

**FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO**



**FEUP**

# **Information Extraction From Medication Leaflets**

**Bruno Lage Aguiar**

Master in Informatics and Computing Engineering

Supervisor: Eduarda Mendes Rodrigues (PhD)

11<sup>th</sup> July, 2012



# **Information Extraction From Medication Leaflets**

**Bruno Lage Aguiar**

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

Chair: João António Correia Lopes (PhD)

External Examiner: Maria Benedita Campos Neves Malheiro (PhD)

Supervisor: Eduarda Mendes Rodrigues (PhD)

---

11<sup>th</sup> July, 2012



# Abstract

With the constant growth of medical electronic systems, including decision support systems and personal wellbeing applications, the need for machine understandable information has increased.

However, much of the data currently available is in free-form text, which is a convenient way for people to express concepts and events, but is especially challenging for machines to process. Information extraction can relieve some of the problems related with processing free-form text, by providing a semantic interpretation and abstraction of texts.

This thesis presents the PharmInX information extraction system, which aims to automatically extract information from pharmacological texts, more precisely medication leaflets. The system was designed to target several different kinds of information regarding pharmacological products, particularly their posology, side effects and indications. The primary goal is to provide high-quality and machine understandable information, which is currently not available for medical electronic systems. With such information, these systems could provide better well care services for patients, and enhance decision support systems for health care professionals.

The PharmInX system was designed and developed with these goals in mind. It includes 6 components each with different capabilities: 1) text pre-processing, 2) document reader, 3) general natural language processing, 4) named entity recognition, 5) relation extraction and finally 6) information consumers. The reasoning of these components relies in rules, regular expressions, searches in external resources and machine learning. Once all these stages are completed, we can then access the information extracted through an ontology which was carefully developed to support the pharmacological information that we intended to extract.

For the purpose of both development support and evaluation of the system, some pharmacological documents were manually annotated and used as gold standard. The results achieved by the system resorting to this evaluation indicate that pharmacological and clinical information can successfully be extracted from free-form texts in Portuguese, presenting a F1 score of 99.23% when recognizing entities and a F1 score of 97.43% when extracting relations between those entities.



# Resumo

Com o constante crescimento de sistemas médicos eletrónicos, incluindo sistemas de apoio à decisão e de aplicações de cuidados pessoais, a necessidade de informação compresensível por máquinas tem sido crescente.

No entanto, grande parte dos dados disponíveis atualmente estão em formato de texto livre, que embora seja um formato útil para as pessoas expressarem conceitos e eventos, apresenta desafios para um processamento automático. A extração de informação pode aliviar alguns dos problemas relacionados com o processamento de texto em formato livre, proporcionando uma interpretação semântica e abstracta desses textos.

Esta tese apresenta o sistema de extração de informação PharmInX, que visa extrair automaticamente informações a partir de textos farmacológicos, mais precisamente folhetos de medicamentos. Nós pretendemos abordar vários tipos de informação a respeito dos produtos farmacológicos, particularmente a sua posologia, efeitos adversos e indicações. Com isso, o nosso objetivo é disponibilizar informação de alta qualidade e passível de processamento automático, tendo em conta que este tipo de informação não está actualmente disponível para sistemas médicos eletrónicos. Com essa informação, estes sistemas poderão fornecer melhores serviços de cuidados pessoais para os pacientes, e melhorar os sistemas de apoio à decisão para profissionais de saúde.

O sistema PharmInX foi desenhado e desenvolvido tendo em conta estes objetivos. Este sistema inclui 6 componentes diferentes e com diferentes capacidades: 1) pre-processamento do texto, 2) leitor de documentos, 3) procedimentos gerais de processamento de linguagem natural, 4) reconhecimento de entidades, 5) extração de relações e finalmente 6) consumidores de informação. O processamento destas componentes incluem a regras, expressões regulares, acessos a recursos externos e aprendizagem computacional. Depois de todas estas etapas estarem concluídas, podemos aceder às informações extraídas através de uma ontologia que foi cuidadosamente desenvolvida para suportar a informação farmacológica que pretendemos extrair.

Para efeitos de apoio ao desenvolvimento e também para permitir uma avaliação adequada do sistema, uma quantidade relevante de documentos farmacológicos foram manualmente anotados. Posto isto, fomos capazes de comparar os resultados fornecidos pelo sistema com dados considerados corretos. Os resultados alcançados pelo sistema de recorrendo a esta avaliação indicam que as informações farmacológicas e clínicas podem ser extraídas com êxito a partir de textos livres em Português, apresentando um *F1 score* de 99,23% ao reconhecer as entidades e um *F1 score* de 97,43% ao extrair relações entre essas entidades.





# Acknowledgements

My first words of thanks go to my supervisors, Dr. Liliana Ferreira and Prof. Eduarda Rodrigues, whose guidance was imperative through all the phases of this thesis.

To Fraunhofer AICOS Portugal for the opportunity and for providing the necessary means to complete this research.

With all my heart, to my parents and brother, for the continuous support and encouragement through all these years and to my beloved Diana.

Finally, to my friends for the last amazing few years and the ones to come.

Bruno Lage Aguiar



*“Information is the currency of democracy”*

Thomas Jefferson



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Context . . . . .	1
1.2	Research Objectives and Motivation . . . . .	2
1.3	Contributions . . . . .	4
1.4	Dissemination of Research Results . . . . .	4
1.5	Outline . . . . .	5
<b>2</b>	<b>Background and Related Work</b>	<b>7</b>
2.1	Information Extraction . . . . .	7
2.1.1	Named Entity Recognition . . . . .	8
2.1.2	Coreference Resolution . . . . .	9
2.1.3	Template Element construction . . . . .	9
2.1.4	Template Relation . . . . .	9
2.1.5	Scenario Template . . . . .	9
2.2	Architecture of IE Systems . . . . .	10
2.3	Approaches to Information Extraction . . . . .	11
2.4	Information Extraction in the Portuguese Language . . . . .	12
2.5	Information Extraction in the Pharmacological Domain . . . . .	12
2.5.1	Posology . . . . .	13
2.5.2	Drug Adverse Reactions . . . . .	15
2.5.3	Drug Indications . . . . .	15
2.5.4	Drug-drug Interactions . . . . .	16
2.6	Summary . . . . .	18
<b>3</b>	<b>Resources and Tools</b>	<b>19</b>
3.1	Resources . . . . .	19
3.1.1	UMLS . . . . .	19
3.1.2	Pharmacological Therapeutic Records . . . . .	20
3.2	Processing tools . . . . .	22
3.2.1	UIMA . . . . .	22
3.2.2	Protégé . . . . .	24
3.2.3	Jena . . . . .	24
3.2.4	Knowtator . . . . .	24
3.3	Summary . . . . .	25
<b>4</b>	<b>Knowledge Representation</b>	<b>27</b>
4.1	Ontology Development Method . . . . .	27
4.2	PharmInx Ontology . . . . .	28

## CONTENTS

4.3	Summary . . . . .	33
<b>5</b>	<b>PHARMacological INformation eXtraction system</b>	<b>35</b>
5.1	PharmInx Architecture . . . . .	35
5.2	Document Pre-Processing . . . . .	37
5.3	Document Reader . . . . .	38
5.4	General Natural Language Processing . . . . .	38
5.5	Named Entity Recognition . . . . .	41
5.6	Relation Extraction . . . . .	45
5.7	Information Consumers . . . . .	47
5.8	Summary . . . . .	48
<b>6</b>	<b>Performance and Evaluation</b>	<b>49</b>
6.1	Evaluation Metrics . . . . .	49
6.2	Evaluation Setup . . . . .	50
6.2.1	Annotation Process . . . . .	51
6.2.2	Test Set Characterization . . . . .	52
6.3	Evaluation Results . . . . .	54
6.3.1	Identified Mismatches . . . . .	55
6.3.2	Precision, Recall and F-measure . . . . .	55
6.3.3	Results for each Category . . . . .	56
6.3.4	Results for each Relation . . . . .	59
6.4	Discussion . . . . .	59
6.5	Summary . . . . .	60
<b>7</b>	<b>Conclusions</b>	<b>61</b>
7.1	Future Work . . . . .	62
<b>A</b>	<b>Submitted Papers</b>	<b>63</b>
A.1	StudECE . . . . .	63
	<b>References</b>	<b>67</b>

# List of Figures

1.1	Specific information regarding a pharmacological product . . . . .	3
2.1	Common IE architecture (adapted from Ronen Feldman and James Sanger (2006) [FS06]) . . . . .	10
3.1	UIMA System Architecture (original figure from Liliana Ferreira (2011) [dSF11])	23
4.1	Tree of entities, types and subtypes considered in the knowledge representation of the information regarding pharmacological products . . . . .	28
4.2	Partial view of the knowledge representation model, focusing in the classes <i>Posology</i> , <i>Dosage</i> , <i>Time</i> , <i>AdminRoute</i> and <i>PersonClass</i> . . . . .	31
4.3	Partial view of the knowledge representation model, focusing in the classes <i>Posology</i> , <i>Drug</i> and <i>UMLS</i> . . . . .	32
4.4	Partial view of the knowledge representation model, focusing in the classes <i>Posology</i> , <i>Restriction</i> and <i>Characteristic</i> . . . . .	33
4.5	PharmInx ontology in Protégé . . . . .	34
5.1	System Architecture . . . . .	36
5.2	Types of annotations present in PharmInx type system . . . . .	37
5.3	Sample of the generated XML with the pharmacological therapeutic records content	38
5.4	Flowchart for NER . . . . .	39
5.5	Example of sentence and token annotations . . . . .	40
5.6	Flowchart for NER . . . . .	41
5.7	General work-flow chart of the UMLS annotator reasoning . . . . .	44
5.8	Example of UMLS leaf and composite annotations . . . . .	45
5.9	Example of a <i>Characteristic</i> annotation . . . . .	45
5.10	An example of a posology after the NER task . . . . .	46
5.11	Tree inferred from example in Figure 5.10 . . . . .	47
5.12	Populated PharmInx ontology . . . . .	48
6.1	Knowtator layout when annotating the Test Set . . . . .	51
6.2	Distribution of the entities found by PharmInx according to their categories . . .	53
6.3	Distribution of the relations found by PharmInx according to their relations . . .	54
6.4	Precision, Recall and F1-Score of the system for the NER task . . . . .	57
6.5	Precision, Recall and F1-Score of the system for the RE task . . . . .	57
6.6	F1-Score for each of the categories present in PharmInx . . . . .	58
6.7	F1-Score for each of the relations present in PharmInx . . . . .	59

## LIST OF FIGURES



# List of Tables

2.1	Results presented by Gold <i>et al</i> [GEZ <sup>+</sup> ]	14
2.2	Results presented by Xu <i>et al</i> [XSD <sup>+</sup> 10]	15
3.1	UMLS semantic types and PharmInx semantic groups	21
3.2	Statistics regarding the content of the information topics in the document where the extraction process is held	22
4.1	Classes of the PharmInx Knowledge Representation	29
4.2	Relations of the PharmInx Knowledge Representation	30
5.1	Short sample of the abbreviations used in the corpus	39
6.1	Number of documents, characters, tokens and sentences in the PharmInx corpus and its subsets	50
6.2	Detailed statistics regarding the amount of entities extracted from the <i>Test Set</i>	52
6.3	Detailed statistics regarding the amount of relations extracted from the <i>Test Set</i>	54
6.4	Detailed statistics regarding the errors detected when comparing the entities extracted with the manually annotated entities	55
6.5	Detailed statistics regarding the amount of relations extracted from the <i>Test Set</i>	56
6.6	Overall results for the NER and the RE tasks	56
6.7	Detailed statistics regarding the errors detected in the Posology category	58

## LIST OF TABLES

# Abbreviations

AICOS	Assistive Information and Communication Solutions
ADE	Adverse Drug Event
CAS	Common Analysis Structure
CO	Coreference Resolution
DeCS	Descritores em Ciências da Saúde
DDI	Drug-Drug interactions
GATE	General Architecture for Text Engineering
HMM	Hidden Markov model
ICD	International Classification of Diseases
ICD-9	International Classification of Diseases, Ninth Revision
ICD-10	International Classification of Diseases, Tenth Revision
IE	Information Extraction
ICF	International Classification of Functioning, Disability and Health
IR	Information Retrieval
MedLEE	Medical Language Extraction and Encoding system
MeSH	Medical Subject Headings
MUC	Message Understanding Conferences
NER	Named Entity Recognition
NLP	Natural Language Processing
OWL	Web Ontology Language
PDF	Portable Document Format
POS	Part-Of-Speech
RDF	Resource Description Framework
RLS	Regularized Least-Squares
SPARQL	SPARQL Protocol and RDF Query Language
ST	Scenario Template
SVM	Support Vector Machines
SW	Semantic Web
TE	Template Element
TR	Template Relation
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System
USNLM	United States National Library of Medicine
W3C	World Wide Web Consortium
WHO	World Health Organization
XML	Extensible Markup Language



# Chapter 1

## Introduction

This thesis aims to develop an information extraction system capable of extracting specific information from the Portuguese medical therapeutic records [IM]. These records are available through Infarmed, the National Authority of Medicines and Health Products and the Portuguese Health Ministry, and they contain many different kinds of information about all the pharmacological drugs currently in use in Portugal. All this information is currently expressed in natural language text, more specifically in Portuguese, so in order to provide concrete and valuable information for computer applications we need to be able to structure the information desired. The main goal of this thesis is to design and develop an information extraction system capable of outputting structured information about pharmacological drugs currently in use in Portugal, relying on the natural language records available.

This chapter introduces the context in which this project fits. It presents the scope of the project, the motivation for this thesis, as well as the objectives of our thesis work. It also summarizes the main contributions of our work and the dissemination of our results. Finally, it outlines the structure of the document.

### 1.1 Context

This thesis, "Information Extraction from Medication Leaflets", was held as a master dissertation project under the Master in Computer Engineering from Faculty of Engineering, University of Porto. This thesis was proposed by Fraunhofer Assistive Information and Communication Solutions (AICOS) Portugal, where the development phase was also held.

One of the current applications areas of Fraunhofer AICOS Portugal is denominated "Care, Well-Being and Inclusion" and some of its main goals include helping people with chronic (and other) conditions to live more independent lives, increasing the responsiveness and efficiency of the health care services and contributing to the creation of a competitive health care industry in Europe. Considering these goals, the need for accessing pharmacological information naturally

emerged. In the current context, the pharmacological information needed emphasis more in the information that is important for patients. Nevertheless, we bear in mind the potential usefulness of this information for other domains and individuals. Therefore, since this information has been considered vital for further developments and with great potential, the thesis was proposed.

## 1.2 Research Objectives and Motivation

As stated above, Fraunhofer AICOS Portugal has been investing in applications with an health-care character. Therefore, some of these applications presented the need for specific medical and pharmacological information. Since information regarding pharmacological products is currently only available in a natural language format, a system capable of transforming this machine readable information into machine understandable became a major necessity. The aim of this thesis is to fulfill this need.

As we can see in Figure 1.1, which represents a fragment of the information available for a specific pharmacological product in the documents where the information extraction tasks are to be held, there are several different sorts of information that characterize a pharmacological product. The main components of information to be extracted are here described:

- **Posology:** Describes how the pharmacological product should be administrated, including therefore the advised dosage, administration route, frequency and duration of administration, for which disorders is recommended and even some restrictions regarding the time of administration (e.g. before meals) or patient characteristics (e.g. weight, age);
- **Indications:** Aims to indicate the drugs target and its common uses;
- **Adverse Reactions:** Describes the harm associated with the use of a given medication at a normal dosage during normal use;
- **Contraindications and cautions:** Some warnings regarding possible dangers when administering the product in patients with certain conditions or with other medications prescribed;
- **Interactions:** Information regarding interactions between this product and other pharmacological products. This information is commonly referred in the literature as drug-drug interactions (DDI).

All the existing information about pharmacological products can be very useful for health-care applications, being helpful not only for patients, to know more about the medication they are taking, but also for health-care professionals, through decision support system.

Regarding the patients, we must emphasize the information regarding posology, since this information is important to remind the patient about the administration methods and more importantly about dosages. Nevertheless, information regarding adverse reactions can also be helpful as a remainder of the reactions to be expected, preventing in some cases new doctor visits from the worried patients. Indications are also of great importance for patients, indicating the common

■ BENZILPENICILINA POTÁSSICA

*Ind.:* Infecções por agentes penicilino-sensíveis, nomeadamente faringite, amigdalite, otite média, pneumonia, endocardite estreptocócica e meningite meningocócica ou pneumocócica.

*R. Adv.:* Reacções de hipersensibilidade incluindo febre, urticária, dores articulares; angioedema. Leucopenia e trombocitopenia, usualmente transitórias. Choque anafilático apenas em doentes com hipersensibilidade às penicilinas.

*Contra-Ind. e Prec.:* História de hipersensibilidade às penicilinas. Reduzir a posologia no doente com IR.

*Interac.:* A probenecida inibe competitivamente a secreção tubular das penicilinas causando um aumento significativo das suas concentrações séricas.

*Posol.:* [Adultos] - Via IV (perfusão intermitente) ou IM: 300.000 a 1.200.000 UI/dia, a administrar de 3 em 3 ou de 4 em 4 horas.

Via IV (perfusão intermitente ou perfusão contínua): 10.000.000 a 24.000.000 UI/dia, a administrar a intervalos de 2 em 2 horas ou por perfusão contínua, no tratamento de infecções graves.

Reduzir a posologia no doente com IR (Cl cr < 50 ml/min).

Via intratecal - Não recomendada.

[Crianças] - Via IV: < 12 anos: 25.000 a 400.000 UI/kg/dia, a administrar de 4 em 4 ou de 6 em 6 horas.

Figure 1.1: Specific information regarding a pharmacological product

uses of the pharmacological drug. Since this thesis aims initially to fulfill the patients information needs, these three components will be the main and initial goals.

For the purpose of improving decision support systems for health-care professionals, many of this pharmacological information can be used. Analyzing some statistics regarding medication errors in hospitals we can better understand the importance of this information. For instance, a study [SDM<sup>+</sup>] states that diverse adverse drug events (ADE) cause more than 770 000 injuries and deaths each year and cost up to \$5,6 million per hospital. It additionally states that patients who suffered unintended drug events remained in the hospital an average of 8 to 12 days longer than patients who did not experience such mistakes. These added days led to increased stay cost for the hospital from \$16 000 to \$24 000. Other studies, as [KCD00] state that medication errors kill 7000 per annum only in USA. We should also consider that 5% of this medication errors are due to DDI [LBC<sup>+</sup>95]. Therefore, with information regarding pharmacological product we could aim to lower medication errors, reminding and making available to health-care professionals information regarding indications, adverse reactions, contraindications and even DDI.

As stated above, we aim mainly to fulfill the information needs of patients, being that the posology, indications and adverse effects were identified as the most useful topics for such a matter. However, indications and adverse effects are usually a list a disorders, i.e. disorders for which

the product is recommended and disorders that can be triggered by the product, being that the posology is a much more useful and complete source of information. Thus, our system was primarily designed for extracting posology information. By targeting first the posology we are aiming for more important and useful information and we can also assure that some of the components developed can be properly reused for the extraction of indications and adverse reactions.

### 1.3 Contributions

This thesis presents an information extraction system able to extract information from medication leaflets written in Portuguese, the PharmInX system. This system allows, not only the automatic extraction of information from pharmacological texts, but also the automatic instantiation of a knowledge representation model for the medication leaflets. As far as we know, this system is the first information extraction system for the automatic processing of Portuguese medication leaflets.

The implementation of PharmInX lead to some other innovative contributions that can be summarized as follow:

1. Knowledge resources in Portuguese:

- development of a new knowledge representation model for the pharmacological domain.

2. Language processing applications:

- development of a named entity recognizer for Portuguese pharmacological texts, identifying dosages, frequencies, durations, age groups, restrictions and administration routes;
- development of an annotator of clinical procedures, disorders and anatomical sites on Portuguese texts;
- development of a relation extractor system for the pharmacological domain.

3. Evaluation:

- development of a test set for the posology topic of pharmacological products;
- development of an annotation schema for the pharmacological domain.

### 1.4 Dissemination of Research Results

In order to disseminate our scientific contributions, we submitted a paper to a national conference and are currently preparing another paper for submission to an international conference.

The first paper, a short summary of the work developed and presentation of the main results, was accepted for the StudECE 2012<sup>1</sup> proceedings and is attached to this thesis in appendix A.

---

<sup>1</sup>More details at <http://paginas.fe.up.pt/StudECE2012/>



The second paper, still under development, is aiming for the Second International Workshop on Managing Interoperability and compleXity in Health Systems (MIX-HS'12)<sup>2</sup>, which is held in conjunction with the 21st ACM International Conference on Information and Knowledge Management (CIKM)<sup>3</sup>.

Besides the scientific dissemination, possible approaches for the market could be interesting. Considering the nature of the results of this thesis, the easier way to reach the market would be by integrating these results in health-care systems currently under development at Fraunhofer AICOS Portugal. The integration with eCAALYX (Enhanced Complete Ambient Assisted Living Experiment)<sup>4</sup> and S4S (Smartphones for Senior)<sup>5</sup> have been seriously discussed and should happen in a near future.

## 1.5 Outline

This thesis is organized in three main parts, with the following structure:

**Part 1: State of the Art.** This first part contains a literature review on the background of information extraction, on related works and an analysis in the main resources that were useful for the development of this thesis. It is composed by two chapters:

- Chapter 2, *Background and Related Work*, where a revision on the background and related works is performed, emphasizing information extraction systems in general and then their application for the Portuguese language and in the pharmacological domain.
- Chapter 3, *Resources and Tools*, identifies and provides a brief description of the resources that were identified as useful for the development of this thesis.

**Part 2: Implementation.** The second part of this thesis describes our approach to the project, including how we managed to represent the knowledge to be extracted and also including the system technical details.

- Chapter 4, *Knowledge Representation*, describes the adaptive model used for representing the knowledge we aim to extract, in order to be later available for other applications.
- Chapter 5, *PHARMacological INformation eXtraction system*, describes the system technical design and architecture, describing its several components.

**Part 3: Results and Conclusions.** The last part of this thesis describes the evaluation procedures held, as well as the results obtained. We finish with some conclusions and future work.

