
<td>Tratamento</td>
<td>89</td>
<td>75</td>
<td>100</td>
<td>84.3</td>
</tr>
<tr>
<td>Substância Ativa</td>
<td>221</td>
<td>155</td>
<td>98.1</td>
<td>68.8</td>
</tr>
<tr>
<td>Marca de Fármaco</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.3: Auxiliary table to help in Precision and Recall results

<table>
<thead>
<tr>
<th>Entities</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fármacos</td>
<td>217</td>
<td>16</td>
<td>127</td>
</tr>
<tr>
<td>Doenças Reumatológicas</td>
<td>29</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Diagnóstico</td>
<td>494</td>
<td>62</td>
<td>141</td>
</tr>
<tr>
<td>Bactéria</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Hormona</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hospital</td>
<td>64</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Sintoma/Problema Clínico</td>
<td>607</td>
<td>13</td>
<td>260</td>
</tr>
<tr>
<td>Tratamento</td>
<td>75</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Substância Ativa</td>
<td>152</td>
<td>3</td>
<td>69</td>
</tr>
<tr>
<td>Marca de Fármaco</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Example 5.2: Chart that shows the preliminary evaluation vs. the final evaluation of the frame.
Conclusions

Contents

6.1 Main Contributions ......................................................... 61
6.2 System Limitations and Future Work ................................. 62
The medical texts that we want to process have information relevant to the prognosis and definition of treatments, as well as, to the health of the patient, but are not currently used and duly analyzed because they require techniques of NLP in the Portuguese language and, more specifically, of medical terms.

This thesis presents a sequence of steps, starting from medical running texts (from the *Reuma.pt* database and obtained in partnership with SPR), until generating structured information from these medical texts. To this, five steps were taken, adopting a hybrid approach: Dealing with Acronyms (Chapter 4.1), Anonymization (Chapter 4.2), Named Entity Recognition (Chapter 4.3), Spell Checking (Chapter 4.4) and Structuring the Clinical Notes (Chapter 4.5).

This chapter overviews the main contributions and highlights possible directions for future work, as well as, listing some limitations of the STRING chain.

### 6.1 Main Contributions

The main contribution of this thesis was to construct a pipeline with five stages that lead us to our goal which consists of obtaining a structured frame from unstructured clinical notes. The stages are the following:

- **Dealing with Acronyms**: It was made a list with 492 entries of attribute-value pairs of acronyms, in which the attribute is the acronym and the value is the extensive full form of that acronym. This list allows us to substitute an acronym by their full form, hence facilitating the next stages. It also helps us to better understand the text of the clinical notes;

- **Anonymization**: Changes were made in the STRING original code to only anonymize our sensitive information – person names. Now, the STRING is able to anonymize, individually, names, persons and organizations;

- **Named Entity Recognition**: The STRING tool was successfully enriched, as previously mentioned, with a wide range of medical terms of determined entities in our domain, specifically: names and drugs brands, symptoms, treatments, hormones, bacteria, diagnoses, hospitals, active substances and rheumatological diseases. In Table 6.1 we showed precisely the number of entries added on STRING, in each list of entities. In addition to the added terms, rules have also been added to relate base terms, thus adapting the STRING chain to a clinical domain. It was added, more precisely, 75 rules to enlarge the medical terms;

- **Spell Checking**: Adopting the idea of *Carvalho and Curto* (2014), it was possible for us to correct misspelled terms of the clinical domain. For that we use as auxiliary, not only, but also some lists made by us in the previous step;
Table 6.1: Number of entities that we add to the STRING

<table>
<thead>
<tr>
<th>Entities</th>
<th>Number of Entries in List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fármacos</td>
<td>1314</td>
</tr>
<tr>
<td>Doenças Reumatológicas</td>
<td>19</td>
</tr>
<tr>
<td>Diagnóstico</td>
<td>104</td>
</tr>
<tr>
<td>Bactéria</td>
<td>178</td>
</tr>
<tr>
<td>Hormona</td>
<td>83</td>
</tr>
<tr>
<td>Hospital</td>
<td>83</td>
</tr>
<tr>
<td>Síntoma/Problema Clínico</td>
<td>1803</td>
</tr>
<tr>
<td>Tratamento</td>
<td>143</td>
</tr>
<tr>
<td>Substância Ativa</td>
<td>717</td>
</tr>
<tr>
<td>Marca de Fármaco</td>
<td>160</td>
</tr>
<tr>
<td>Corpo Humano</td>
<td>174</td>
</tr>
</tbody>
</table>

- Structuring the Clinical Notes: It was made more 244 rules to catch larger expressions that we want to use as value in the frame. Furthermore, we did a parser for the output of STRING chosen by us.