- Chapter 6, *Performance and Evaluation*, presents our system evaluation, including the evaluation metrics, setup and finally the results. A brief discussion over the results is also held.

---

<sup>2</sup>More details at <http://informatics.mayo.edu/CNTRO/index.php/Events/MIXHS12>

<sup>3</sup>More details at <http://www.cikm2012.org/>

<sup>4</sup>More details at [http://www.fraunhofer.pt/en/fraunhofer\\_aicos/projects/government\\_contractresearch/ecaalyx.html](http://www.fraunhofer.pt/en/fraunhofer_aicos/projects/government_contractresearch/ecaalyx.html)

<sup>5</sup>More details at [http://www.fraunhofer.pt/en/fraunhofer\\_aicos/projects/industry\\_contractresearch/s4s.html](http://www.fraunhofer.pt/en/fraunhofer_aicos/projects/industry_contractresearch/s4s.html)

## Introduction

- Chapter 7, *Conclusions*, is the last of this thesis and summarizes our work by reviewing the proposed objectives of this thesis and their achievement. The considered future work is also described.

## Chapter 2

# Background and Related Work

During the initial phase of this master thesis a deep analysis on the current state of the art regarding Information Extraction (IE) was performed. Therefore, a thorough analysis was held on the background knowledge needed to correctly understand information extraction and also on the information extraction systems from where we could withdraw any kind of synergies.

We start in Section 2.1 by providing some background knowledge about information extraction, providing a brief definition and identifying the principal inherent tasks. Later, on Section 2.2 a common architecture for information extraction systems is presented. Section 2.3 presents some of the most common approaches to the IE task. Section 2.4 presents a resumed state of the art regarding information extraction for the Portuguese language, with more emphasis in clinical domains which are more similar with the domain in study in this thesis. Section 2.5 intends to present some important works that constitute the state of the art on information extraction from pharmacological domains, including the extraction of posologies, adverse reactions, indications and DDI. Finally, in Section 2.6 a brief summary of this chapter is presented.

### 2.1 Information Extraction

Information extraction is a sub-area of Natural Language Processing (NLP), which goal is to extract information from natural language text, and the aim is to be able to do it without requiring the user to analyze the text. It contrasts with Information Retrieval (IR) which focuses on retrieving documents, whereas IE focuses on retrieving information or facts. Some examples of IR systems are search engines, using these systems the user must read the documents retrieved to access the information or facts contained in those documents. However, using IE we are able to directly tabulate those facts and information [AM05].

IE is considered the most prominent technique currently used for text-mining pre-processing [FS06], since it allows the system to transform a machine readable document into a machine understandable document.

## Background and Related Work

According to [AM05], IE can be divided in two phases. In the first one, the system should be able to take natural language text and extract some facts, using a model to guide this process. The system will attempt to retrieve information according to this model, ignoring other types of data. Secondly, it should be able to represent the facts previously extracted in a predefined template, filling the slots of this template according to the content extracted.

Using IE there are four basic types of elements that can be extracted from the natural language text:

- **Entities:** These are the basic building blocks that can be found, representing the subjects we want to represent, as for example drugs, people, companies and locations.
- **Attributes:** These are features of the previously extracted entities, as titles of a person, type of a company and so on.
- **Facts:** These facts are the existing relations between the entities, as for example incompatibility relationships between drugs.
- **Events:** These represent activities or occurrences in which the entities participate, such as birthday parties or a merging between two companies. This type of elements is not a priority for this project, since the information to be extracted does not contain any newsworthy events.

Some of the most comprehensive work for information extraction has arisen with the promotion of the Message Understanding Conferences (MUC), specially from MUC-6 [GS96] and MUC-7 [Chi98] conferences. In these conferences, information extraction has been introduced as the aggregate of five tasks, which are below described.

### 2.1.1 Named Entity Recognition

Named Entity Recognition (NER) is the basic task-oriented phase of any IE system. During this phase the system tries to identify in the text all the proper names, such as person and companies names, and sometimes even special types of entities, such as dates, times or other values. Generally this entities are classified into particular categories.

NER recognition can be performed at up to around 95% accuracy, and considering that human annotators do not perform 100% level, as measured in MUC by inter-annotator comparisons, we can say that NER recognition function at human performance levels [Cun05].

In recent years, NER in the biomedical scientific literature has become the focus of many research. NER in clinical notes feature some further challenges, as the idiosyncratic abbreviated symbolic expressions, the ambiguous names, the synonyms and, finally, the variations of the names often found in the text.

In general, the current approaches in biomedical name recognition can be roughly divided into the following four groups [AM05]:

- **Dictionary-Based approaches:** Try to find names of the well-known nomenclatures in the texts;
- **Rule-Based approaches:** Manually or automatically construct patterns or rules to directly match them to candidate NE in the literature
- **Machine Learning approaches:** Develop statistical models for name recognition, employing machine learning techniques, such as Hidden Markov Models (HMM) and Support Vector Machines (SVM).
- **Hybrid approaches:** try to deal with the different aspects of NER by merging two or more of the previous approaches generally in a sequential way.

### 2.1.2 Coreference Resolution

Coreference Resolutions (CO), also known as anaphora resolution, is the process of matching expressions that refer to the same entity in the real world. The most common type of coreference is the pronominal anaphora, which deals with pronouns such as he, she and they. This task is critical to the proper function of advanced text mining pre-processing systems [FS06].

Resolving some kinds of coreference is usually a difficult task, thus this is not an high-performance task for IE. Even humans only achieve about 80% [FS06].

### 2.1.3 Template Element construction

The Template Element (TE) task builds on NE recognition and CO resolution, extracting identifying and descriptive attributes of the entities identified in NE. TE tasks still are open to some improvement, since good scores for these tasks are around 80%, whereas humans can achieve results around 90% [Cun05].

### 2.1.4 Template Relation

The Template Relationship task (TR) aims to find domain-independent relationships between entities, as compared with TE that just identify entities themselves. The goal of the TR task is to find the relationships that exist between the template elements previously extracted during the TE task. In general, good TR systems are able to score around 75% [Cun05].

### 2.1.5 Scenario Template

Scenario Templates (ST) try to express domain and task-specific entities and relations. ST tie together TE entities and TR relations into more complex event descriptions. This is a difficult task in IE, being that the best MUC systems scored around 60%. Even the human score for this task can be considered low, reaching as low as 80%, which illustrates the complexity involved in this task [Cun05].

## 2.2 Architecture of IE Systems

The general architecture of IE systems is based on a pipeline processing where a set of modules is executed in a given order. An output of one module is an input for another. The order in which the subsequent modules are executed is essential because some modules provide or require information that is required or provided by other modules. Depending on the task specification and the language characteristic different types of modules are plugged into the pipeline. However, a generalized architecture for information extraction systems is shown in Figure 2.1, noting that the sub-components are colored according to their priority within the full system.

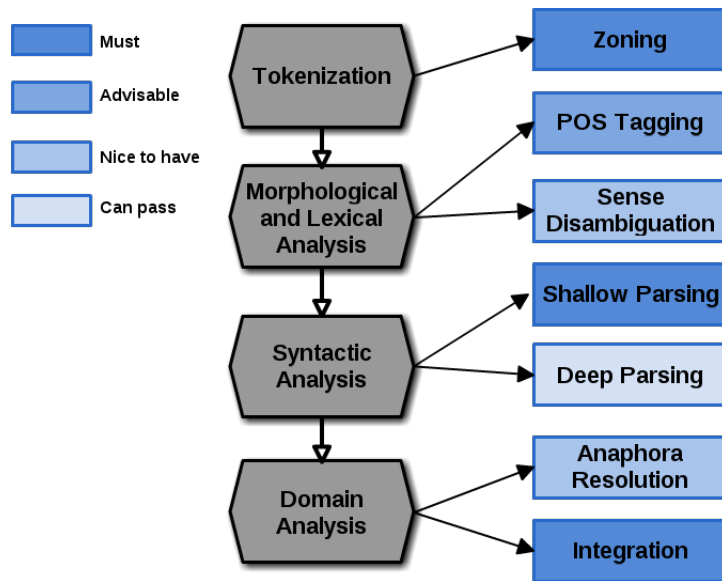


Figure 2.1: Common IE architecture (adapted from Ronen Feldman and James Sanger (2006) [FS06])

Typically, IE systems have three or four major components. The first component is a tokenization or zoning module, which splits the input text into basic building blocks. These building blocks are normally words, sentences and paragraphs. In some special cases, we may have some higher building blocks like sections or even chapters.

The second component aims to perform morphological and lexical analysis. In this module activities such as part of speech (POS) tagging are executed. The various words in the document are tagged with a value, based on definition and context of the word. Disambiguation of both words and sentences is also performed in this module.

The third component is responsible for syntactic analysis. This part of an IE system establishes the connection between the different parts of each sentence. This is done either by doing full parsing or shallow parsing. However, shallow parsing is usually sufficient for IE tasks and therefore preferable to full analysis because of its far greater speed and robustness [FS06].

A fourth and increasingly more common component in IE systems performs domain analysis, which is a module in which the system combines all the information collected from the previous

components and creates complete facts that describe relationships between entities. Advanced domain analysis modules include anaphora resolution components.

### 2.3 Approaches to Information Extraction

One of the most common approach to IE in the last years is based on pattern-matching, which exploits basic patterns over a variety of structures, using regular expressions. These patterns are generally searched in the text, however slightly deeper analysis can be made, as for example, by first tagging text for part-of-speech and then apply the regular expressions over these tags [AM05] [OHTT01]. Despite its simplicity, experiments have shown that this technique produces surprisingly good results. The regular expressions are usually identified manually, requiring a person or a group of persons to read most of the documents of the corpus and then infer suitable extraction rules. This is obviously a tedious and time consuming task, which together with the difficulty of its extension to other domains form the majors drawbacks of this approach. In order to deal with these issues, some systems implement machine learning algorithms.

Therefore, the need for more adaptive IE systems lead to the development of different approaches [ABD<sup>+</sup>95] [CDIW03], mainly based on supervised machine learning. Thus, the effort required to redesign and implement a new system was replaced by the effort of creating a new batch of training data and applying learning algorithms [Tab10]. This effort comprises mainly manual annotation of text, which usually requires less skilled personnel and therefore brings down the cost of re-targeting the system for another domain. Holding with the pattern-matching approach, several systems implement a supervised learning of extraction patterns and rules, being therefore able to overcome some of the main drawbacks of this approach [DSM00]. However, there are other approaches, based on supervised learning of sequential classifier models, that consider IE as a classification problem that can be tackled using sequential learning models [DSM00]. Some investment has also been done regarding weakly or even unsupervised approaches, being therefore capable of skipping the effort of the creation of new batches of training data. However, these approaches still need further investment and study [DSM00].

More recently an interesting approach has emerged, ontology-driven information extraction, which aims at using ontology to guide and constrain text processing [HRS] [AM05]. In this approach, ontologies play a crucial role as they provide formal and explicit specifications of conceptualizations [WD10]. Since Berners-Lee [BLCL<sup>+</sup>94] [BLF99] proposed and started to endorse ontologies as the backbone of the Semantic Web (SW), a whole research field evolved around the fundamental engineering aspects of ontologies such as the generation, evaluation and management of ontologies. Therefore, several IE groups have focused on developing extraction methods that use ontologies content and semantics to guide and constrain the extraction task [AM05]. An interesting and useful side effect of this approach, in the case of pattern-matching, would be the significant decrease of linguistic rules needed for the extraction task [AM05].

## 2.4 Information Extraction in the Portuguese Language

The review of the literature on this topic for Portuguese written texts revealed a severe lack of studies developed to process the Portuguese language, in any of its different variants. Nevertheless, the number of studies in this matter has been growing in the last years. In this section a brief summary of the most relevant studies for the Portuguese language will be presented, although emphasizes will be made in studies related with clinical domains, due to the similarities with the domain of this thesis.

In European Portuguese, for the best of my knowledge the most relevant study found is a work that aims to automatically extract information from free-text discharge summaries [dSF11]. This work used Unstructured Information Management Architecture (UIMA) [FL04] framework to develop both rule-based and machine-learning based components. This system used an ontology-driven approach to the task of IE from Portuguese clinical narratives and, from the use of ontologies, it is able to convert a complete report into a relational information model and to increase the expressive power of its extraction rules. The evaluation of the system, based on 50 reports, validated by 7 different reviewers, presented an admirable F-measure of 95% in the task of identifying and classifying clinical entities.

Some other relevant studies were found such as a work on the development of ontologies to model neurovascular reports in Portuguese [BQ10]. In this work, an ontology to represent the information of the reports was developed and a set of NLP based tools were used for its automatic population. Also available is an experiment developed for the automatic population of a vaccination ontology [TC10]. This study described the methods used to model and populate the vaccination ontology and also a system that identifies relevant vaccination information in medical texts, particularly, in the Portuguese National Vaccination Programme. The work also focuses on clinical text mining, as it describes a first approach to extract information regarding inter-class relations using Association Rule Mining [AS94].

In the Brazilian variant of Portuguese, the most relevant study found was developed in University of São Paulo, Brazil, and intended to apply a medical language processor, MedLEE, on a set of Portuguese medical text and X-ray reports [CFM07]. However, since MedLee was developed for the English language, the Portuguese medical text set was machine translated to English. These results show that successful application of systems developed initially for one language to another is not straightforward.

## 2.5 Information Extraction in the Pharmacological Domain

Medication information is considered to be one of the most important types of clinical data in electronic medical records, being critical for health-care safety and quality, as well as for clinical research based on electronic medical record data. However, medication data are commonly recorded in a free-text format, being therefore not accessible to computer applications that rely on coded data.



As a consequence of the above, in the last few years many authors have been focused on developing systems capable of producing machine understandable data from the clinical narratives regarding medication information. This section aims to enumerate and describe some of the most relevant approaches found for this matter.

### 2.5.1 Posology

Information regarding drugs posology or dosage has been a matter of interest for quite some years by now. This information has been considered crucial not only to assist patients and health-care professionals, but also to research purposes. In this section we will summarize some of the more important approaches to this matter.

One early and important study regarding this matter dates back from 1996 [EBHC96], where the development and evaluation of an automated procedure for extracting drug-dosage information from clinical narratives was performed. In this work Evans *et al* designed a model for the development of this kind of studies, defining 5 essential steps:

1. Establishing a model or definition of the concept to be extracted (viz., drug dosage);
2. Preparing data (medical text) for training and testing;
3. Creating an IE module, including preparing new resources;
4. Processing the data to extract dosage references;
5. Scoring results.

When defining the concept, Evans *et al* defined an object as a drug or pharmacological substance, and its allowable attributes as any of the sub-expressions that play some role (dose-level, frequency, necessity, purpose, route or duration). For the extraction process, a set of about 50 rules were defined. These rules were designed to match textual expressions of the object and attributes in the drug-dosage concept definition.

A relevant study in this matter performed an attempt to extract medication information from discharge summaries using a set of regular expressions as parsing rules and a user-configurable drug lexicon [GEZ<sup>+</sup>]. These authors based their work on Evans *et al* study [EBHC96] referred above, considering their study a proof of concept and using Evans *et al* study as a model to follow. Therefore, following [EBHC96] methodology, but with some modifications of their own, they resumed their work in the following 6 steps:

1. Defined the concepts to be extracted;
2. Built the parser and parsing rules;
3. Prepared data for testing;
4. Had two physicians annotate the test data to create a gold standard;

## Background and Related Work

5. Processed test data with the parser;
6. scored the results.

Their parser received as parameters a set of parsing rules formatted as regular expressions and a lexicon of drug names, which in their experiments was a lexicon derived from RxNorm entries in the Unified Medical Language System (UMLS) [oM08]. The parser had the ability to extract information regarding the context of the drug reference, if mentioned as an allergy, if it was discontinued or prescribed and taken. Information regarding where the drug was administered, at home or in the emergency room, was also extracted and denominated context. However, the most important feature of this work in the context of this project is the ability to extract the drug name, dose, route, frequency and necessity. In order to evaluate their study, two physicians created the gold standard by sequentially annotating the test data. Their evaluation showed an average precision of 94% and 83% for recall in the extraction of medication information task. However, as we can see in Table 2.1, these results would be higher if the extraction of the context of administration was not intended, as in our case. Therefore, analyzing these results we can be optimistic about the effectiveness of parsing rules for the extraction of drugs dosage information.

Table 2.1: Results presented by Gold *et al* [GEZ<sup>+</sup>]

	Context	Dose	Route	Frequency	Necessity
<b>Correct</b>	65.9%	83.7%	88.0%	83.2%	98.6%
<b>Partially Correct</b>	N/A	5.8%	0.0%	1.4%	1.0%
<b>Wrong</b>	34.1%	10.6%	12.0%	15.4%	0.5%

Another study aimed to tackle the limitations of regular expression based approach when processing complicated medication text that contains multiple signatures and contextual level information such as status or temporal data. In this study, denominated MedEx [XSD<sup>+</sup>10], Xu *et al* used a new method to parse medication text based on a sequential tagger and a combined parser. The sequential tagger developed, which combined lookup, regular expression, and rule-based disambiguation components, is said to provide a more robust tagging method, which caused great improvements in the accuracy of semantic labeling of drug names and signatures. The developed parser, combining a Chart parser and a regular expression Chunker, has also improved the ability to parsing more complicated medication text. Xu *et al* evaluation showed that MedEx has presented excellent results so far in the information extraction task for this scope with high F-measures (93.2%-96.0%). More detailed information about the results, which take in account 25 clinical notes, is presented in Table 2.2.

In this section we were able to identify some important studies from where many synergies can be taken. We can see that the studies above described presented excellent results using pattern-matching approaches, showing that this kind of approaches performs well for the extraction of information regarding posology of drugs.

Table 2.2: Results presented by Xu *et al* [XSD<sup>+</sup>10]

Finding Types	Total #	Precision (%)	Recall (%)	F-measure(%)
Drug Name	200	97%	88%	92%
Strenght	94	95%	95%	95%
Route	54	96%	87%	91%
Frequency	102	97%	89%	93%

### 2.5.2 Drug Adverse Reactions

Drugs adverse reactions, commonly known as ADR, are officially identified in the leaflets of drugs, being an important information for the patients. Having this information available, these patients are able to foresee what effects they will probably suffer from the medication, being therefore able to take the proper precautions or simply not being caught completely unaware.

Aramaki *et al* presented a study that aimed at extracting adverse drug effects from clinical records [AMT<sup>+</sup>10]. Although the clinical records used were written in Japanese, the authors claim that their approach is language independent and could therefore be applied to any other language. When approaching the problem, Aramki *et al* identified two separate tasks, the term identification, which aimed for identifying symptoms and drugs, and the relation identification task. For the first one, a typical NER task, they used a state-of-the art method, conditional random fields (CRF). For the second task, they implemented both a pattern-based method and a machine-learning (SVM) method. Regarding their results, for the term identification task an average 83.4% F-measure was achieved. The relation extraction task results presented an average F-measure of 65.0% for the pattern-based approach and of 59.8% for the machine learning approach, which used SVM. The authors claimed that the machine learning approach could possibly present better results if a bigger training data were available.

In the end of this section, we can conclude that both patter-matching and machine-learning approaches presented similar results, existing only a slight difference on advantage for patter-matching approaches. Although there are not many other studies specific for drug side effects extraction, many studies more focused on other topics, like posology or DDI, claim that the same methodologies could be adopted to drug side effects extraction, like for instance the study presented by Rubrichi *et al* [RSGQ10].

### 2.5.3 Drug Indications

Drug indication refers to what disease(s) a drug may treat, and is therefore an information frequently sought by biomedical researchers, health-care professionals and the general public.

Névél *et al* developed a study [NL10] which focused on automatically extracting and integrating "TREAT" relationships between drugs and diseases. To extract drug/disease relationships from biomedical text, they relied on the semantic representation program SemRep [RF03], which is a tool under development at the National Library of Medicine. This tool extracts relationships between biomedical entities based on linguistic analysis of text and domain knowledge in the

Unified Medical Language System (UMLS). Since anaphora and ellipsis are frequently used in biomedical corpus, they developed a pre-processing method for ellipsis, which consisted in automatically restoring drug names when these were absent. To deal with anaphora they developed a post-processing inference method, consisting in a simple heuristic to infer more specific relationships. The results presented that 7670 unique "TREATS" relationships were extracted between 4666 drugs and 1293 diseases, with an estimated overall correctness of 77% and specificity of 84%.

Another work [ZLL03], by Zhu *et al*, presented a system capable of automatically extracting disease-chemical relationships from the UMLS Co-occurrence table. These disease-chemical relationships could be subdivided for diagnostic use, which represents the target of the drug, and for adverse drug effects. Among the other kind of relationships that this work intended to extract, which are not relevant for our scope, this relationships presented best results, with a 93% sensitivity.

Chen *et al* presented a study [CHX<sup>+</sup>08] that aimed to apply automated NLP and statistical techniques for identifying disease-drug associations within biomedical literature and clinical narratives. Two different NLP systems, BioMedLEE and MedLEE, were applied to Medline articles and discharge summaries, respectively, where disease and drug entities were identified using the NLP systems and MeSH annotations for the Medline articles. Later, co-occurrence statistics were applied to compute and evaluate the strength of the associations between drugs and diseases.

Some relevant studies were analyzed in this section, showing that in average good results are achieved in this task.

### 2.5.4 Drug-drug Interactions

Drug drug interactions (DDI) occur when one drug influences the level of activity of another. These interactions are not always necessarily prejudicial, although negative DDI can be extremely dangerous and are therefore an important field of research. This research is crucial for both patient safety and health-care cost control. This sections aims to present some of the state of the art approaches in this field.

Segura-Bedmar *et al* proposed a pattern-matching approach to extract relations between drugs from biomedical texts using also shallow parsing for the linguistic analysis [SBMdPS10]. They entrusted the task of defining the domain-specific lexical patterns to a pharmacist, trusting in his professional experience. Their approach was based on Huang *et al* work, which proposed a set of syntactic patterns to split long sentences into clauses where pattern matching algorithms could successfully extract relations. With this approach, the system would be capable of detecting appositions, coordinate constructions and relative clauses. After the sentence simplification phase, DDI lexical patterns were applied. Segura-Bedmar *et al* stated that even considering the richness of natural language expressions, DDI were often expressed by a limited number of constructions, being therefore appropriate to pattern-based approaches. Despite their confidence in a pattern based approach, the results obtained presented low recall (14.07%) and a reasonable precision (67.30%).

## Background and Related Work

Proceeding with their previous work Segura-Bedmar *et al* proposed another approach [SBMdPS11], using an IE method based on supervised machine learning and kernel-methods. Just as in their previous work, the authors used the DrugBank database [WKG<sup>+</sup>08] as the source of unstructured textual information about drugs and interactions between each other. In this approach they identify the DDI relation task as supervised learning problem, more particularly as a drug pair classification task. For the development, they used the shallow linguistic kernel defined in Guiliano *et al* [GLR06], which is the combination of two different sequence kernels, a global context kernel and a local context kernel. The global context kernel analyzes entire sentences in order to detect the presence of relations, whereas the local context kernel based its analysis in the context information of the candidate entities. Depending in the configuration parameters of the global context kernel (n-gram) and the local context kernel (windows-size) the authors achieved different results, presenting a maximum precision of 52.07% but with low recall and a maximum recall of 78.63% but with low precision. Therefore, after maximizing the F-measure for testing all the models, the authors presented a F-measure of 60.01%, with 51.03% and 72.82% for precision and recall respectively.

Another research study, by García-Blaso *et al* [GBMVDR11], proposed a different approximation for DDI detection on biomedical texts based in automatically determining patterns that identify DDI from the training set of documents. Considering the many different ways of possibly describe a drug drug interaction using natural language, this approaches intends to find patterns that are repeated in the large amount of biomedical texts. The method presented is language and domain independent, and relies on Maximal Frequent Sequences, which are used to extract the patterns that will allow the automatic extraction of DDI. For each Maximal Frequent Set extracted it is determined how likely it is that it describes a DDI, and then apply it to a set of biomedical documents to see its performance. This approach demonstrated reasonable results, with an average 51% precision and 67.25% recall. These results are promising, especially since this approach can be adapted independently from both language and context.

Björne *et al* [BAPS11] developed a system capable of extracting DDI for drug mention pairs found in biomedical texts, approaching the DDI problem as a classification task using machine-learning approaches. The system relies heavily on deep syntactic parsing to build a representation of the relations between drug mentions, abstracting event and relation extraction by using an extensible graph format. The system extracts information in two main steps: detection of trigger words (nodes) denoting entities in the text and detection of their relationships (edges). However, before building the machine learning examples, all sentences were processed with a deep syntactic parser. For the classification task, both support vector machines (SVM) and regularized least-squares (RLS) were tested, since they are regularized kernel methods. Björne *et al* present a final F-score of 62.99%, stating that the basic machine learning approaches are suitable for this task.

Segura-Bedmar *et al* have also developed an anaphora resolver to be applied for DDI extraction in pharmacological documents [SBCDPSM09], attempting to improve the recall of the extraction methods. This approach for anaphora resolution uses Centering Theory [GWJ95] in order to se-

lect the scope of the anaphoric expressions and assign the correct antecedent. A simpler heuristic that selects the closer nominal phrase has also been experimented in this domain for some types of expressions, relative pronouns and possessive nominal anaphors. One of the most important components in this approach was the use of several domain resources, including MMTx biomedical parser and the UMLS-thesaurus, taking advantage of the shallow syntactic information provided by MMTx and using UMLS to identify the anaphors and implement semantic restrictions to candidate resolutions. An F-score of 76% was achieved, which is an admirable result for this kind of tasks.

Since DDI are one of the main research topics in the pharmacological domain, we were able to identify many studies. In average, we can see that machine learning approaches presented better results than pattern-matching. The machine learning approaches success would be predictable, mainly due to the complexity of the interactions between drugs, being therefore more difficult to design rules for the extraction. We can also state that this is clearly the most difficult and complex topic of information regarding pharmacological products.

## 2.6 Summary

This chapter introduced the main concepts needed to better understand the information extraction task. The essential tasks inherent in IE systems were identified and described, providing essential information for the development of an IE system.

A common architecture is described, identifying a pipeline process usual in IE systems. The different modules that integrate a IE systems, their importance and their execution order is also explained.

There are different kinds of approaches to the IE task, being that the most important ones were described and evaluated.

Bearing in mind that one of the main difficulties of this thesis undergoes by the use of the Portuguese language in the texts where the extraction tasks will be performed, a summary of the most relevant approaches to the IE task in Portuguese was made. In this summary, systems that targeted clinical texts were favored due to the resemblances with the domain of this thesis. This summary revealed a severe lack of studies that targeted the Portuguese language, although in the past few years a few improvements have been noticed.

Finally, an analysis of the state of the art on information extraction in the pharmacological domain is performed. In this analysis, we were able to identify several approaches in this domain, specially targeting the pharmacological products posologies, indications, adverse effects and DDI.

The study performed in the background of information extraction and the research held for similar studies was imperative for the development of this thesis, since a great deal of the knowledge needed for the development of such a system was obtained from these studies and many synergies from similar studies were exploited.

## Chapter 3

# Resources and Tools

This chapter intends to present some of the resources that were imperative for the development of this thesis. Firstly, in Section 3.1 are presented some of the resources that were used in the NER task. In Section 3.2, a brief description of the processing tools that were used through the development of this thesis is performed. These tools range from the NLP infrastructures chosen for the development of the system, to the tools used for the ontology management task.

### 3.1 Resources

Information extraction systems, particularly the NER task, often presents the need for extensive gazetteers, listing, for example, names of people, organizations, locations or other named entities. This compilation of high-quality gazetteers is commonly claimed to be the bottleneck of IE systems, essentially due to the time and human effort required [MGM98]. This limitation can be overcome with the use of external knowledge, like for instance UMLS, in order to improve semantic category label extraction. In this section a brief introduction to the corpus is also performed.

#### 3.1.1 UMLS

The United States National Library of Medicine (USNLM) started in 1986 a long-term research and development effort known as UMLS. This effort was encouraged by the anticipation of an increasing amount of biomedical information available in electronic systems, being therefore needed a way of facilitating the development of advanced information systems able to retrieve and integrate information from many sources such as bibliographic databases, patient record systems, factual data-banks and knowledge bases. The major barrier for effective retrieval and integration of information from multiple sources was identified as the "naming problem" or the variety of different ways that the same concepts are expressed in different information sources and by different information seekers [SJNH].

Currently, the UMLS [oM08] is a synopsis of the many controlled vocabularies in biomedical sciences. It can be considered a comprehensive thesaurus and ontology of biomedical concepts, providing a mapping structure among the vocabularies.

There are currently three major UMLS knowledge sources:

1. Metathesaurus, which is the core database of UMLS, collecting concepts and terms from many biomedical vocabularies and their relationships;
2. Semantic Network, consisting in a set of sensible relationships among the broad semantic types or categories to which all Metathesaurus concepts are assigned;
3. SPECIALIST Lexicon, a database containing syntactic, morphological, and orthographic information for biomedical and common words in the English language.

Four of the current 150 UMLS source vocabularies are available with a Portuguese translation, being therefore of great importance for any development for the Portuguese language. These vocabularies are listed below:

- The Medical Dictionary for Regulatory Activities Terminology;
- the International Classification of Primary Care;
- the WHO Adverse Drug Reaction Terminology;
- DeCS, the Portuguese translation of MeSH.

From the version used, the 2011 version, we were able to identify 77 112 different concepts and 157 675 unique concepts names with Portuguese translations. In order to allow the extraction system to access the contents of the UMLS records, we developed a SQLite database. All the UMLS records with a Portuguese translation were added to the database, creating therefore a crucial resource for the extraction process here targeted, or for any other system that aims to recognize clinical terms in the Portuguese texts.

During the development of the information extraction system, we noticed that the semantic types used in the UMLS are very specific and granular, which hindered the extraction process, emerging therefore the need for the creation of semantic groups. Hereupon, 3 main semantic groups were created: Disorders, Procedures and Anatomical Sites. According to the semantic type given by the UMLS, when a term was to be found, it would automatically be allocated to one of the groups. The association between the UMLS semantic types and PharmInx semantic groups is given in Table 3.1.

### 3.1.2 Pharmacological Therapeutic Records

The pharmacological therapeutic records is assumed as a set of guidelines for the use of therapeutic drugs and is an imperative resource for the rational use of medicines, supporting the prescription of pharmacological products. This document contains information regarding all the drugs marketed



Table 3.1: UMLS semantic types and PharmInx semantic groups

Semantic Groups	UMLS Semantic Types
Anatomical Site	Anatomical Structure, Body Location or region, Body Part, Organ, or Organ Component, Body Space or Junction, Body System, Cell, Cell Component, Embryonic Structure, Fully Formed Anatomical Structure, Organism Attribute, Tissue;
Disorder	Acquired Abnormality, Anatomical Abnormality, Bacterium, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Finding, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Sign or Symptom;
Procedures	Diagnostic Procedure, Educational Activity, Health Care Activity, Laboratory Procedure, Molecular Biology Research Technique, Research Activity, Therapeutic or Preventive Procedure;

in Portugal and available in the community pharmacy, as well as some products only available in hospitals.

This document organizes 1357 pharmacological products through 19 chapters, according to the main target of each product. In each chapter, that can contain other smaller chapters, each product has an individual section (see Figure 1.1) where indications, adverse effects, counter indications, drug-drug interaction and the posologies are presented. All these topics were briefly described in Chapter 1.

It is noteworthy to state that the text targeted within this document is very formal, and usually uses short and direct sentences, easing therefore the extraction process. Although, the document contains many elliptical sentences, where the pharmacological product is omitted since it can easily be inferred from the document structure. Regarding the posology information, usually the information is presented in a descriptive manner, from where we can infer some patterns.

This document is an annual edition, being that the file used was the last version, from the year 2011. The information is available online in an HyperText Markup Language (HTML) format or available to download in Portable Document Format (PDF). In order to allow an easier access to the information, we used the PDF file as the source document.

In order to be able to process the text present in the document, some pre-processing tasks were performed (see 5.2). Thus, an Extensible Markup Language (XML) file with the information relevant for the purpose of this work was generated from the original PDF file.

Table 3.2 presents some statistics about the content of each information topic present in the document.

Analyzing the table we can see that there is a considerable amount of data present in the document. The information topic with more content is the posology. It is interesting to notice that although adverse effects presents the highest number of sentences, the number of characters is relatively low (characters for sentences ratio of 38), supporting the idea that this topic is generally expressed with many short sentences.

Table 3.2: Statistics regarding the content of the information topics in the document where the extraction process is held

Topic	#Characters	#Tokens	#Sentences	#Characters/#Sentences
Indications	108 333	18 500	2148	50
Adverse Reactions	135 856	24 118	3567	38
Counter Indications	132 116	24 027	2604	51
Drug-Drug Interactions	118 999	20 227	2170	55
Posology	199 480	49 154	2595	77
<b>Total</b>	694 784	136 026	13084	53

## 3.2 Processing tools

An important and decisive challenge in the development of this thesis concerned the decision of which tools should be used and for which purposes, considering the several different components that compose this work. Generally, NLP tools are complex, aim to solve very specific problems, and aim to fit in other specific systems. In the last years several repositories have been created, including for example the Linguistic Data Consortium [ldc], the Natural Language Software Registry [nls], the European Language Resources Association [elr] and the distributed language resource center for Portuguese, Linguatca [lin]. However, it is extremely difficult and unlikely to attempt to reuse modules from these repositories for the construction of NLP systems with some interoperability.

The tools that were imperative for the development of this thesis are described in this section, including also some of the reasons that determined their specific choice.

### 3.2.1 UIMA

The Unstructured Information Management Architecture (UIMA) is an open, industrial-strength, scalable and extensible framework that supports the definition and integration of software modules that perform analysis on unstructured data such as text documents or videos [FL04]. UIMA has been originated at IBM, however it has already moved on to be an open source project which is currently incubating at Apache Software Foundations [uim].

Recently, an increasing number of members of the NLP community have adopted UIMA [Gue08] [HBL<sup>+</sup>08] [KR08] as a platform for the creation of reusable NLP components, which can be assembled to address different NLP tasks depending on their order, combination and configuration.

An UIMA application can be roughly divided in two phases: analysis and delivery. In the analysis phase, collections of documents are collected and therefore analyzed, being that the results are stored in one or more forms, depending on the needs of the delivery phase. This phase may include tokenization, semantic class detection on the input documents and may use structured sources, such as dictionaries or ontologies, to find and annotate named entities.

Afterwards, in the delivery phase, all the analysis results and possibly even the original documents or other structured information are made accessible. In this phase, the application can present the results, by for example present a query interface that enables the user to search for a combination of tokens, entities, and relationships through a semantic search engine, or by saving the information in a machine-understandable format, such as a database or an ontology, for later analysis.

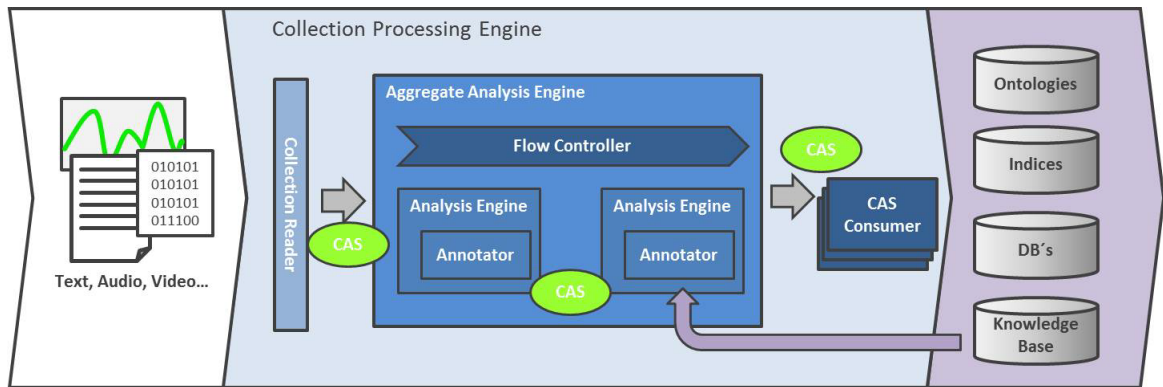


Figure 3.1: UIMA System Architecture (original figure from Liliana Ferreira (2011) [dSF11])

Two key components of UIMA are the Common Analysis Structure (CAS) and the UIMA type system. The UIMA type system is a declarative definition of an object model [VJRH09] and serves two main purposes:

1. to define the kinds of meta-data that can be stored in the CAS;
2. to support description of the behavior of a processing module, or analytic, through specification of the types it expects to be in an input CAS and the types it inserts as output.

The CAS is the basic data structure in which both the unstructured information being analyzed and the meta-data inferred for that information are stored. Therefore, providing CAS contents to analysis engines, enabling read access to the artifact being analyzed (text, video) and read/write access to the analysis results, these engines can discover and add more meta-data to the CAS. Analysis engines may be organized and composed together to form reusable components that encapsulate rich workflow of cooperating engines.

There are two more primary components extremely important and useful in UIMA which are the Collection Reader and the CAS Consumer. The Collection Reader is able to connect and iterate through a source collection, acquiring documents and initializing the CAS for analysis.

CAS Consumers function at the end of the flow and are responsible for the final CAS processing. A CAS Consumer may be implemented, for example, to add CAS contents to an ontology, a database or even to index the results in a search engine.

The complete flow from source, through document analysis, and to CAS Consumers supported by UIMA is illustrated in Figure 3.1.

Other development frameworks, like for instance GATE [CMBT02], were considered and duly analyzed. However, UIMA was chosen mainly due to its interoperability, and for providing easier development and integration of several components.

### 3.2.2 Protégé

Protégé is the ontology development tool chosen to support ontology development in this work. Protégé was developed by the Stanford Medical Informatics group at Stanford University, and began as a small application designed for a medical domain (protocol based therapy planning), but has evolved into a general-purpose set of tools [GMF<sup>+</sup>03]. Protégé is an extensible and customizable framework for constructing knowledge bases, and has an open architecture, allowing an easy integration with many other applications. One of the major benefits of Protégé is that it is knowledge representation independent, contrarily to many other ontology development tools, as Ontolingua Server [FFR96], the OntoSaurus [SPKR96] and WebOnto [Dom98] tools. Since Protégé is written in Java, it is able to run under a wide variety of operating systems. Protégé presents an intuitive user interface that allows developers to create and edit domain ontologies. The numerous plugins available provide alternative visualization mechanisms, enabling management of multiple ontologies, the use of inference engines and even problem solvers with Protégé ontologies, along with many others functionalities.

### 3.2.3 Jena

Jena [jen] is a Java framework for building Semantic Web applications based on W3C recommendations for Resource Description Framework (RDF) and Web Ontology Language (OWL). It provides a programmatic environment for RDF, RDF Schema, OWL, SPARQL Protocol and RDF Query Language (SPARQL) and includes a rule-based inference engine. Jena is open source and was created in the HP Labs Semantic Web Research Programme.

The Jena Ontology Application Programming Interface (API) can be extremely useful in this work, since it defines object classes to represent RDF graphs, resources, properties and literals. The Jena API will therefore be used to develop a programmatic representation of the ontology to be created, modeling, structuring and providing access to its content and ensuring its correct population.

### 3.2.4 Knowtator

Knowtator [Ogr06] is a general-purpose text annotation tool implemented as a Protégé plug-in and running in the Protégé environment. Knowtator is not the only text annotation tool that has emerged in the last few years, there are also tools such as WordFreak [wor], MMAX2 [mma] and even GATE [CMBT02], which is a software architecture for NLP that incorporates many components, including one for text annotation. However, the need to create a complex annotation schema, with hierarchical and constrained relationships among the annotation types lead us to use Knowtator.

Since Knowtator runs within Protégé, it can avail Protégé’s knowledge representation capabilities to specify more complex annotation schemas, using Protégé’s classes, instances and slots. Hereupon, Knowtator can model both semantic (*e.g.* drug-target interactions) and linguistic phenomena (*e.g.* co-reference resolution).

### 3.3 Summary

In this chapter some imperative resources are identified and described, starting by the information resources. The information resources described had an extremely important role specifically in the NER task, supporting the identification and consequent classification of the entities present in the text.

The tools described were the main tools used for the development of the system and of some of the artifacts inherent for the good functioning of the IE system, like the ontologies.

Therefore, we believe that this chapter presented the fundamental knowledge regarding some of the resources imperative for this thesis.

## Resources and Tools

## Chapter 4

# Knowledge Representation

The representation of the information extracted is of extreme importance for this information extraction system, since only with an appropriate representation we can ensure that after the execution of the system such information will be conveniently persisted.

In this chapter we start by explaining the methodology used for the development of the ontology. Afterwards, we describe the extensible knowledge model used to structure the entities and respective relations extracted from the pharmacological texts, providing an organized and structured representation of the information. This model is referred as the *PharmInx* ontology.

### 4.1 Ontology Development Method

Ontologies must be designed in order to be internally consistent and to support interoperability while offering a wide coverage of the domain. Thus, special attention was given to the standard principles of ontologies development, namely *consistency*, *completeness*, *conciseness*, *expandability* and *sensitiveness* [SS04]. Efforts were made so that the ontology developed in PharmInx would not allow any contradictory conclusions from any of the definitions or axioms, nor would it store unnecessary definitions..

The modelling process of an ontology is not a trivial task, considering that there is not a single correct methodology for ontology development. According to the purpose and content of an ontology, different methodologies must be considered and applied. Nevertheless, there are some basic guidelines proposed by Noy and McGuinness [NM01] that we considered when developing the ontology for PharmInx. According to these guidelines, the following steps were performed:

1. Identification of the purpose and scope of the ontology, *i.e.*, clarifying the reason why the ontology is being built and identifying its intended uses and which are the relevant terms in the domain;
2. Identification of important terms and respective properties that appear in the domain;

3. Definition of the classes and class hierarchy;
4. Definition of the properties for each class.

## 4.2 PharmInx Ontology

The primary goal of this ontology was to model information of the pharmacological products and the concepts they describe. Currently, the model is primarily focused on the information about posologies, indications and adverse reactions. Hereupon, the model of knowledge representation includes general medical entities such as procedures, disorders, and anatomical sites, but also with some more specific entities as for example dosages and frequencies of administration.

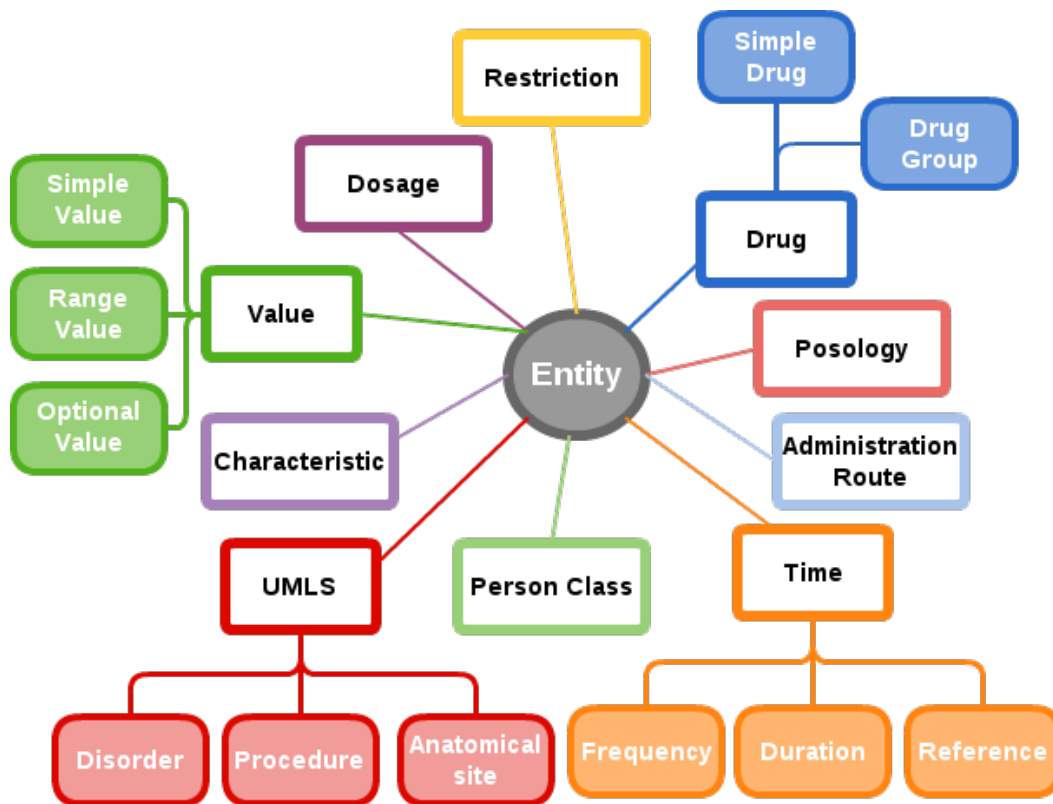


Figure 4.1: Tree of entities, types and subtypes considered in the knowledge representation of the information regarding pharmacological products

In Figure 4.1 we can see a summarization of the classes and subclasses present in the PharmInx ontology, along with their hierarchical representation.