With this insertion of medical terms, targeting to rheumatological texts, and following all the mentioned steps, it was possible to move from a running medical text to a structured medical text. It opens a lot of exploration and analysis hypothesis that we will describe on the next section.

6.2 System Limitations and Future Work

Relatively to the system limitations, as already mentioned before, a bug was detected in STRING, regarding the possible outputs returned by it. All the outputs are being well returned, except the XML output that does not consider rules done before, which hindered our final step (structure the clinical notes) because the rules are essential and could not be put aside.

A great difficulty of this work was the fact that the data is written in Portuguese and the major part of the resources only supports data in English. We know that the fact that we were dealing with very specifics domain and medical language, also contributes to the difficulty of this work.

With regard to future work, we considered:

- Implement on the Reuma.pt database a feature that auto corrects misspelled terms written by the user, using, for instance, an algorithm similar to what done by us;

- To facilitate the data sharing between the health care units and between the hospitals and health centers, we should use a list with all the acronyms present and frequently used in the medical texts, in order to transform them, automatically, in their extensive full form, and, after that, to use a tool capable of omitting the patient names in this field, as well as doctors and other people involved;
• Apply the Name Entity Normalization (NEN) task in the clinical notes. This consists of choosing a specific standard to each one of the medical terms that can be mentioned in two or more different ways. As already referred to in this document, one example is the case in which, in the Portuguese language, we could write “cefaleia” and “dor de cabeça” that both refer to a headache. In order to not have to enter all these terms individually, we could do a link between all these forms and choose one of them to be a standard. This idea was retrieved from a CodaLab competition\(^1\), which theme is very similar to our and which have the NEN task as the main goal;

• Instead of structuring all the clinical notes by chronological order and randomly by patients, it can be structured by patient ID, in order to have access to a structured information from the clinical notes by the evolution of medical appointments by patients, facilitating the analysis of the medical appointments progression and the patient’s own illness;

• With the data in a structured format, we could resort to the application of inference based techniques on this extracted data, more precisely, Data Mining. We could try to apply Data Mining techniques to find patterns/relations between structured information, in a way to extract knowledge, which could auxiliate the doctors in the quality of the diagnose and treatment of patients, finding an association between the extracted pieces of information \(^{[35]}\).

\(^1\text{https://competitions.codalab.org/competitions/19350}\)
Bibliography


A

Some characteristics about the Patients in the Reuma.pt
<table>
<thead>
<tr>
<th></th>
<th>Artrite Reumatoide</th>
<th>Espondilite Anquilosante</th>
<th>Artrite Psoriática</th>
<th>Artrite Idiopática Juvenil</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Número de doentes (total)</strong></td>
<td>3187</td>
<td>732</td>
<td>450</td>
<td>552</td>
</tr>
<tr>
<td>- com biológico</td>
<td>898</td>
<td>404</td>
<td>229</td>
<td>123</td>
</tr>
<tr>
<td>- outras opções terapêuticas</td>
<td>2289</td>
<td>328</td>
<td>221</td>
<td>429</td>
</tr>
<tr>
<td><strong>Idade atual</strong></td>
<td>60,9 ± 14,2</td>
<td>45,0 ± 12,6</td>
<td>53,4 ± 12,9</td>
<td>19,5 ± 11,3</td>
</tr>
<tr>
<td><strong>Idade no diagnóstico</strong></td>
<td>48,1 ± 14,8</td>
<td>32,7 ± 12,3</td>
<td>41,0 ± 13,1</td>
<td>7,7 ± 5,3</td>
</tr>
<tr>
<td><strong>Idade início do biológico</strong></td>
<td>52,4 ± 12,7</td>
<td>40,1 ± 12,2</td>
<td>47,2 ± 11,2</td>
<td>17,6 ± 9,6</td>
</tr>
<tr>
<td><strong>Duração da doença na última consulta (anos)</strong></td>
<td>14,1 ± 10,3</td>
<td>17,6 ± 10,7</td>
<td>15,4 ± 9,7</td>
<td>12,8 ± 10,2</td>
</tr>
<tr>
<td><strong>Duração da doença no início do biológico (anos)</strong></td>
<td>10,8 ± 9,9</td>
<td>13,1 ± 9,9</td>
<td>11,7 ± 9,2</td>
<td>10,2 ± 9,2</td>
</tr>
<tr>
<td><strong>Sexo feminino (%)</strong></td>
<td>82,3%</td>
<td>39,2%</td>
<td>47,4%</td>
<td>68,2%</td>
</tr>
<tr>
<td>- com biológico</td>
<td>86,7%</td>
<td>34,6%</td>
<td>49,0%</td>
<td>73,5%</td>
</tr>
<tr>
<td>- outras opções terapêuticas</td>
<td>80,5%</td>
<td>44,8%</td>
<td>45,7%</td>
<td>67,0%</td>
</tr>
</tbody>
</table>