Table 4.1 presents more detailed information about the classes present in the model, while Table 4.2 describes all the relations between classes.

Figures 4.2, 4.3 and 4.4 present some partial views in order to ease the interpretation of the model. Figure 4.2 shows a partial view focused mainly on the posology, duration, frequency and dosage, including their respective relations. In this partial view we can see that both dosage and



Table 4.1: Classes of the PharmInx Knowledge Representation

Category	SubClasses	Description
Time	Duration Frequency Time Reference	Temporal expressions, including durations ( <i>e.g.</i> two weeks, ten years), frequencies ( <i>e.g.</i> 5 times a week) and temporal references ( <i>e.g.</i> morning, evening) typically used in medical texts.
Dosage		Identifies the dosage for a given pharmacological product, including a value and its respective unit;
Administration Route		The route of administration of a given pharmacological product;
Person Class		Represents an age group for which the pharmacological product is indicated ( <i>e.g.</i> adults, children, newborns);
Restriction		Describes a restriction, which generally focus on temporal restrictions (before or after an event) or age and weight restrictions regarding the patient;
UMLS Entity	Disorder Procedure Anatomical Site	All references to medical entities. These medical entities can for instance be disorders, which represent health states, diseases, symptoms, pathologies or any other problems manifested by the patient. Therapeutic procedures are also identified by this class, including all kind of actions that can be performed with the purpose of measuring or treating some aspect of a disorder. Lastly, all references to anatomical structures or locations are also incorporated in this class;
Posology		All references to a possible posology of a pharmacological product. Posologies basically consist in a large set of relations where the other entities are assembled in order to generate more consistent and relevant information;
Characteristic		All references to terms that characterize another class individual;
Drug	DrugGroup Simple Drug	All references to pharmacological groups and also for specific pharmacological products;
Value	SimpleValue RangeValue OptionalValue	All forms of quantifications, including simple quantifications ( <i>e.g.</i> two), a range of values ( <i>e.g.</i> five to ten) or even optional values ( <i>e.g.</i> five or six);

Table 4.2: Relations of the PharmInx Knowledge Representation

ObjectProperty	First Argument	Second Argument	Description
belongsTo	Drug	Drug Group	Relates a Drug with the DrugGroup where it belongs. This property is transitive;
hasAdminRoute	Posology	Administration Route	Describes the recommended administration route for a specific posology of a pharmacological product;
hasCharacteristic	PharmInxEntity	Characteristic	Provides a characterization of the domain class through a characteristic;
hasChild	UMLS	UMLS	Relates an UMLS individual with another more specific individual due to the addition of a characteristic ( <i>e.g.</i> hypertension <i>hasChild</i> severe hypertension);
hasDosage	Posology	Dosage	Describes the dosage that should be applied for a given Posology;
hasDuration	Posology	Duration	Identifies which is the duration of administration to be applied in a given Posology;
hasFrequency	Posology	Frequency	Relates a Posology with its valid frequencies of administration;
hasOperand	Restriction	PharmInxEntity	Considering that besides the operator a restriction also includes operands, this relation aims to identify those operands, being that <i>hasLeftOperand</i> and <i>hasRightOperand</i> further specify the respective position of the operand regarding the operator;
hasPersonClass	Posology	PersonClass	Identifies the age group for which the posology is recommended;
hasPosology	Drug	Posology	Relates a Drug with its respective possible posologies;
hasRestriction	Posology	Restriction	Identifies the Restrictions present in a given posology;
hasTarget	Posology Drug	UMLS	Relates a Posology or a Drug with a medical entity (UMLS) for which it is recommended. When the UMLS is a Procedure we understand that the pharmacological product aims to support such a procedure, whereas in the case of Disorders the drug aims to work as treatment;
hasValue	Dosage Duration	Value	Relates a Dosage or a Duration with their respective Values;
includesUmls	UMLS	UMLS	Relates an UMLS individual with its smaller aggregated UMLS individuals, which is common with longer medical terms ( <i>e.g.</i> genital infection by Chlamydia <i>includesUmls</i> genital infection and <i>includesUmls</i> Chlamydia). This relationship is needed since long medical terms are often missed by the external knowledge resources, being therefore needed the concatenation of smaller individuals;
involvesTherapeutic	Posology	Procedure	Relates a Posology with the necessary procedures for its proper appliance;

duration are associated with a value element, which can represent simple values, optional values or even range values. However, *Frequencies* are not modeled similarly since their value is divided into the administration value and the time value, which are both simply integers. Thus, in our system frequency means: “AdminValue dose for each TimeValue TimeUnit” (e.g. 2 doses every 1 week or 1 dose every 12 hours). Further analyzing the figure, we can also see that for a valid *Posology* only a *Dosage* and at least one *Frequency* are mandatory, and that we can only have one *Duration* but many *Administration Routes*, *Person Classes* and *Frequencies*.

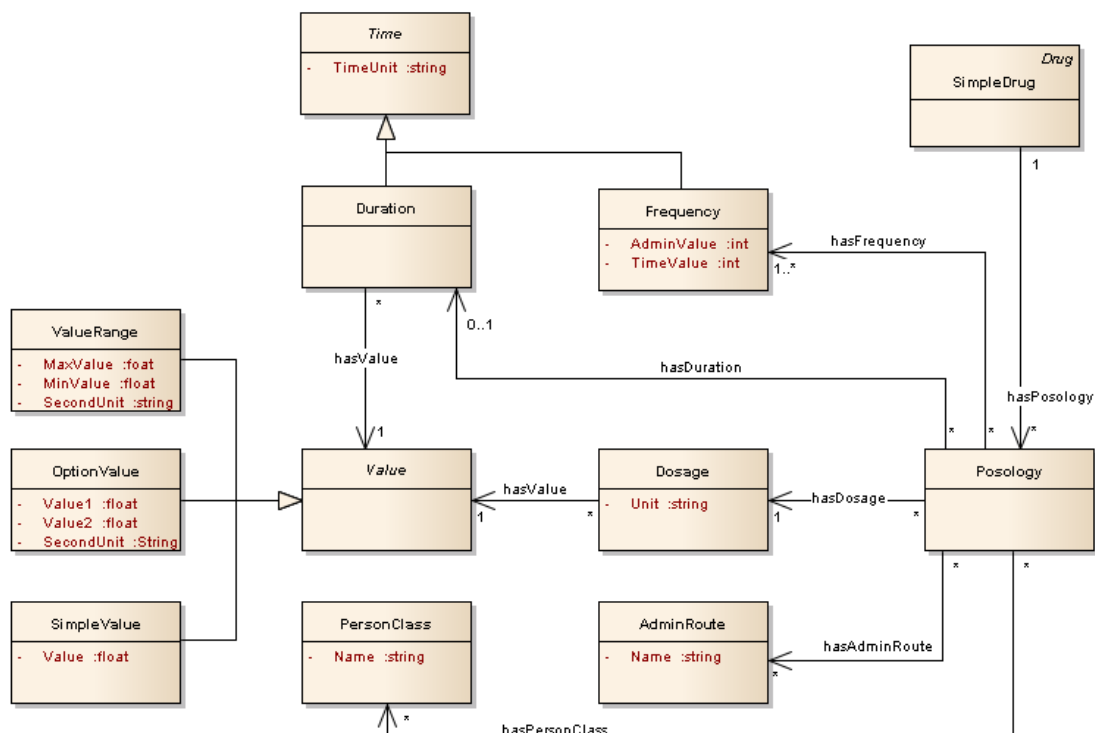


Figure 4.2: Partial view of the knowledge representation model, focusing in the classes *Posology*, *Dosage*, *Time*, *AdminRoute* and *PersonClass*

We can see in Figure 4.3 that, according to the subclasses of the UMLS, different relations can be established with both the posology or the pharmacological product. Generally, when the UMLS entity is a *Disorder*, the drug or a specific posology of the drug, can treat that disorder and, when the UMLS is a *Procedure*, the drug supports such a procedure, being therefore both of this situations modeled by the *hasTarget* relationship. However, a posology can also include a set of procedures to support the correct administration of the drug, which is modeled by the *includesTherapeutic* relationship. A pharmacological product can also be related to a disorder when this disorder is an adverse reaction for the administration of the product, being such a relation modeled by the *hasAdverseEffect* relationship. Through the diagram we can also understand that the modeling of the more complex UMLS terms, terms that involve more than one match from the UMLS external source, is conveniently performed with the *includesUMLS* definition. We can easily access the compounds that aggregate a complex UMLS term. In order to allow the access

of all the *UMLS* terms that were derived from a previous one through the addition of Characteristics, the *hasUMLSCChild* relationship was also included. These two relations, *hasUMLSCChild* and *includesUMLS*, are very important in order to ease afterwards queries to the ontology, being therefore easier to know where an UMLS term is used. As an example, consider that you want to search for all pharmacological products that aim to treat “respiratory infections”. With the use of this relation we can also include in the results terms as “severe respiratory infections” or “treatment of respiratory infections”, which would also be of interest.

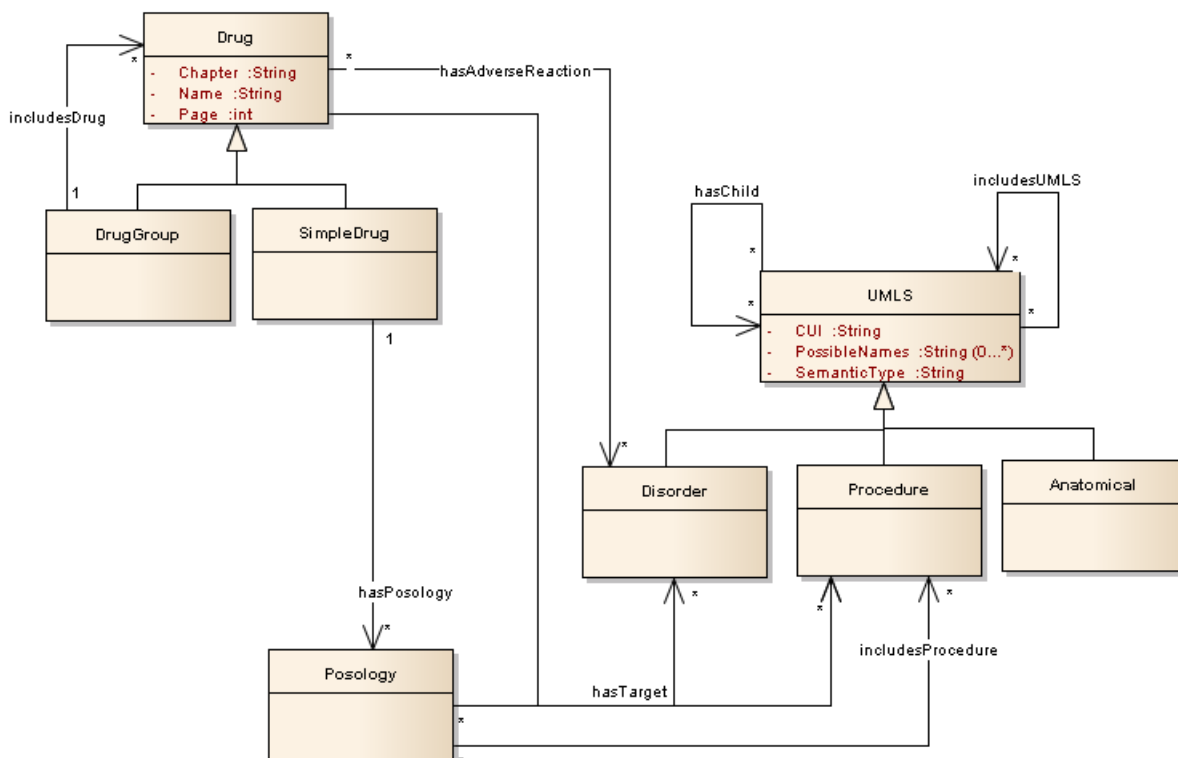


Figure 4.3: Partial view of the knowledge representation model, focusing in the classes *Posology*, *Drug* and *UMLS*

In Figure 4.4 we can see a partial view of the model focused in the restrictions and characteristics. For a correct analysis of the diagram, we should consider that the *PharmInxAnnotation* class can represent any of the other classes, excluding restrictions, characteristics, drugs and posologies. We can see that the restrictions must have one operator and can have up two operands. This allows us to model more complex restrictions (e.g. 30 min before a surgical intervention). We can also understand that almost every annotation can have characteristics, allowing more detailed and complex descriptions.

PharmInx ontology was designed and maintained resorting to the Protégé framework, whereas the population of the ontology was performed with the Jena API. The resulting hierarchy in Protégé is presented in Figure 4.5, along with some individuals from the *Disorder* class. It is noteworthy that the population of the ontology with individuals items is successfully performed in a

## Knowledge Representation

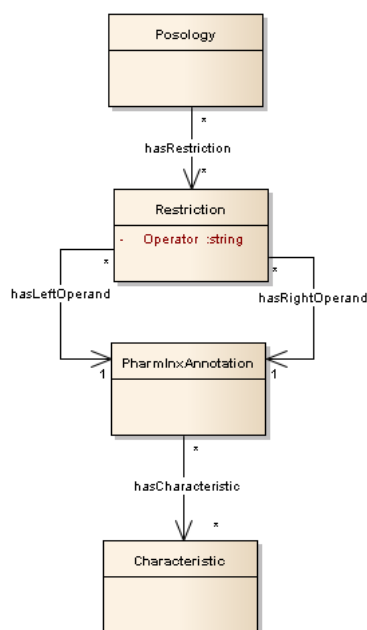


Figure 4.4: Partial view of the knowledge representation model, focusing in the classes *Posology*, *Restriction* and *Characteristic*

completely automatic process, using the information extracted by the IE system PharmInx, which is introduced in Chapter 5.

### 4.3 Summary

The success of an IE system strongly correlates with the coverage and overall quality of the knowledge it is able to output. A proper knowledge representation of the information being handled is imperative to the quality of an IE system.

The current chapter presents the methodology used for the development of a pharmacological ontology, briefly describing some of the efforts performed in order to ensure the overall quality of the output from the system. The ontology developed for the purpose of representing the knowledge extracted by PharmInx was described, through the presentation of its classes and relations and by explaining the most relevant design decisions. The ontology was developed in order to be easily extensible by adding additional concepts and relations, being therefore reusable in different scenarios and easily upgraded in case further progresses in the extraction system are performed.

## Knowledge Representation

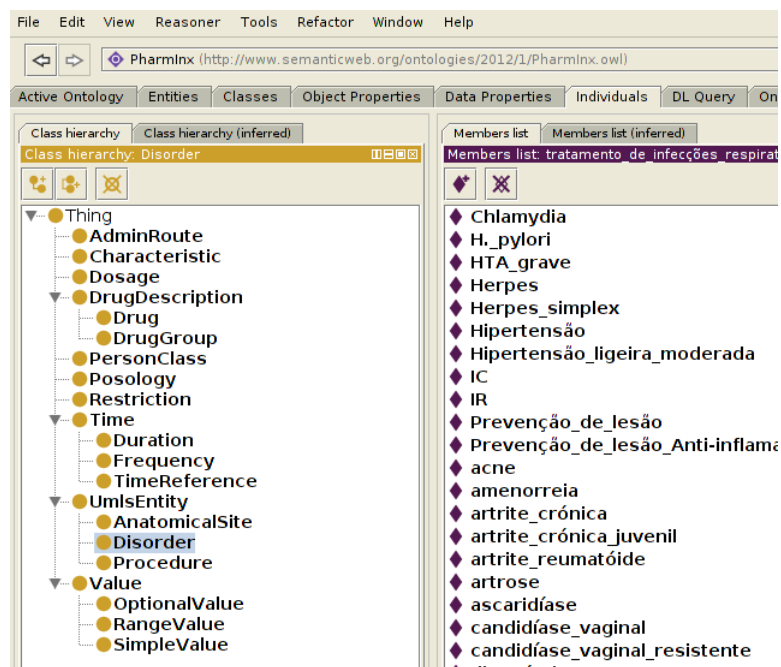


Figure 4.5: PharmInx ontology in Protégé

## Chapter 5

# PHARMacological INformation eXtraction system

We developed PharmInx, an information extraction system that fulfills the goals set out for this project. This system is able to receive as input the documents where the extraction task is held and also some external resources, as ontologies or terminologies to assist the extraction task. The system outputs the concepts identified within the document under analysis and their respective relations with each other.

PharmInX is tailored for clinical domains, specially for the pharmacological domain. However, PharmInX architecture was designed in order to allow flexibility, including several different components that can be rearranged in its work-flow.

The following sections aim to describe the architecture of PharmInX and explain in detail its aggregated subprocesses.

### 5.1 PharmInx Architecture

The analysis of natural language documents is a complex task, that typically involves several steps. The system developed in this thesis implements a set of components that reflect such steps.

In this section we provide a brief description of the general architecture of PharmInX, identifying its main subprocesses and respective interactions.

The main goal of this specific IE system is the extraction of several different entities and their relations from free-form text written in Portuguese. The components of this system follow the fundamental NLP principles, and provide several mechanisms to read, process and utilize external resources, such as ontologies and terminologies. We can divide the possible architecture of this IE system in 6 main components:

1. **Document Pre-Processing:** Although this component is not directly inserted in PharmInX workflow, it is still an indispensable task in order to appropriately execute the extraction

## PHARMacological Information eXtraction system

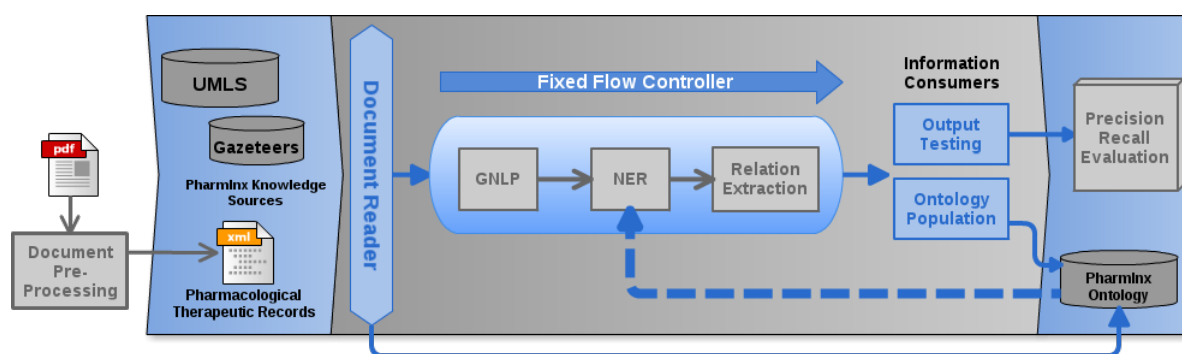


Figure 5.1: System Architecture

task, being therefore executed in beforehand. This task aims to process the initial text, reading its content in order to be able to save it in a XML file, adding tags representing the implicit meaning regarding the document structure.

2. **Document Reader:** This component is able to read the XML input file and load its content and respective embedded tags into annotations for further analysis. The information implicit in the structure of the document is atypically important for this system, since the corpus has a great number of elliptic sentences, where the pharmacological product addressed is usually omitted since it can be easily inferred from the document structure.
3. **General Natural Language Processing:** This component is responsible for the non domain specific NLP tasks, which include sentence discovery, tokenization, stemming and part-of-speech tagging.
4. **Named Entity Recognition:** This component is able to construct annotations that identify entities in the text. In order to fulfill this task, regular expressions, rules, external resources for lookups and the contextual surrounding information are used by several different annotators.
5. **Relation Extraction:** Considering the entities previously identified, this component is capable of inferring possible relations existent between these entities using contextual information and the type of the entities.
6. **Information Consumers:** This component is responsible for ending the process and creating the desired output, using the previous analysis and results. There can be several different information consumers in an IE system. In this particular IE system the two main information consumers are:

- **Ontology Population**, which populates the PharmInX ontology. This component produces an OWL file with the information extracted.



- **Output Testing**, that is responsible for comparing the outputted data with previously and manually annotated information, presenting then the precision, recall and F1-score of the system.

All PharmInX components run within the UIMA framework [FL04], using therefore its type system definition to define all the kind of metadata that can be recognized by the system, as well as, the types each annotator expects to find in the CAS in order to execute its reasoning and the types it will then output through the CAS. Figure 5.2 presents a partial representation of the types used in PharmInX.

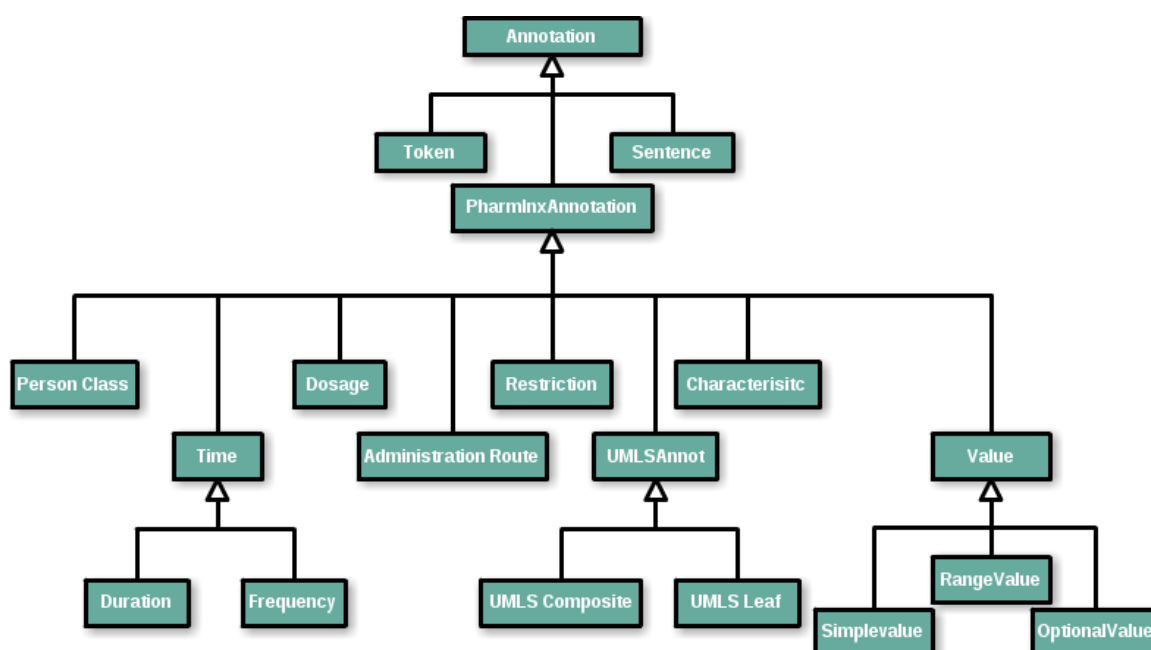


Figure 5.2: Types of annotations present in PharmInx type system

## 5.2 Document Pre-Processing

Considering the relevance and importance of the structure in which the information is presented in the original document and considering the inherent difficulties of preserving this information due to the format in which the document is available, PDF, an attempt to contact Infarmed was performed. This contact aimed to verify the possibility to access the information in the document in an easier to extract format, however, such a request was denied. Thus, a considerably costly and hard working preprocessing task was necessary and executed, resulting in a XML file containing all the information from the PDF. This XML file is conveniently structured to favor the access to the text, embedded with structural information in the XML tags. In order to be able to collect the information inferred in the document structure, several rules and regular expressions were used.

```

▼<Med chapter="1.1.1.1." name="BENZILPENICILINA POTÁSSICA" page="23">
  ▼<indications>
    Infecções por agentes penicilino-sensíveis, nomeadamente faringite, amigdalite, otite média,
    pneumonia, endocardite estreptocócica e meningite meningocócica ou pneumocócica.
  </indications>
  ▼<adverse_reactions>
    Reações de hipersensibilidade incluindo febre, urticária, dores articulares; angioedema.
    Leucopenia e trombocitopenia, usualmente transitórias. Choque anafilático apenas em doentes
    com hipersensibilidade às penicilinas. Contra-Ind.
  </adverse_reactions>
  ▼<interactions>
    A probenecida inibe competitivamente a secreção tubular das penicilinas causando um aumento
    significativo das suas concentrações séricas.
  </interactions>
  ▼<counter_indications>
    História de hipersensibilidade às penicilinas. Reduzir a posologia no doente com IR.
  </counter_indications>
  ▼<posology>
    [Adultos] - Via IV (perfusão intermitente) ou IM: 300000 a 1200000 UI/dia, a administrar de 3
    em 3 ou de 4 em 4 horas. Via IV (perfusão intermitente ou perfusão contínua): 10000000 a
    24000000 UI/dia, a administrar a intervalos de 2 em 2 horas ou por perfusão contínua, no
    tratamento de infecções graves. Reduzir a posologia no doente com IR (Cl cr < 50 ml/min). Via
    intratecal - Não recomendada. [Crianças] - Via IV: < 12 anos: 25000 a 400000 UI/kg/dia, a
    administrar de 4 em 4 ou de 6 em 6 horas.
  </posology>

```

Figure 5.3: Sample of the generated XML with the pharmacological therapeutic records content

### 5.3 Document Reader

Once the preprocessing of the document is done, a collection reader is responsible to connect and iterate through the data stored in the XML file. When the collection reader acquires an artifact to analyze it initializes the CAS, immediately adding structural information in the form of annotations. This information immediately added corresponds to the information that can be inferred from the XML tags, distinguishing the topic under analysis. Therefore, after the processing of this component, the CAS is outputted containing as many annotations as topics that describe the pharmacological product, and containing also the drug name. These annotations are of the utmost importance so that the entities and relations are extracted according to the topic under analysis.

This collection reader also includes an atypical task in this kind of components, which consists in reading the entire input file before any analysis takes place, populating the ontology with the names and structural information (hierarchical relations, chapter and page) about the drug products present in the document. This task enables the system to identify drug references in the text.

### 5.4 General Natural Language Processing

Before any analysis can take place over the corpus, several NLP general tasks must be executed. These tasks aim to identify basic knowledge about the text to process, which are crucial for subsequent analysis in the NLP pipeline. This basic knowledge consists initially of the identification of sentences and tokens, and later of the calculation of the part-of-speech tag and stem of each token. The identification of the stop-words present in the text and the expansion of possible abbreviations are also performed in this component. The general natural language component of PharmInx is

composed by a tokenizer, sentence segmentator, a stop-words finder, a abbreviation expander, a stemmer and finally a part-of-speech tagger .

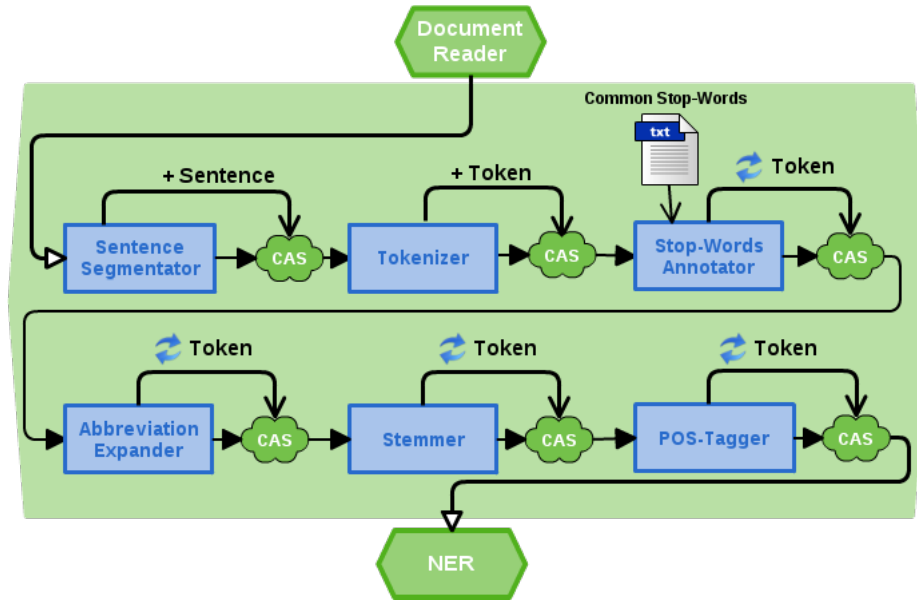


Figure 5.4: Flowchart for NER

Regarding the sentence segmentator and the tokenizer, the basic heuristics considering white space characters and punctuation detection were applied, resorting to the UIMA example packages annotators. Such simple heuristics were successful since special cases like abbreviations are later conveniently handled.

Following the tokenization, all the generated tokens are checked in a list of common Portuguese stop-words, identifying all the known stop-words in the text in order to be used in further analysis.

After the stop-words identification, a search for abbreviations and consequent expansion is executed. This task resorts to a list of the common abbreviations used in the document, which are made available by Infarmed [IM]. This list of abbreviations contains 68 abbreviations and their respective expanded form. A sample of the list of abbreviations used is presented in Table. 5.1.

Table 5.1: Short sample of the abbreviations used in the corpus

Abbreviation	Expansion	English Translation
AL	Anestésicos locais	Local anesthetics
ECG	Electrocardiograma	Electrocardiogram
HTA	Hipertensão Arterial	Hypertension
IH	Insuficiência Hepática	Hepatic Insufficiency
IM	Intramuscular	Intramuscular
Inj.	Injectável	Injectable
RN	Recém-nascido	Newborn

After that, a stemmer is applied to each token not identified as a stopword. The stemmer used is a snowball implementation for the Portuguese language [sno]. After the application of the stemmer on the texts where the extraction process should be held, some common mistakes were identified and resulted in some improvements in order to enhance the performance of the stemmer.

Regarding the POS tagger, it is known that the application of standard POS taggers parametrized through the use of general language corpora originates a great loss of performance when applied in specific domains, such as the clinical domain here targeted [CJ01]. Several corpora were already developed in order to train POS taggers with clinical narratives written in English [LCHC07] [HW04]. However, once again there is a lack of studies in the area for the Portuguese language. Therefore, the POS tags are carefully used in the system, being reserved for situations where any other reliable source can be useful.

Concerning the part-of-speech tagger implementation, an add-on component of UIMA was applied, the Hidden-Markov Model (HMM) tagger annotator [hmm]. This annotator implements a HMM tagger and then employs the Viterbi algorithm [RN93] to calculate the most probable sequence of tags over a sentence and respective sequence of tokens. This implementation receives as input statistical information from a model file, which must be externally generated. Therefore, the model training was executed using the corpus gently provided by Pablo Gamallo [Gam], which consists of a general language corpus tags calculated by his Treetagger implementation for the Portuguese language. The use of this TreeTagger implementation was initially considered, however such implementation would imply the installation of specific software in any device with the intent of executing the system, being therefore a serious portability concern.

Considering the information obtained from the GNLP, excluding the sentence segmentation information, all the other data are stored in the token annotation, which is the underlying annotation since it will be essential for almost every other annotation.

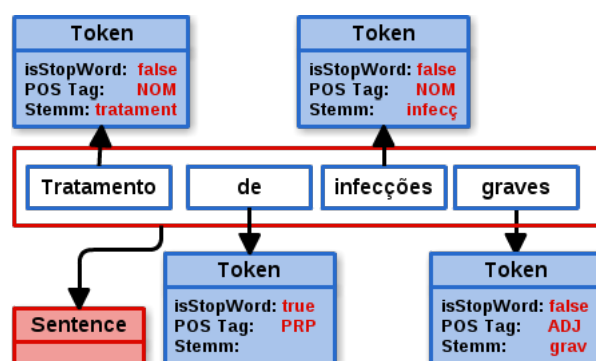


Figure 5.5: Example of sentence and token annotations

## 5.5 Named Entity Recognition

There are several different entities to identify in the text, which requires several different annotators, each one specialized to find and annotate specific entities. Considering the complexity of some of the entities that PharmInx aims to extract, some annotators rely on previous outputs in order to conveniently build more complex and relevant entities.

The NER task is definitely the most expensive task of PharmInx, since several external resources are consulted. Therefore, an attempt to minimize the number of queries to external resources is needed, in order to minimize the time cost of this task. Besides the natural order of the annotators, considering the dependencies regarding input and output, the annotators are set according to a fixed flow also considering their time cost. Hereupon, the annotators that consult larger and consequently slower external resources are saved for last, only analyzing tokens that were not annotated by any of the previous annotators.

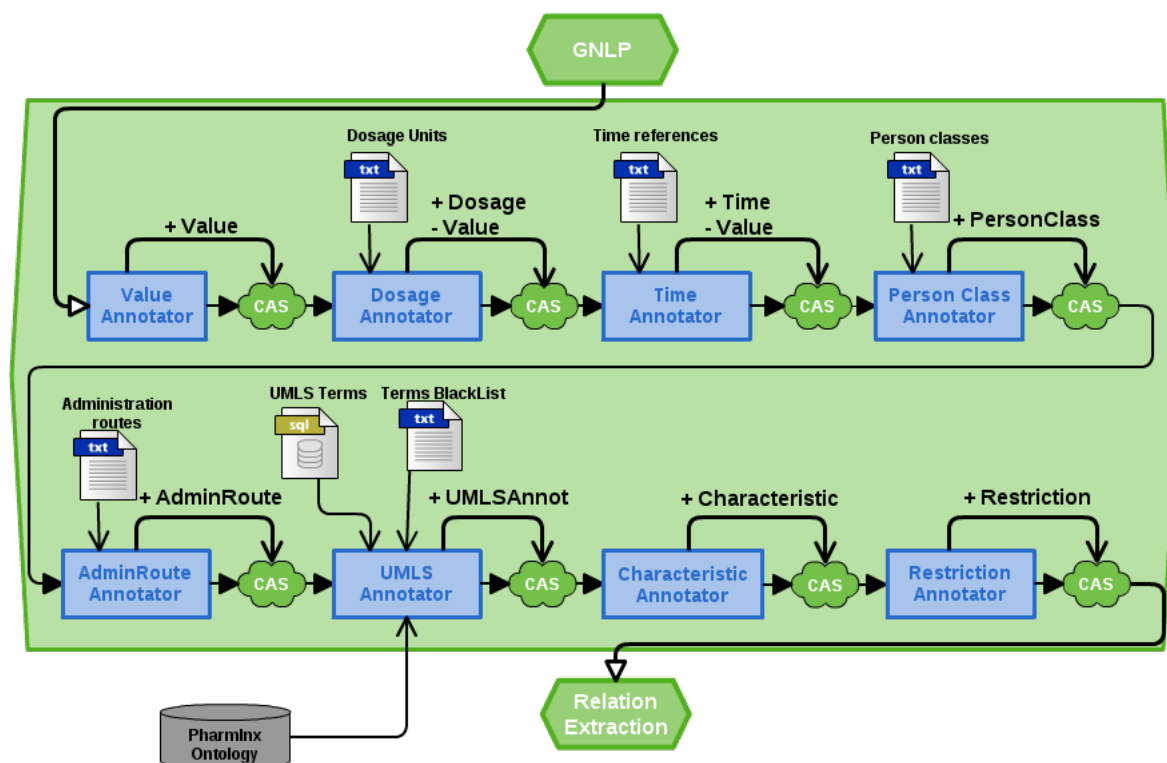


Figure 5.6: Flowchart for NER

### Value annotator

This is the only annotator in PharmInx that does not rely on the tokens outputted by the tokenizer. This annotator, unlike the others, relies exclusively on rules and regular expressions to identify values expressed in the text, being able to identify not only simple values (*e.g.* 10) but also range

values (*e.g.* 10 a 20) and optional values (*e.g.* 10 ou 20). The values extracted by this annotator also include fully spelled numbers (*e.g.* *uma ou duas*).

This is also the only annotator whose annotations should never reach the end of the work-flow, but should rather be consumed by other annotators in order to produce more valuable annotations, as dosages or time references (durations or frequencies).

### **Time annotator**

The annotator aims to identify temporal entities existing in the text. In order to carry out the annotation process, this annotator relies in an external gazetteer list of common temporal references (*e.g.* *horas, dias, mensalmente*) and a set of rules and regular expressions, in order to conveniently discern frequencies, durations or simple temporal references and their respective values. The values of these temporal references are identified by the Value Annotator, which is therefore necessarily processed before this annotator.

### **Dosage annotator**

The main goal of this annotator is to identify dosage indications in the text. In order to do so, this annotator uses an external gazetteer list with the existent dosage units, so that it can then identify the value and associate it with the respective dosage unit, using several rules and regular expressions. Thus, this annotator also relies on the output generated by the Value Annotator.

One concern identified in this annotator, is that any reference that includes dosage units (*e.g.* person weight) is naturally identified a dosage. Thus, a careful analysis of the surroundings of this annotations must be made in order to correctly use its information.

### **Administration Route annotator**

This annotators aims to identify administration routes in the text. The reasoning of this annotator consists simply in comparing the text with terms present in a small external gazetteer that contains the more frequent administration routes.

### **Person Class annotator**

Using a simple access to an external gazetteer list, this annotator is able to identify classes of persons referenced in the text. The classes of persons here targeted can be for instance, adults, children, elderly or even newborns.

### **UMLS annotator**

This annotator aims to match the terms in the text being analyzed with UMLS terms, in order to identify therapeutic procedures, disorders, anatomical sites, or any other kind of clinical information existing in the UMLS records.

This annotator starts its processing by filtering all tokens present in the text, in order to exclude all stop-words and tokens already annotated, and then sorting them in sets for further analysis. These sets are constructed so that tokens in different sentences or far from each other will not be searched together. Then, starting by the first token in each set, the annotator queries the UMLS terms sources for terms starting with similar text as the one present in the token. Then, these results are sequentially filtered by iterating through the tokens and verifying if there are still results that match these multiple tokens. This iterative filtering ends when the number of matched terms drops to zero or if the group of tokens in analysis is finally over. Regardless of the stop condition, the result chosen will be the result with more tokens that scores the best when compared with the text under analysis using the Eq. 5.1, which uses the Levenshtein distance [Lev66]. Nevertheless, one result that aggregates less tokens can be chosen if the other results do not present a score above 85% in the evaluation method above identified. This evaluation method favors the resemblance between the text (Levenshtein distance) but also considers the size of the two different texts, presenting a reasonable results in a percentage format. It is noteworthy that this evaluation method includes a pre-processing to the texts being compared, eliminating any stop-words and replacing special characters and uppercases.

$$similarity(str1, str2) = \frac{(str1.length + str2.length) \times 100}{LevenshteinDistance(str1, str2) + str1.length + str2.length} \quad (5.1)$$

This annotator relies on a particularly time consuming external resource, which is the UMLS SQLite database (see 3.1.1). Considering the time consumed by queries in such a database, some procedures were considered in order to reduce the number of queries done to this database. This reduction was possible resorting to querying the annotations stored in the CAS (the annotations done so far in the text under analysis) and querying the PharmInx ontology, which contained all the annotations extracted from previous texts. With these two measures we were able to achieve a large reduction in the time consumed by this annotator.

Figure 5.7 illustrates the general work-flow of the processing of this annotator, already considering the three different information sources.

After the annotation is finally created, one last verification must be performed, aiming to verify that the annotation being saved does no overlap with any other UMLS annotation previously identified. In cases where such situation is found, a merge of these annotations is performed, resulting in a final UMLS annotation which aggregates the previous ones, enabling therefore an easier processing for further analysis, such as relations extraction. This merge also enables the generation of new complex terms which will certainly be more accurate to properly express the specific meanings in which the simpler terms appear together. As an example of this situation, consider the example in Figure 5.8.

In the UMLS vocabularies there is not any term that matches the entire text from Figure 5.8. However, we are able to identify two terms that can partially annotate the text, which are represented as *UmlsLeaf1* and *UmlsLeaf2*. After the second term is found by the annotator, when

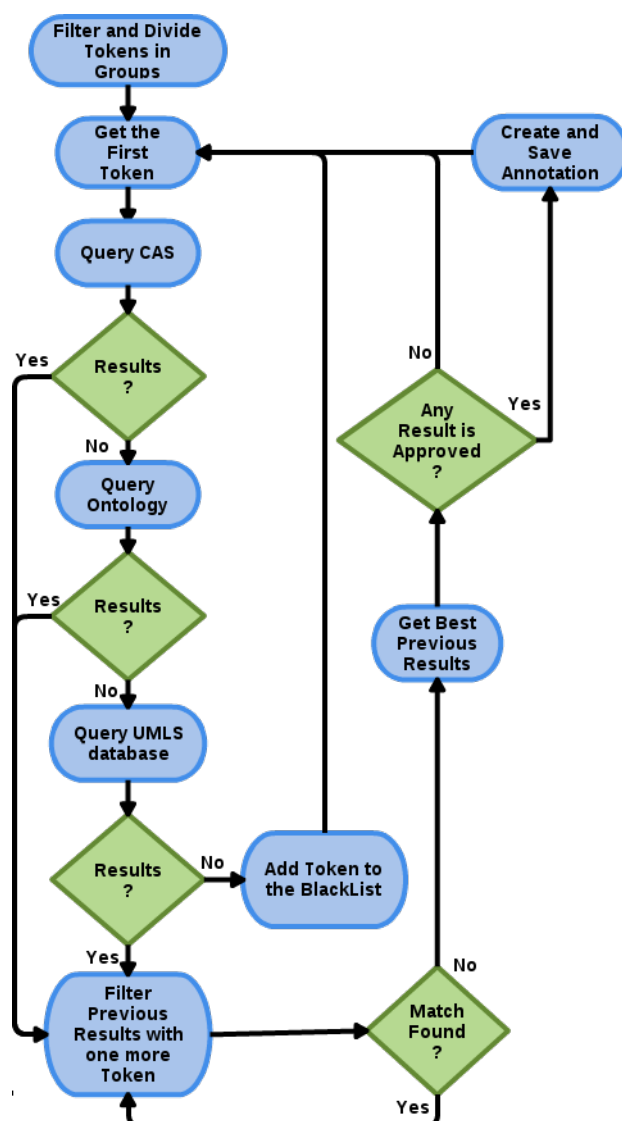


Figure 5.7: General work-flow chart of the UMLS annotator reasoning

saving the annotation the merge will be performed, being, therefore, saved in the ontology the two simple and the UMLS resulting from the merge, the *UmlsComposite* in the figure. Hereupon, the meaning of the term is certainly more accurately saved and when such a text is found again we can immediately recognize it, or even if only one of the simpler terms is found it is also immediately identified, since all the terms are conveniently saved in the PharmInx ontology.

### Restrictions annotator

This annotator relies on regular expressions to find restriction operators (*e.g.* >, <, after, before) among the text under analysis. However, since only the existence of an operator cannot be consider a valid restriction, it immediately searches the neighborhood of the operator in order to find other



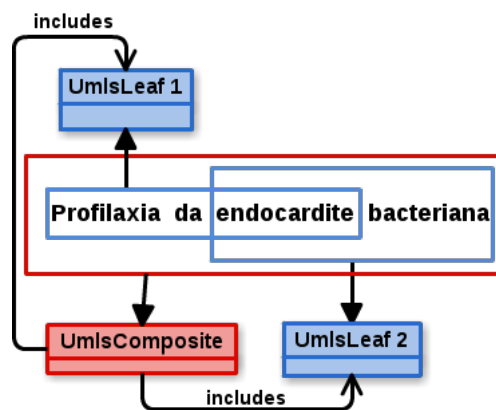
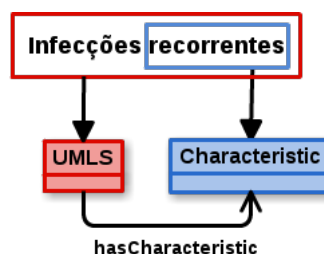


Figure 5.8: Example of UMLS leaf and composite annotations

annotations. If annotations are found, these are associated with the restrictions, classified as left or right operand according to their position over the operator.

### Characteristics annotator

After all the other annotators have finally finished their processing, all the tokens are analyzed one last time, verifying if they belong to any of the annotations found. For those who are not contained in any annotation, are not stop-words, are adjectives or possible adverbial characteristics and are immediately after a given annotation, we consider them as a characteristic of that annotation.

Figure 5.9: Example of a *Characteristic* annotation

Considering the example in Figure 5.9, only the word “*infecções*” is identified by the UMLS annotator as a disease, however, the Characteristics annotator will then correctly identify “*recorrentes*” as being a characteristic of the identified *Disorder*, resulting in a single UMLS annotation which includes the characteristic. Although characteristics are clearly more frequent in UMLS annotations, some other types of annotations can also include characteristics.

## 5.6 Relation Extraction

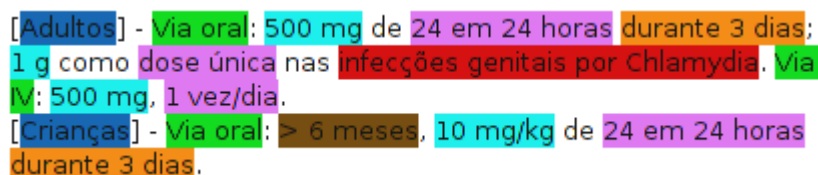
After the correct identification and categorization of the entities in the text, the next step in the pipeline would be the extraction of relations between these entities.

The kind of relations present in PharmInx can be roughly divided in two different types. The first type, the hierarchical relations, are the *is-a* relations between an entity type and its main category. These kind of relations are directly handled by UIMA through the definition of an hierarchical type system, similar to the one illustrated in Figure 5.2, which easily can for instance represent the *is-a* relation between Duration and Time.

The second type of relations present in PharmInx are the ones that represent relations between entities of different categories. These relations are therefore more difficult to find, being the main reason for the existence of a relation extractor component. During the NER task some basic relations are immediately extracted while the correct identification of entities is performed, as it is the case for the relation between values and time or dosages, characteristics with theirs respective annotation or even the restrictions, which immediately identify the operands with annotations in the neighborhood.

However, some more complex relations are needed, as the relations between the posology and its components. Due to the complexity of the relations possible for the posology, a simple search for nearby entities would not be sufficient, requiring a complex method for organizing the entities before these are correctly associated to the posology with the according relation. Thus, a tree structure as been designed, in order to correctly represent the order of the annotations and the scope of their influence.

Considering the example given in Figure 5.10, we can see several different entities correctly identified and categorized, being that the next step in the flow is now to relate these entities with the topic under analysis, in this case the posology.



[Adultos] - Via oral: 500 mg de 24 em 24 horas durante 3 dias;  
 1 g como dose única nas infecções genitais por Chlamydia. Via  
 IV: 500 mg, 1 vez/dia.  
 [Crianças] - Via oral: > 6 meses, 10 mg/kg de 24 em 24 horas  
 durante 3 dias.

Figure 5.10: An example of a posology after the NER task

After a simple analysis of the text we can understand that the scope of influence of the entities is not always the same and, generally, the influence of an entity will remain until another entity of the same category presents itself to take its place. Considering the example, we understand that the influence of the entity *Adultos* remains until the entity *Crianças* appears, and so forth with the other categories.

Noticing these characteristics, our solution consists in relying on the construction of a tree with all the entities, where these are carefully added according to their order in the text, the existence of previous entities of the same category and some more specific rules. Figure 5.11 illustrates such a tree, which is inferred after the NER task and aims to facilitate the extraction of relations.

After the tree is entirely constructed, we can easily notice that starting from the root node and moving directly downwards to the leafs, each path between the root and a leaf represents a

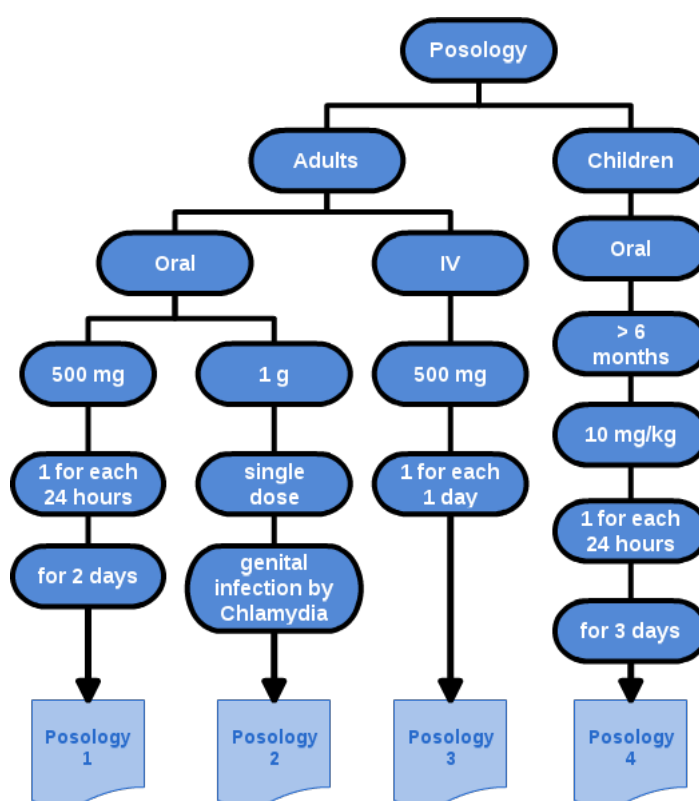


Figure 5.11: Tree inferred from example in Figure 5.10

different posology instance existent in the text. All the nodes that compose this path are features of the posology, being therefore added relations to relate them to the final posology.

## 5.7 Information Consumers

The information consumers component is responsible for extracting the results from the previous analysis and to select the information to be outputted. The output can take any format and does not necessarily mean that the information extracted will be saved, as the information can, for instance, be used for testing.

PharmInx includes two main information consumers were developed, the ontology population consumer and the tester consumer.

The ontology population consumer outputs the information extracted via an ontology, through the population of new individuals. Therefore, all entities and respective relations will be conveniently saved in a valid OWL file. This consumer uses a model of an ontology previously created (see Knowledge Representation 4) and relying on the Jena API creates new instances and fills their respective properties. This consumer is of extreme importance, since it is the consumer that will output the machine understandable data and because the information here saved will be used to enhance later analysis in new text, through lookups in the ontology to find known medical terms. The annotated PharmInx ontology is illustrated in Figure 5.12.

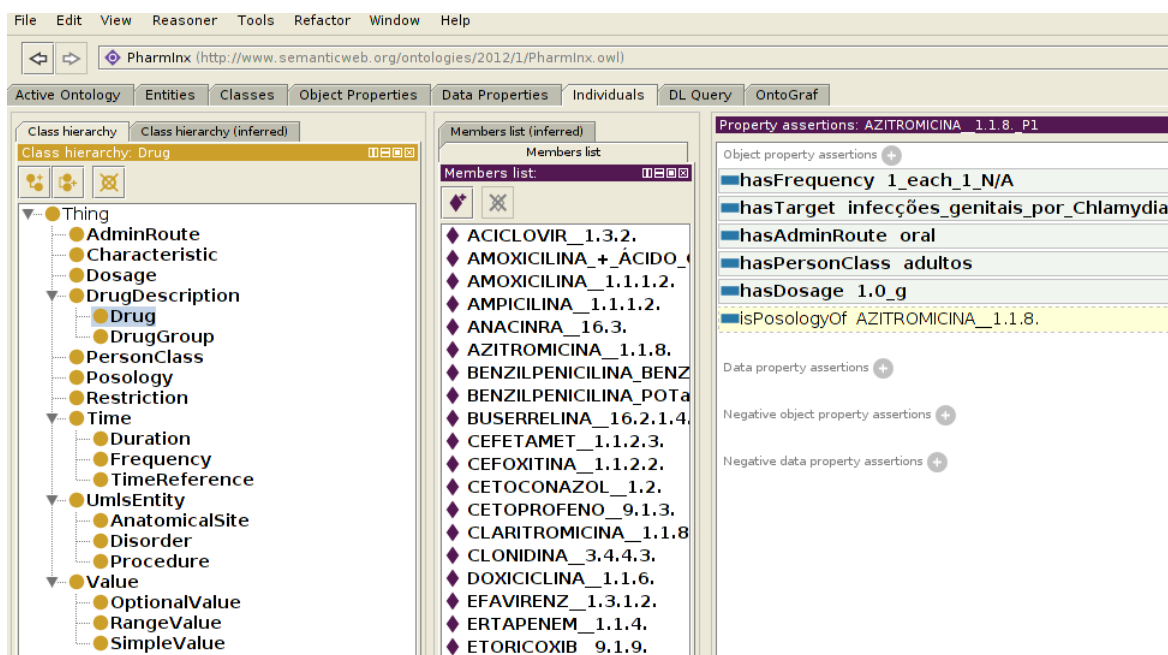


Figure 5.12: Populated PharmedX ontology

The tester consumer does not save the information in any format but rather processes the information extracted and resorting to manually annotated data, compares the extracted and the inputted results. The output of this consumer is therefore an evaluation of the system, calculating IE evaluation metrics such as precision, recall and F1-score of the system, in the overall and for each of the different categories and relations. Further details regarding the evaluation process will be given in Chapter 6.

## 5.8 Summary

This chapter introduced the architecture of the NLP system proposed in this thesis, PharmedX.

PharmedX components are based on the open source framework UIMA and rely in NLP principals to automatically extract information from pharmacological texts, identifying clinical and other entities and their respective relations that result in precious information. The system provides mechanism to read and process the pharmacological texts and to automatically populate the PharmedX ontology with the information extracted, using information from several sources to help the IE task.

The PharmedX components use several different methods to extract information from the text, including rules, regular expressions, lookups in external clinical information sources and even some machine learning techniques. PharmedX also involves complex methods to organize the entities identified in order to enhance the relation extraction phase, resulting in high quality structured output from the system.

## Chapter 6

# Performance and Evaluation

In order to evaluate the performance of the system proposed and consequently developed in this dissertation, several evaluation procedures were undertaken. This chapter aims to describe the evaluation process, starting in the data preparation until the analysis and discussion of the results. It is noteworthy that the results presented in this chapter are only concerned with the posologies of pharmacological products. Indications and adverse affects were not targeted by the evaluation due to time restrictions, even though the extraction system has been successfully applied to both these topics.

This chapter starts by describing the evaluation metrics that are used, and then thoroughly describes and analyze the evaluation setup. Later, the results of the evaluation are presented and a brief discussion is also held.

### 6.1 Evaluation Metrics

When constructing an IE system proper performance measurement metrics are essential in order to evaluate the system under development, detecting and therefore fixing substitution, deletion and insertion errors. The standard metrics commonly used for evaluation and, therefore, used in the evaluation of this thesis are precision, recall and F-measure [MKS99]. In particular, precision is calculated as the ratio of the correct findings in all findings of the system (see Eq. 6.1) and deals with substitution and insertion errors. On the other hand, recall is calculated as the ratio of correct findings within the total numbers of all expected findings (Eq. 6.2) and deals with substitution and deletion errors. The F-measure is the weighted average of the precision and recall (Eq. 6.3). The most popular F-measure used, known as F1 score, represents the harmonic mean of precision and recall, and is represented in eq. 6.4.

$$precision = \frac{\#ofCorrectFindingsExtracted}{\#ofFindingsExtracted} \quad (6.1)$$

$$recall = \frac{\#ofCorrectFindingsExtracted}{\#ofAllExpectedFindings} \quad (6.2)$$

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall} \quad (6.3)$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (6.4)$$

## 6.2 Evaluation Setup

The main goal of PharmInx is the identification of both categories and relations of semantic entities in the text, and typically this kind of task relies on human annotated documents to provide material to support development and to provide evaluation standards.

Since the corpus used was the pharmacological therapeutic records (see 3.1.2), there were exactly 1084 pharmacological products with information regarding the posology available. However, only 999 had the posology information in a text format, while the others had only a reference to another pharmacological product with a similar posology. Considering this large amount of data and the expense of the human annotation process, the data duly annotated and used for evaluation had to be a relatively small portion of the whole corpus.

Given so, the set of pharmacological products corpus was divided in two main subsets, the *Development Set* and the *Test Set*. The *Development Set* was used in the definition, development and refinement of the system, whereas the *Test Set* was used in the evaluation of the performance of the system. The *Test Set* consisted in approximately 3.5% of the whole corpus, which consists in exactly 35 documents, while the remaining 964 documents composed the *Development Set*.

In an initial phase of development, 14 documents from the *Development Set*, with a total of 43 posologies of pharmacological products, were randomly chosen and duly annotated in order to provide consistent data for the purpose of performing unit tests. These tests were an important tool and were able to ease and accelerate the development of the system.

Table 6.1: Number of documents, characters, tokens and sentences in the PharmInx corpus and its subsets

	<b>Development Set</b>	<b>Test Set</b>	<b>Total</b>
#Documents (#Drugs)	964	35	999
#Characters	184 765	8789	193 554
#Characters/Documents	192	251	194
#Tokens	45 218	2466	47 684
#Tokens/Documents	47	70	48
#Sentences	2071	80	2151
#Sentences/Documents	2	2	2

Table 6.1 presents some statistic about the PharmInx corpus and its subsets when dealing only with the extraction of posologies. Through this statistics, we can see that the average number

of tokens in the *Test Set* is relatively higher than the numbers presented in the *Development Set* and in the overall. This occurrence happens since in the selection process of the pharmacological products for manual annotation, preference was made to longer texts that would therefore include more content and would probably present a greater challenge for the system.

### 6.2.1 Annotation Process

In order to support the annotation process an annotation schema was developed. This annotation schema was similar to the knowledge representation model of PharmInx, although some simplifications were needed to allow an easier annotation process. Thus, the annotation schema included the main entities aimed by PharmInx (*e.g.*, frequencies, durations, dosages) and the relations between them.

The annotation schema and the consequent annotation process were performed resorting to Knowtator (see 3.2.4), which is a Protégé plug-in. In Figure 6.1 we can see the Knowtator interface, which presents in the left side the annotation schema developed, whereas the center shows the sample of text being annotated and the right side is where the appropriate relations and attributes of each annotation are inserted.

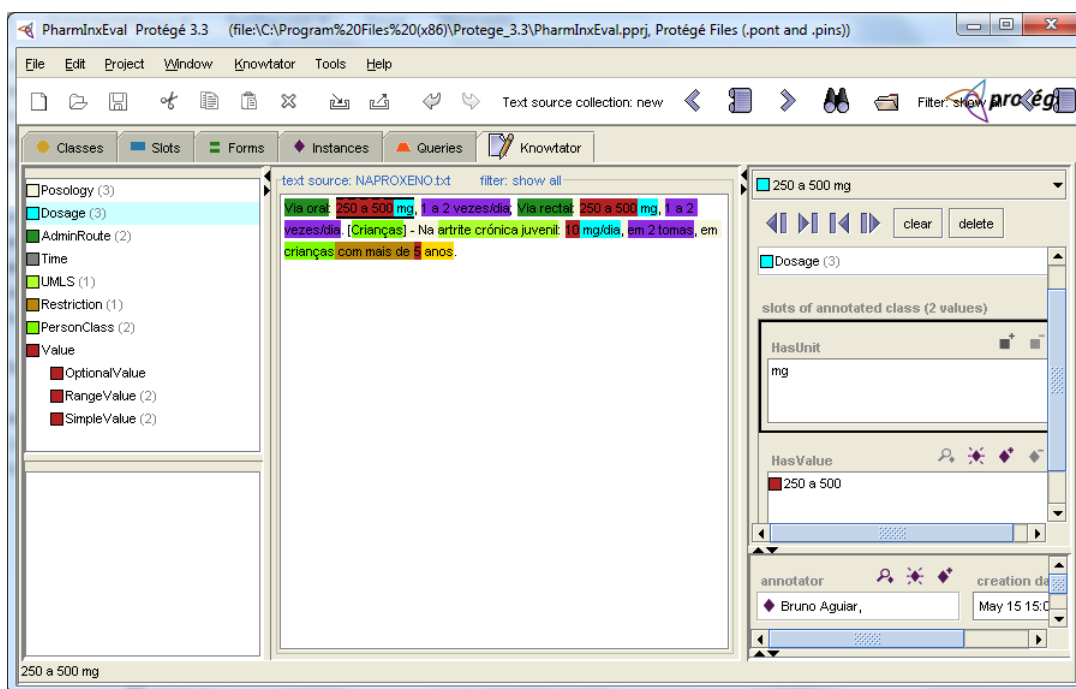


Figure 6.1: Knowtator layout when annotating the Test Set

After the annotation process took place, the annotation data was exported to XML files using Knowtator functionalities. Later, a module for parsing and processing the content of such XML files was developed and added to PharmInx components, so that PharmInx tester information consumer would be able to load the data from the XML files and process the comparisons with the results from the extraction process.

In order to properly perform the comparison between the automatically extracted and the manually annotated entities, a specific and complex module was developed. Iterating through all the manually annotated entities, a search was made in the extracted entities in order to find an entity with the same span of text, being therefore considered the same slots. If no match was found, a deletion error would be processed. If a match was found, a correct extraction would be processed if there were a positive match of the content, and a substitution error otherwise. The extracted entities that were never matched were then processed as insertion errors. A similar process is then held for the processing of the comparisons between relationships.

## 6.2.2 Test Set Characterization

Considering the 35 documents that compose the *Test Set*, we aim in this section to present some statistics about its content.

Thus, some descriptive statistics regarding the number of entities extracted from the *Test Set* by the PharmInx system are presented in Table 6.2. This table presents the total number of entities extracted in each category (N), the average (Mean), the standard deviation (Std. Dev.), the minimum (Min) and the maximum (Max) number of entities identified in the documents.

Table 6.2: Detailed statistics regarding the amount of entities extracted from the *Test Set*

Category	N	Mean	Std. Dev.	Min	Max
Administration Route	84	2	2	0	8
Duration	64	2	2	0	7
Frequency	<b>151</b>	<b>4</b>	<b>3</b>	<b>1</b>	12
Person Class	44	1	1	0	4
Posology	130	3	2	<b>1</b>	11
Simple Value	140	<b>4</b>	<b>3</b>	0	<b>15</b>
Optional Value	2	0	0	0	1
Range Value	68	2	2	0	9
Time Reference	7	0	1	0	2
UMLS	79	2	2	0	8
Restriction	57	2	2	0	10
Dosage	146	<b>4</b>	<b>3</b>	<b>1</b>	<b>15</b>

Analyzing the data presented in the table, we can see that the most frequent categories are *Frequency*, *Dosage* and *SimpleValue*, closely followed by the *Posology* category. These results could be predicted, since simple values are always common and considering that each document should have at least one posology and that each posology must have one dosage and one or more frequencies. Therefore, these are the only three categories that are always present in any of the documents tested, since all documents always contain at least a posology and consequently one dosage and at least one frequency.

Figure 6.2 presents a box plot chart with the number of entities found for each of the different categories, allowing a better analysis and understanding of the distribution of the entities



through the documents. We can see, for instance, that the more frequent number of posologies per document is 3, as well as for the dosages and frequencies.

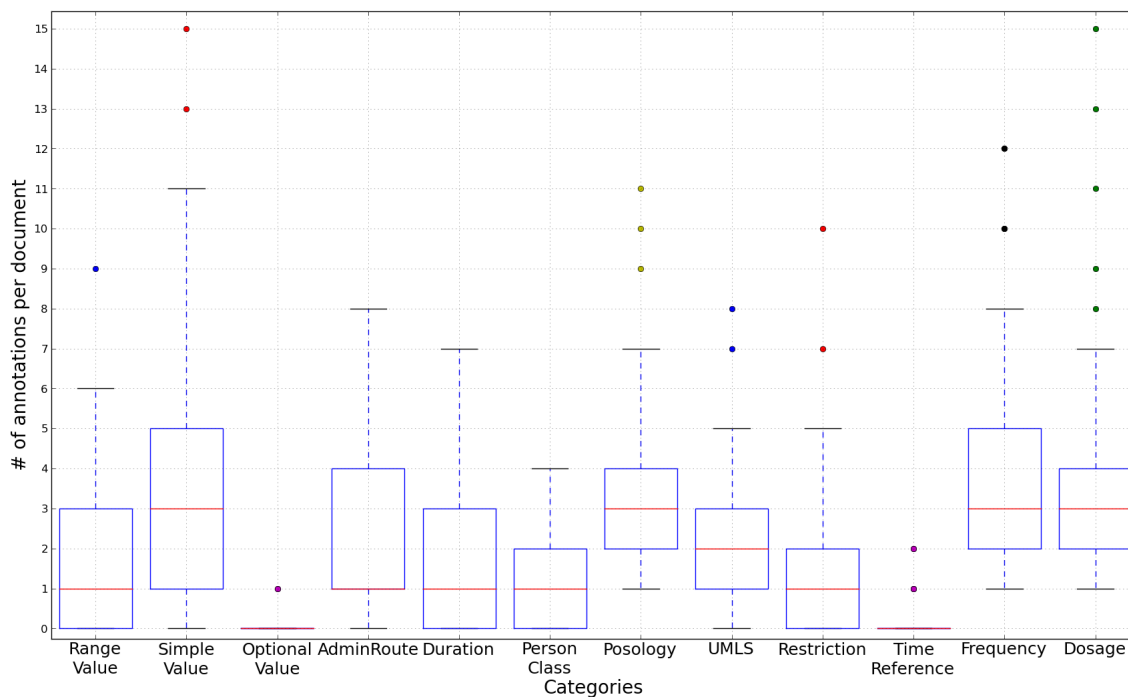


Figure 6.2: Distribution of the entities found by PharmInx according to their categories

A similar analysis can be performed regarding the relationships identified by the system when applied over the *Test Set*. Table 6.3 presents the number of relations extracted from the *Test Set* documents, showing that the most common relations are *hasValue*, *hasFrequency*, *hasAdminRoute* and *hasDosage*. The frequency of the *hasValue* relationship can easily be justified, since all *Dosage* and *Duration* annotations are guaranteed to have an *hasValue* relationship. Since the number of *Dosage* and *Duration* annotations in the *Test Set* is of 146 and 64 respectively, the 210 *hasValue* relationships are duly justified. Regarding the *hasDosage* and *hasFrequency* relationships, since there cannot be a *Posology* without these two relationships, these are frequent and as expected are never null in any document. Furthermore, considering that a *Posology* cannot have more than one *hasDosage* relationship but can have more than one *hasFrequency* relationship, it is expected that the number of *hasFrequency* relationships would be slightly higher.

Figure 6.3 presents a box plot chart of the number of relations extracted from the *Test Set*, allowing therefore an easier and more detailed analysis of their distribution. We can understand from the chart that the only relations that always appear in any document are the *hasDosage* and *hasFrequency*, and that the *hasFrequency* is in average more frequent. It is also interesting to see that restrictions have more commonly right operands than left operands.

Table 6.3: Detailed statistics regarding the amount of relations extracted from the *Test Set*

Relation	N	Mean	Std. Dev.	Min	Max
hasAdminRoute	132	4	3	0	12
hasDuration	30	1	1	0	4
hasFrequency	150	4	3	1	12
hasPersonClass	101	3	3	0	11
hasValue	<b>210</b>	<b>6</b>	<b>4</b>	<b>1</b>	<b>22</b>
hasTarget	59	2	2	0	9
involvesTherapeutic	9	0	1	0	3
hasRestriction	64	2	3	0	12
hasLeftOperand	15	0	1	0	3
hasRightOperand	56	2	2	0	10
hasDosage	126	4	3	1	11

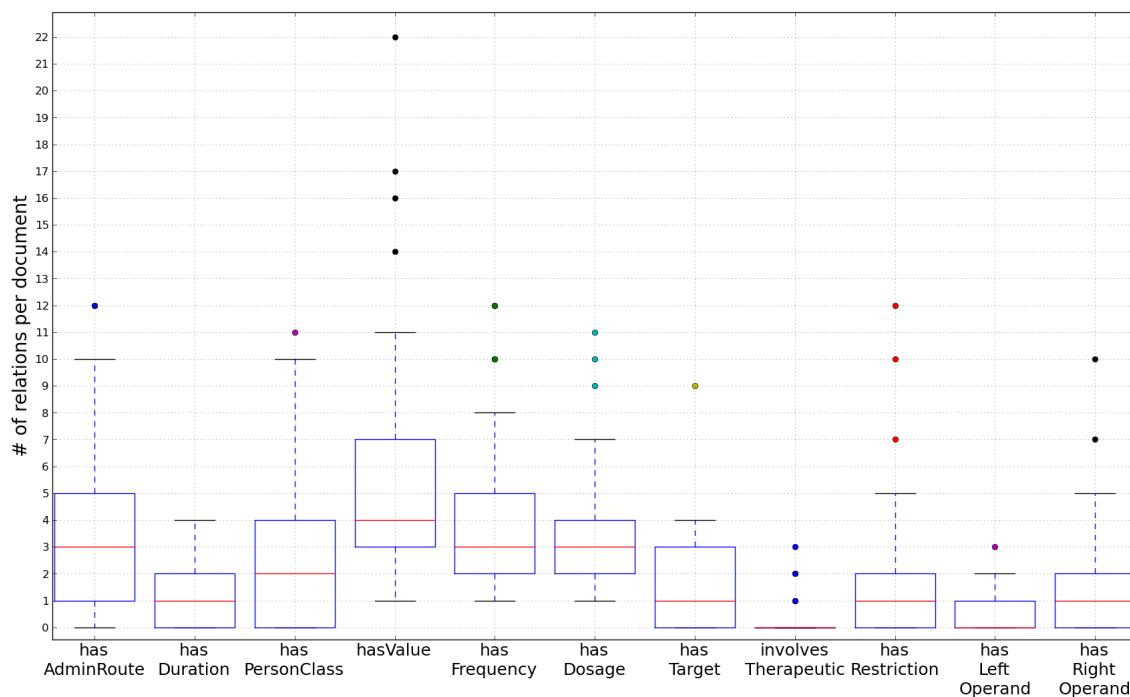


Figure 6.3: Distribution of the relations found by PharmInx according to their relations

## 6.3 Evaluation Results

After the test set was entirely annotated, we could then proceed to the comparison of the results of PharmInx with the manually annotated data. Several evaluation procedure were performed in order to comprehend the quality of the results of the developed system.

### 6.3.1 Identified Mismatches

Considering the evaluation metrics previously described, we aimed to properly identify and categorize 3 types of errors: *substitution*, *insertion* and *deletion* errors. Descriptive statistics regarding the errors found in the NER task are presented in Table 6.4. This table presents the total number of entities, both extracted by the IE system (M) and loaded from the manually annotated data (N), the number of correct extractions and the number of substitution, insertion and deletion errors.

Table 6.4: Detailed statistics regarding the errors detected when comparing the entities extracted with the manually annotated entities

Category	N	M	Correct	Substitution	Insertion	Deletion
Administration Route	84	84	84	0	0	0
Duration	64	64	64	0	0	0
Frequency	151	151	151	0	0	0
Person Class	44	44	44	0	0	0
Simple Value	140	140	140	0	0	0
Optional Value	2	2	2	0	0	0
Range Value	68	68	68	0	0	0
Time Reference	7	7	7	0	0	0
UMLS	80	79	74	4	1	2
Restriction	57	57	56	0	1	1
Dosage	146	146	146	0	0	0
<b>Total</b>	<b>843</b>	<b>842</b>	<b>836</b>	<b>4</b>	<b>2</b>	<b>3</b>

Analyzing the table, we can immediately understand that the systems shows a very good performance, being able to correctly extract 836 entities among the 841 loaded from the manually annotated data. The most common errors consist in substitution errors, 4 errors, and we can also see that deletion errors are more frequent than insertion errors.

A similar comparison can be performed for the relations extraction task, comparing therefore the relations extracted by the system with the relations established in the data manually annotated. The results of such a comparison are presented in Table 6.5.

### 6.3.2 Precision, Recall and F-measure

The overall results of PharmInx in terms of precision, recall and F1-score are presented in Table 6.6. The results are divided according to the two main tasks of an extraction system, the named entity recognition and the relation extraction.

The results presented indicate an high performance of the system in both tasks. For the NER task, all the evaluation metrics presented high results, in all cases above 99%. There is also a slight difference between the precision and recall in the NER task, favoring the precision. Figure 6.4 presents the distribution of the results, precision, recall and F1-Score, of the system when computed according to the evaluation of each of the documents of the *Test Set*. Once again, the figure shows the similarity between the evaluation metrics, being only noteworthy the difference

Table 6.5: Detailed statistics regarding the amount of relations extracted from the *Test Set*

Relation	N	M	Correct	Substitution	Insertion	Deletion
hasAdminRoute	132	126	125	1	0	6
hasDuration	28	30	28	0	2	0
hasFrequency	149	150	148	1	1	0
hasPersonClass	98	95	91	1	3	6
hasValue	210	210	210	0	0	0
hasTarget	52	56	49	0	7	3
involvesTherapeutic	8	9	8	0	1	0
hasRestriction	61	61	57	1	3	3
hasLeftOperand	15	15	15	0	0	0
hasRightOperand	56	55	55	0	0	1
hasDosage	127	127	125	2	0	0
<b>Total</b>	<b>936</b>	<b>934</b>	<b>911</b>	<b>6</b>	<b>17</b>	<b>19</b>

Table 6.6: Overall results for the NER and the RE tasks

IE Task	F1-Score	Precision	Recall
Named Entity Recognition	99.23%	99.29%	99.17%
Relation Extraction	97.43%	97.54%	97.33%

between the outliers, which are the cause of diminution of the results. It is also noticeable that the most frequent value for all the evaluation metrics is 100% and the worst result for *F1-Score* is 88.89%, indicating once again the high performance of the system.

Regarding the RE task, although the evaluation results present lower values than the NER task, the results are very positive. The *F1-Score* almost achieves the 97.5%, being that the precision is higher than that and the recall slightly lower. Figure 6.5 presents a box plot chart with the results of the evaluation metrics through all the documents of the *Test Set*. Just as the analysis on the NER evaluation, we can see that the most common evaluation results is 100%. Although in the RE task the evaluation presents an higher lower quarter, the existence of more relevant outliers, F1-Score worse result of 72.22%, leads to worse results when comparing with NER.

### 6.3.3 Results for each Category

In order to allow a more detailed analysis on the evaluation results, Figure 6.6 presents the F1-Score for each of the different categories present in PharmInx. We can now see the differences on the performance of the extraction according to the category.

As we can see on the figure, the category with worse results is the UMLS, around 93%, followed by the Restriction category, around 98%. The difficulty behind the extraction of UMLS terms is evident, due to the complexity of the medical terms present in the texts and the heavy dependency of the system in external resources. Restrictions also present a lower performance since there are many different ways to express restrictions, being difficult to cover all possibilities.

## Performance and Evaluation

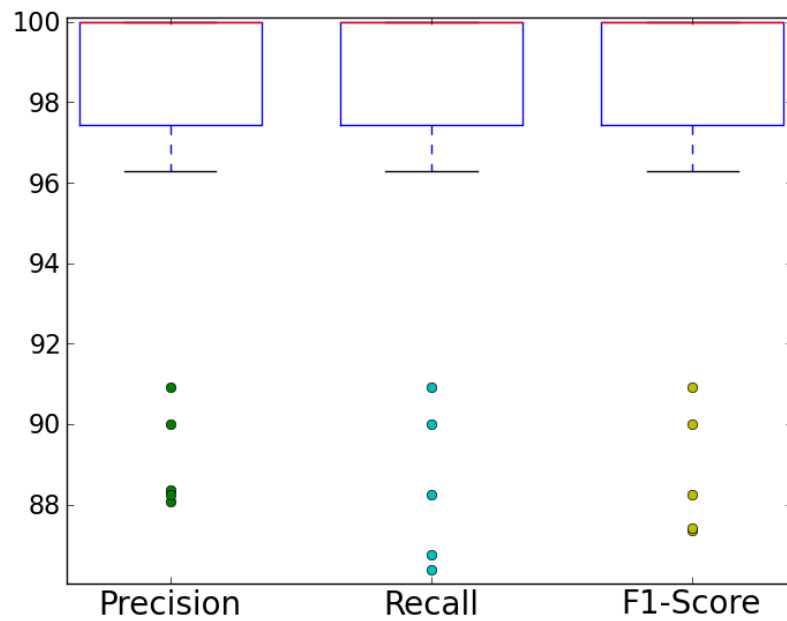


Figure 6.4: Precision, Recall and F1-Score of the system for the NER task

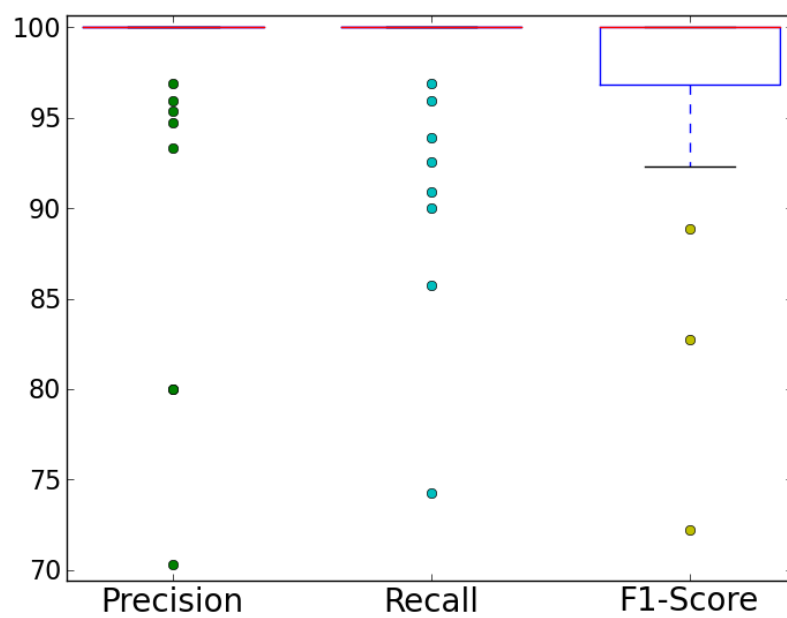


Figure 6.5: Precision, Recall and F1-Score of the system for the RE task

Disregarding these two categories, all the other present an 100% evaluation, indicating that they

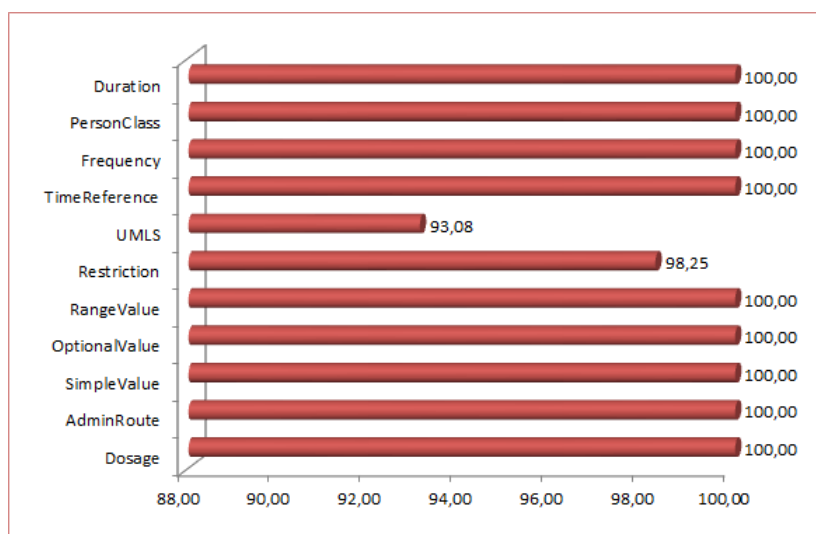


Figure 6.6: F1-Score for each of the categories present in PharmInx

are always correctly identified.

### 6.3.3.1 Results for Posology

Although the main evaluation metrics allowed us to have an appropriate overview of the performance of the system developed, considering that we aimed to extract the posologies of the pharmacological products, it would be to the best of interests to analyze the performance of such an output. The evaluation of the *Posology* category is presented separately since it is a category entirely composed with relationships. The errors found in this category are presented in Table 6.7.

Table 6.7: Detailed statistics regarding the errors detected in the Posology category

Category	N	M	Correct	Substitution	Insertion	Deletion
Posology	131	130	117	11	2	3

The *Posology* includes many relationships, as the *hasAdminRoute*, *hasPersonClass*, *hasDosage*, *hasFrequency*, *hasDuration*, *involvesTherapeutic* and *hasTarget*. An error in any of these relationships would make the posology invalid and therefore be accounted as an error. Therefore, it is expected that the *Posology* would present lower results than the general results presented in the NER and RE task.

Considering the values on Table 6.7, we can calculate the precision of the *Posology* to be around **90%** and the recall close to **89%**. As expected, these values are lower than the ones present in the NER or even the RE task (see 6.3.4), however, these results still present an high performance on the main task here targeted, the extraction of complete posologies from natural language text about pharmacological products.

### 6.3.4 Results for each Relation

Relationships can also be better analyzed if the evaluation for each one of them is presented individually. Therefore, Figure 6.7 presents such an evaluation, with each relations and the respective *F1-Score* calculated.

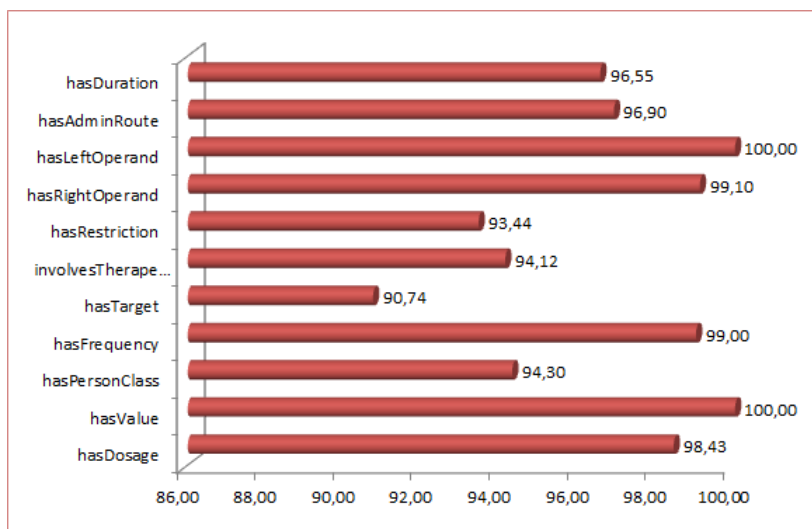


Figure 6.7: F1-Score for each of the relations present in PharmInx

In this case, only the *hasValue* and *hasLeftOperand* relationships present an 100% score, being that the other results are scattered between 90% and 99%. The *hasTarget* evaluation presents the lower value. More than half the relationships present an evaluation score below the overall evaluation (97.43%), however, the most frequent relationships are above the overall value.

## 6.4 Discussion

PharmInx evaluation was performed resorting to a set of posology documents, which were annotated by the author through human manual inspection. However, manual inspection of this sort of documents is an extremely demanding task, which together with the lack of human resource for the task was the main cause of the relatively short set of documents manually annotated and integrated into the *Test Set*.

The general results, using the Precision, Recall and F1-Score, indicate a very good performance of the system, which in the NER task presents results above 99% and in the RE task around 97.5%. Although the comparison with similar systems developed for other languages and naturally working with other corpus is not straightforward, PharmInx presents itself clearly at the level of international state of the art systems. Gold *et al* presented a system with an overall precision of 94.1% and recall 82.5%, while Xu *et al* presented F1-Score results between 93.2 and 96%. Hereupon, we can see that two of the most relevant studies regarding the extraction of posologies presented slightly lower results than the ones presented by PharmInx, allowing us to position

PharmInx as one of the best medication information extraction systems, with more emphasis in posology extraction.

Considering that PharmInx aims to output an ontology, providing machine understandable information for other applications, both the precision and recall were constantly a major concern. The difference between the precision values and the recall values are not noteworthy, although these are in all results slightly above the recall values. Without high precision a system is incapable of providing a correct, consistent and concise ontology for further analysis. However, high recall values are important to ensure the completeness of the ontology. Hereupon, the values of the results obtained by PharmInx allow us to ensure that the ontology follows the standard principles in its creation.

When analyzing the results obtained for each of the categories, the *UMLS* category highlights due to a low result compared with the other categories, being the category that presents worst results. The UMLS Metathesaurus has been used to support the extraction of this kind of entities, and although the UMLS presents a great amount of information regarding several different categories (e.g. diseases, procedures, anatomical sites), some of these subcategories contain a more reduced and disperse information, hindering the extraction process. Hereupon, the utilization of more knowledge sources developed specifically for the Portuguese language would certainly present advantages.

In the relation extraction task the system presented excellent results, justifying the complex relation extraction component developed. However, it is important to note that not all kind of posologies are currently supported by the system, neither by the knowledge representation or the relation extraction component. Such exceptions are the existence of more complex posologies, where an initial dosage is prescribed and then an increment is added according to a prescribed frequency, or the existence of posologies where different pharmacological products should be taken together with a mixed dosage. Although these specific cases are not currently supported, the flexibility of the system and of the knowledge representation will surely ease the implementation of this upgrades.

## 6.5 Summary

This chapter presented the evaluation performed over the PharmInx system, beginning with the identification of the evaluation metrics to be used and with a detailed description of the evaluation setup, the annotation process and the methodologies used. Later, the results of the system were presented, indicating that the system is very precise, reaching 99.29% and 97.54% of precision in the NER and RE tasks respectively. The recall of the system was also very high, achieving 99.17% and 97.33% for the same tasks. Besides these usual evaluation metrics, this chapter presented the precision and recall of the main target of the extraction, the posology, which is a set of complex relationships. These results presented 89.5% F1-Score, indicating that the posologies extracted by the system are in fact reliable.



## Chapter 7

# Conclusions

Medical electronic systems, mainly personal wellbeing applications and decision support systems, are becoming more frequent worldwide and are expected to improve health care quality through improved information access. However, unstructured texts remain a common source of information, restricting the information usable by these systems. Hereupon, the need for systems capable of automatically extract information from narrative data has been growing.

PharmInX, the information extraction system presented in this document, is the first information extraction system that aims to extract information from medication leaflets. The development of this system complies with the growing need of generating structured information from narrative data, in order to support medical electronic systems.

Considering the initial goals proposed (see 1.2), we are now able to perform a balance with the goals accomplished. Since the system here described is able to automatically extract pharmacological information from natural language text in Portuguese, and afterwards save all that information in a machine understandable format, we can claim that the main goals were accomplished. However, the system is not able to extract all kinds of pharmacological information, being currently restricted to information regarding posologies, indications and adverse reactions, which were the main goals of this thesis. The extraction of drug-drug interactions and of counter-indications, which were initially considered as secondary goals, was not accomplished. Moreover, the extraction of indications and adverse reactions were not properly evaluated, being therefore not proved the effectiveness of the system with these topics.

The evaluation performed on the posology topic, which included 130 manually annotated posologies, presented extremely good results, with an F1-Score of 99.23% on the recognition of entities and 97.43% on the extraction of relationships, giving good indications that the system would present a good performance with the other topics.

In conclusion, the development of a system with an entire work-flow, including the extraction of natural language text and consequent persistence of machine understandable data, was executed.

Furthermore, the information topics more important for our target group, the patients, are properly extracted by the system developed, showing an high performance.

### 7.1 Future Work

There are numerous opportunities for both improving and extending this work. Regarding improvements for this work, one of the most important improvements would be the development of development and test sets for indications and adverse reactions, being therefore able to enhance and then evaluate the performance of the extraction task in these topics. Some other improvements could be done on the extraction process as using new external knowledge sources and creating more rules in order to cover more possibilities currently missed. Another improvement to be considered would be the use of Lucene<sup>1</sup> capabilities to index the great amount of data from the UMLS external resource, probably making the system considerably faster.

As for possible extensions of this work, the most obvious and urgent work to be done would be the extension of the extraction task to the drug-drug interactions and counter-indications topics. If the results for these topics remain considerably good, an attempt to extract information in more general clinical texts could be performed. We can also consider that some of the components of the system can be used individually to enhance other systems as, for example, the *UMLS Annotator*. Since it is capable of annotating any kind of medical term, would certainly be useful for identifying and categorizing disorders, procedures and anatomical sites in other clinical documents, such as patient discharge letters, clinical journals, reports of anatomical pathology and even surgical records.

The application of machine learning approaches for the NER and RE task could also be interesting and could possibly enhance the results. Although, for such approaches to be successful, more data would probably have to be annotated.

Besides the improvements and extensions of the current system, there are other short-term tasks to be performed, as for instance the integration of the results outputted by the system in an health-care application.

---

<sup>1</sup><http://lucene.apache.org/core/>

## **Appendix A**

# **Submitted Papers**

### **A.1 StudECE**

The 1st PhD. Students Conference in Electrical and Computer Engineering (StudECE) is a forum for the presentation of technological advances, research results, work-in-progress research and state-of-the-art in the fields of theoretical, experimental, and applied Electrical and Computer Engineering. The following paper is a brief description of the work performed, including the general architecture of the system and a description of its several different components, and also including the results presented by the system developed. This paper was successfully accepted as a contribution for the conference proceedings.

# Information Extraction from Medication Leaflets

Bruno Aguiar<sup>\*†</sup>, Eduarda Mendes<sup>\*</sup>, Liliana Ferreira<sup>†</sup>,

<sup>\*</sup>FEUP - Faculty of Engineering,

University of Porto

{bruno.aguiar, eduarda}@fe.up.pt

<sup>†</sup>Fraunhofer Portugal AICOS

{bruno.aguiar, liliana.ferreira}@fraunhofer.pt

**Abstract**—With the constant growth of medical electronic systems, including decisions support systems and personal-care applications, the need for machine understandable information has been growing. However, much of the data currently available is in free-form text, being therefore necessary the use of Information Extraction (IE) systems. The IE system here proposed aims to automatically extract information from pharmacological texts, more precisely medication leaflets. After the extraction process, we aim to provide high-quality and machine understandable information, which is currently not available for medical electronic systems, and thus help improving personal-care services for patients, and enhancing decision support systems for health-care professionals. The results achieved by the IE system, which is still under development, indicate that pharmacological and clinical information can successfully be extracted from free-form texts in Portuguese, presenting a F1 score of 99.23% when recognizing entities and a F1 score of 97.43% when extracting relations between those entities.

**Index Terms**—Information Extraction; Natural Language Processing; Medical Language Processing; Ontology; UIMA; Pharmacological products; Drugs

## I. INTRODUCTION

This paper aims to describe an information extraction system developed for extracting specific information from the Portuguese medical therapeutic records [1], which contain different kinds of information about all the pharmacological drugs currently in use in Portugal. Among the existent information about pharmacological products, we aim mainly to extract information regarding posology, side effects, indications, interactions and precautions. All this information is currently in natural language text, more specifically in Portuguese, so in order to provide concrete and valuable information for computer applications we need to be able to structure the information intended. Hereupon, we propose an information extraction system capable of outputting structured information about pharmacological drugs currently in use in Portugal, relying on the natural language records available.

Medication information is considered to be one of the most important types of clinical data in electronic medical records. However since such data is commonly recorded in free-text format many authors have been focused on developing systems capable of producing machine understandable data from clinical narratives in the last few years.

A recent study by Gold et al [2] reported a regular expression based approach for extracting drug names and signature information such as dose, administration route and frequency.

Evaluation on a data set of 26 discharge summaries presented a precision of 94.1% and a recall of 82.5% for drug names identification, although other signature information such as dose and frequency had much lower precisions. A later study, based on a similar approach but using a set of manually defined rules for disambiguation [3] achieved F-measures around 93.2% and 96.0% for similar signature information about drugs.

In order to support the development of the information system intended, we resorted to an open, industrial-strength, scalable, and extensible framework, the Unstructured Information Management Architecture (UIMA) [4].

## II. SYSTEM ARCHITECTURE

The main goal of this specific IE system is the extraction of several different entities and their respective relationships from pharmacological free-form text written in Portuguese. Hereupon, we aim to obtain information such as the dosage, duration and frequency of administration, the administration routes, for who the medication is recommended (e.g. adults or elderly), for which purposes is the medication suggested, the adverse reactions and the precautions inherent with the use of the medication.

Given so, this system components are based on NLP principles and provide several mechanisms to read, process and employ external resources, such as ontologies and terminologies. Considering that typically the analysis of natural language documents is a complex task and therefore not carried out in a single step, we divided the processing of this system in 6 main different steps:

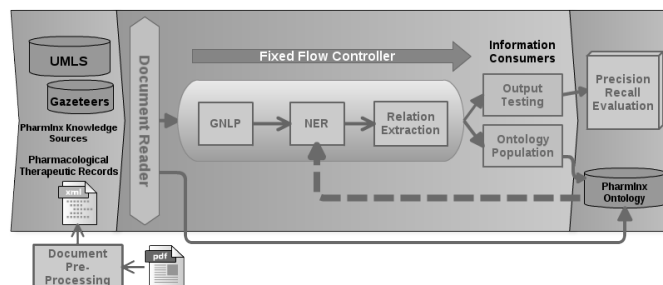


Figure 1. System Architecture

- 1) **Document Pre-Processing:** Since the documents where the extraction task should be held are in a PDF format, a pre-processing task was conducted in order to generate an XML file where the tags added represent the implicit meaning regarding the document structure. The information implicit in the structure of the document is atypically important for this system, mainly due to the vast number of elliptical sentences.
- 2) **Document Reader:** This component should be able to read the XML input file and load its content and respective embedded tags into annotations for further analysis. Another task of this component is to add in beforehand all the pharmacological products in the target document to the ontology, in order to be able to recognize these products when referenced in the text.
- 3) **General Natural Language Processing:** This component includes a sentences discoverer, a tokenizer, an abbreviations expander, a stop-words identifier, an Hidden Markov Model part-of-speech tagger and finally a snowball stemmer.
- 4) **Named Entity Recognition:** This component is able to construct annotations that identify entities in the text, using for that purpose several different annotators which use regular expressions, rules, external resources for lookups and using the contextual surrounding information. The most important external resource is the UMLS [7], which consists in a synopsis of the many controlled vocabularies in biomedical science, providing a mapping structure among the vocabularies. Therefore, this resource is the key for the identification of clinical terms in the text. It is noteworthy that one of the resources used in this component is the ontology generated by the system, thus enabling a faster processing of entities previously encountered in the text.
- 5) **Relation Extraction:** Considering the entities previously identified, this component constructs a tree with all the entities according to their order and scope of influence on the text. Once the tree is finished, all relationships are then easily inferred.
- 6) **Information consumers:** Responsible for ending the process and creating the desired output, using the previous analysis and results. There can be several different information consumers in an IE system, being that in this particular IE system the two main information consumers are:
  - **Ontology Population**, outputs the information extracted by populating an ontology using Jena API [8], generating a valid OWL file.
  - **Output Testing**, that is responsible for comparing the outputted data with previously and manually annotated information, presenting then the precision, recall and F1-score of the system.

All the components run within the UIMA framework, using therefore its type system definition to define all the kind of meta-data that can be recognized. Currently, the meta-

data recognized includes several types of values, dosages, administration routes, restrictions, UMLS records, descriptive characteristics and finally several time references such as frequencies and durations.

### III. RESULTS AND EVALUATION

In order to properly evaluate the system, 35 pharmacological products among different pharmacological groups were randomly chosen for manual annotation, providing a total of 130 different posologies. When the results of the system are compared with the data manually annotated by the first author, a F1 score of 99.23% is achieved for the named entity recognition task, presenting a Precision of 99.29% and a Recall of 99.17%. Regarding the relation extraction task, a F1 score of 97.43% is presented, with a precision of 97.54% and a recall of 97.33%.

Currently, this system has only been applied and tested to the posology information of pharmacological products, however, the results obtained are promising not only for this topic but also to the others, as indications, adverse reactions, counter-indications and even interactions.

### IV. CONCLUSIONS AND FUTURE WORK

A system capable of identifying relevant entities in pharmacological texts and capable of automatically populating a pharmacological ontology, is presented. The several steps of reasoning of the system are identified and briefly described.

The results presented by the system and the flexibility of the architecture described give good indications that this system will show a good performance when faced with pharmacological information that not the posology or even with more general medical text.

Hereupon, it is expected that at short-term the system will be enhanced and then applied to other pharmacological texts and possibly medical narratives. The possible uses of the information extracted are vast, being that the improvement of personal-care services for patients and the enhancement of decision support systems for health-care professionals are the most immediate applications of the structured information outputted by the system.

### REFERENCES

- [1] *Prontuario Terapeutico Online*. <http://m.infarmed.pt/Prontuario/Home.aspx>, Last checked: 30/04/2012.
- [2] Sigfried Gold, Nomie Elhadad, Xinxin Zhu, James J Cimino, and George Hripcsak. Extracting structured medication event information from discharge summaries. AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium, 2008:237241.
- [3] Hua Xu, Shane P. Stenner, Son Doan, Kevin B. Johnson, Lemuel R. Waitman, and Joshua C. Denny. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association, 17(1):1924, January 2010.
- [4] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng., 10:327348, September 2004.
- [5] Hidden Markov Model Tagger Annotator. <http://uima.apache.org/sandbox.html>, Last checked: 30/04/2012
- [6] *Portuguese stemming algorithm*. <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>, Last checked: 30/04/2012.
- [7] United States National Library of Medicine, *Umls knowledge sources*, 2008
- [8] *Apache Jena*. <http://incubator.apache.org/jena>, Last checked: 30/04/2012.

## Submitted Papers

# References

- [ABD<sup>+</sup>95] John Aberdeen, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain. Mitre: description of the alembic system used for muc-6. In *Proceedings of the 6th conference on Message understanding*, MUC6 '95, pages 141–155, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.
- [AM05] Sophia Ananiadou and John Mcnaught. *Text Mining for Biology And Biomedicine*. Artech House, Inc., Norwood, MA, USA, 2005.
- [AMT<sup>+</sup>10] Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kayo Waki, and Kazuhiko Ohe. Extraction of adverse drug effects from clinical records. *Studies in health technology and informatics*, 160(Pt 1):739–743, 2010.
- [AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.
- [BAPS11] Jari Björne, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. Drug-drug interaction extraction with SVM and RLS classifiers. In Isabel Segura-Bedmar, Paloma Martinez, and Daniel Sanchez-Cisneros, editors, *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011*, volume 761 of *CEUR Workshop Proceedings*. CEUR-WS.org, September 2011.
- [BLCL<sup>+</sup>94] T. Berners-Lee, R. Cailliau, A. Luotonen, H.F. Nielsen, and A. Secret. The World-Wide Web. *Communications of the ACM*, 37:76–82, 1994.
- [BLF99] Tim Berners-Lee and Mark Fischetti. *Weaving the web: The original design and ultimate destiny of the world wide web by its inventor*. Harper, San Francisco, 1999.
- [BQ10] Luis Borrego and Paulo Quaresma. Criação de uma ontologia e respectiva povoação a partir do processamento de relatórios médicos. In *Jornadas de Informática da Universidade de Évora*, 2010.
- [CDIW03] Fabio Ciravegna, Alexiei Dingli, José Iria, and Yorick Wilks. Multi-strategy definition of annotation services in melita. In *Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd Int Semantic Web Conf. (ISWC)*, Florida, 2003.

## REFERENCES

- [CFM07] André Coutinho Castilla, Sérgio Shiguemi Furuie, and Eneida A. Mendonça. Multilingual information retrieval in thoracic radiology: Feasibility study. In Klaus A. Kuhn, James R. Warren, and Tze-Yun Leong, editors, *MedInfo*, volume 129 of *Studies in Health Technology and Informatics*, pages 387–391. IOS Press, 2007.
- [Chi98] Nancy A. Chinchor. Proceedings of the Seventh Message Understanding Conference (MUC-7) named entity task definition. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, page 21 pages, Fairfax, VA, April 1998. version 3.5, [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/).
- [CHX<sup>+</sup>08] E. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman. Automated acquisition of disease drug knowledge from biomedical and clinical documents: An initial study. *Journal of the American Medical Informatics Association*, 15(1):87–98, 2008.
- [CJ01] D. A. Campbell and S. B. Johnson. Comparing syntactic complexity in medical and non-medical corpora. *Proc AMIA Symp*, pages 90–94, 2001.
- [CMBT02] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA*, 2002.
- [Cun05] H. Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics*, 2nd Edition, 2005.
- [Dom98] J. Domingue. Tadzebao and webonto: Discussing, browsing, editing ontologies on the web. In *11th Knowledge Acquisition for Knowledge-Based Systems Workshop*, 1998.
- [dSF11] Liliana da Silva Ferreira. *Medical Information Extraction in European Portuguese*. PhD thesis, Universidade de Aveiro, 2011.
- [DSM00] Robert Dale, H. L. Somers, and Hermann Moisl, editors. *Handbook of Natural Language Processing*. Marcel Dekker, Inc., New York, NY, USA, 2000.
- [EBHC96] D. Evans, N. Brownlow, W. Hersh, and E. Campbell. Automating concept identification in the electronic medical record: An experiment in extracting dosage information. In *Proceedings of the AMIA Annual Symposium*, pages 388–392, Washington DC, USA, 1996.
- [elr] European language resources association. <http://www.elra.info> Last checked: 07.02.2012.
- [FFR96] Adam Farquhar, Richard Fikes, and James Rice. The ontolingua server: a tool for collaborative ontology construction. In *International Journal of Human-Computer Studies*, 1996.
- [FL04] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10:327–348, September 2004.



## REFERENCES

- [FS06] Ronen Feldman and James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, Cambridge, MA, USA, December 2006.
- [Gam] Pablo Gamallo. Treetagger para o português. <http://gramatica.usc.es/gamallo/tagger.htm> Last checked: 07.02.2012.
- [GBMVDR11] Sandra Garcia-Blasco, Santiago M. Mola-Velasco, Roxana Danger, and Paolo Rosso. Automatic Drug-Drug Interaction Detection: A Machine Learning Approach with Maximal Frequent Sequence Extraction. pages 51–58, September 2011.
- [GEZ<sup>+</sup>] Sigfried Gold, Noémie Elhadad, Xinxin Zhu, James J Cimino, and George Hripisak. Extracting structured medication event information from discharge summaries. *AMIA Annual Symposium proceedings AMIA Symposium*, 2008:237–241.
- [GLR06] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics EACL2006*, 18:401–408, 2006.
- [GMF<sup>+</sup>03] John H. Gennari, Mark A. Musen, Ray W. Ferguson, William E. Grosso, Monica Crubézy, Henrik Eriksson, Natalya F. Noy, and Samson W. Tu. The evolution of protg&#233;: an environment for knowledge-based systems development. *Int. J. Hum.-Comput. Stud.*, 58:89–123, January 2003.
- [GS96] Ralph Grishman and Beth Sundheim. Message understanding conference - 6: A brief history. In *Proceedings of the International Conference on Computational Linguistics*, 1996.
- [Gue08] Guergana. UIMA-based Clinical Information Extraction System. *LREC 2008: Towards enhanced interoperability for large HLT systems: UIMA for NLP*, 2008.
- [GWJ95] Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225, 1995.
- [HBL<sup>+</sup>08] Udo Hahn, Ekaterina Buyko, Rico Landefeld, Matthias Mühlhausen, Michael Poprat, Katrin Tomanek, and Joachim Wermter. An overview of JCoRe, the JULIE lab UIMA component repository. In *LREC’08 Workshop ‘Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP’*, pages 1–7, Marrakech, Morocco, May 2008.
- [hmm] Hidden markov model tagger annotator. <http://uima.apache.org/sandbox.html#tagger.annotator> Last checked: 17.04.2012.
- [HRS] Udo Hahn, Martin Romacker, and Stefan Schulz. Creating knowledge repositories from biomedical reports: the medsyndikate text mining system. *Pacific Symposium On Biocomputing*, 2002:338–349.

## REFERENCES

- [HW04] Udo Hahn and Joachim Wermter. High-performance tagging on medical texts. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [IM] Infarmed and Portuguese Health Ministry. Prontuario terapêutico online. <http://m.infarmed.pt/Prontuario/Home.aspx> Last checked: 07.02.2012.
- [jen] Jena. <http://incubator.apache.org/jena/> Last checked: 07.02.2012.
- [KCD00] L.T. Kohn, J. Corrigan, and M.S. Donaldson. *To err is human: building a safer health system*. National Academy Press, Washington, 2000.
- [KR08] Manuela Kunze and Dietmar Rösner. Uima for nlp based researchers' workplaces in medical domains. In *Proceedings of Workshop 'UIMA for NLP' at LREC 2008*, pages 20–23. n.b., May 2008.
- [LBC<sup>+</sup>95] L L Leape, D W Bates, D J Cullen, J Cooper, H J Demonaco, T Gallivan, R Hal-lisey, J Ives, N Laird, and G Laffel. Systems analysis of adverse drug events. ade prevention study group. *JAMA*, 274(1):35–43, 1995.
- [LHC07] Kaihong Liu, Wendy Webber Chapman, Rebecca Hwa, and Rebecca S. Crowley. Methods paper: Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *JAMIA*, 14(5):641–650, 2007.
- [ldc] Linguistic data consortium. <http://www.ldc.upenn.edu> Last checked: 07.02.2012.
- [Lev66] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.
- [lin] Linguateca. <http://www.linguateca.pt/HAREM/> Last checked: 07.02.2012.
- [MGM98] Andrei Mikheev, Claire Grover, and Marc Moens. Description of the Itg system used for muc-7. In *Proceedings of 7th Message Understanding Conference (MUC-7)*, 1998.
- [MKSW99] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, pages 249–252, 1999.
- [mma] Mmax2. <http://mmax2.sourceforge.net/> Last checked: 15.06.2012.
- [NL10] Aurélie Névél and Zhiyong Lu. Automatic integration of drug indications from multiple health resources. In *Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10*, pages 666–673, New York, NY, USA, 2010. ACM.
- [nls] Natural language software registry. <http://registry.dfki.de> Last checked: 07.02.2012.
- [NM01] Natalya F. Noy and Deborah L. McGuinness. *Ontology development 101: A guide to creating your first ontology*. Technical report, 2001.

## REFERENCES

- [Ogr06] Philip V. Ogren. Knowtator: A protégé plug-in for annotated corpus construction. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics, 2006.
- [OHTT01] Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(1):155–161, 2001.
- [oM08] United States National Library of Medicine. Umls knowledge sources, 2008.
- [RF03] Thomas C. Rindflesh and Marcelo Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. of Biomedical Informatics*, 36:462–477, December 2003.
- [RN93] Matthew S. Ryan and Graham R. Nudd. The viterbi algorithm. Technical report, Coventry, UK, UK, 1993.
- [RSGQ10] Stefania Rubrichi, Alex Spengler, Patrick Gallinari, and Silvana Quaglini. Preventing adverse drug events by extracting information from drug fact sheets. In Nigel Collier, Udo Hahn, Dietrich Rebholz-Schuhmann, Fabio Rinaldi, and Sampo Pyysalo, editors, *Semantic Mining in Biomedicine*, volume 714 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.
- [SBCDPSM09] Isabel Segura-Bedmar, Mario Crespo, César De Pablo-Sánchez, and Paloma Martínez. Drugnerar : Linguistic rule-based anaphora resolver for drug-drug interaction extraction in pharmacological documents. *Corpus*, pages 19–26, 2009.
- [SBMdPS10] Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez. Combining syntactic information and domain-specific lexical patterns to extract drug-drug interactions from biomedical texts. In *Proceedings of the ACM fourth international workshop on Data and text mining in biomedical informatics*, DTMBIO ’10, pages 49–56, New York, NY, USA, 2010. ACM.
- [SBMdPS11] Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, 44(5):789–804, 2011.
- [SDM<sup>+</sup>] K G Shojania, B W Duncan, K M McDonald, R M Wachter, and A J Markowitz. Making health care safer: a critical analysis of patient safety practices. *Evidence report/technology assessment Summary*, 2001(43):i–x, 1–668.
- [SJNH] Tammy Powell Stuart J. Nelson, MD and Betsy L. Humphreys. The unified medical language system (umls) project. <http://www.nlm.nih.gov/mesh/umlsforelis.html> Last checked: 07.02.2012.
- [sno] Portuguese stemming algorithm. <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html> Last checked: 17.04.2012.
- [SPKR96] Bill Swartout, Ramesh Patil, Kevin Knight, and Tom Russ. Towards distributed use of large-scale ontologies. In *Proceedings of the 10th. Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, Canada, 1996.

## REFERENCES

- [SS04] Steffen Staab and Rudi Studer, editors. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.
- [Tab10] Valentin Tablan. *Toward Portable Information Extraction*. PhD thesis, The University of Sheffield, 2010.
- [TC10] Liliana Ferreira António Teixeira and João Paulo Silva Cunha. Ontology-driven vaccination information extraction. In *Jornadas de Informática da Universidade de Évora*, 2010.
- [uim] Apache uima. <http://uima.apache.org/> Last checked: 07.02.2012.
- [VJRH09] K. Verspoor, W. Baumgartner Jr, C. Roeder, and L. Hunter. Abstracting the Types away from a UIMA Type System. *From Form to Meaning: Processing Texts Automatically*. Tübingen:Narr, pages 249–256, 2009.
- [WD10] D C Wimalasuriya and D Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323, 2010.
- [WKG<sup>+</sup>08] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Database issue):D901–D906, 2008.
- [wor] Wordfreak. <http://wordfreak.sourceforge.net/> Last checked: 15.06.2012.
- [XSD<sup>+</sup>10] Hua Xu, Shane P. Stenner, Son Doan, Kevin B. Johnson, Lemuel R. Waitman, and Joshua C. Denny. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, January 2010.
- [ZLL03] Ai-Ling Zhu, Jian Li, and Tze-Yun Leong. Automated Knowledge Extraction for Decision Model Construction: A Data Mining Approach. In *AMIA 2003 Symposium Proceedings*, pages 758–762. AMIA, AMIA, 2003.