**Liliana
da Silva Ferreira**

**Extracção de Informação Médica em Português
Europeu**

**Medical Information Extraction in European
Portuguese**

**Liliana
da Silva Ferreira**

Extracção de Informação Médica em Português Europeu

Medical Information Extraction in European Portuguese

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Informática, realizada sob a orientação científica do Doutor António Joaquim da Silva Teixeira, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro e do Doutor João Paulo Trigueiros da Silva Cunha, Professor Associado com Agregação do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.

Apoio financeiro da FCT e do FSE no âmbito do III Quadro Comunitário de Apoio.

Ao Bernardo

o júri

presidente

Doutor Joaquim Manuel Vieira
Professor Catedrático da Universidade de Aveiro

vogais

Doutor João Paulo Trigueiros da Silva Cunha (co-orientador)
Professor Associado com Agregação da Universidade de Aveiro

Doutora Diana Sousa Marques Santos
Professora Associada da Universidade de Oslo, Noruega

Doutor Paulo Jorge de Sousa Gomes
Professor Auxiliar da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Doutor António Joaquim da Silva Teixeira (orientador)
Professor Auxiliar da Universidade de Aveiro

Doutor Alberto Manuel Brandão Simões
Professor Adjunto do Instituto Politécnico do Cávado e do Ave

agradecimentos

Ao longo deste trabalho recebi o apoio e a colaboração de diversas pessoas, às quais gostaria de transmitir o meu mais sincero agradecimento. Na impossibilidade de agradecer individualmente a todos aqueles que, de alguma forma, estiveram presentes durante o desenvolvimento do trabalho, reservo as seguintes linhas para agradecer às pessoas e instituições sem as quais este doutoramento não teria sido possível.

Reservo o meu primeiro agradecimento aos meus orientadores, Doutor António Teixeira e Doutor João Paulo Cunha. A Vossa dedicação, os amplos e multidisciplinares ensinamentos, a exigência, a confiança, o constante incentivo e o rigor científico que me transmitiram foi absolutamente determinante em todos os momentos deste processo. Não posso, no entanto, deixar de exprimir um agradecimento especial ao Doutor António Teixeira pelos, já muitos, anos de orientação, pelos inumeráveis ensinamentos, pela disponibilidade e apoio incondicional em todos os momentos da minha passagem pela Universidade de Aveiro. Estou perfeitamente consciente que só a total confiança e liberdade que depositou em mim me permitiu alcançar os sucessos académicos, profissionais, mas também pessoais e familiares que hoje vivencio.

Aos meus amigos e colegas do IEETA agradeço todo o apoio e companheirismo, desde os momentos de discussão científica aos fantásticos momentos de descontração e convívio, a paciência e a ajuda constante: Samuel, Rita, Catarina, Sara, colegas do SIAS... a todos, muito obrigada pelo fantástico ambiente de trabalho e pelos inesquecíveis momentos de experienciámos.

À instituição IEETA e aos seus representantes, devo todo o apoio material e infra-estrutural que colocaram à minha disposição. Agradeço também o ambiente, simultaneamente, descontraído e cientificamente rigoroso que transmitem aos Vossos colaboradores. À Anabela em particular, agradeço toda a disponibilidade e profissionalismo que confere à resolução dos mais variados problemas e desafios.

Ao Dr. Telmo Costa do Hospital Infante D. Pedro agradeço as valiosas conversas orientadoras na área médica, a paciência para responder às mais variadas perguntas e a gentileza que teve ao disponibilizar o seu tempo para participar nas tarefas mais cruciais de modelação, anotação e avaliação deste trabalho. Ainda neste sentido, agradeço aos pacientes revisores que cederam o seu tempo para a avaliação dos resultados do sistema. Sem a Vossa colaboração não teria sido possível realizar este trabalho.

À Fundação para a Ciência e Tecnologia agradeço o apoio financeiro que me concedeu, viabilizando, não só a realização do trabalho, mas também a minha participação em conferências, eventos científicos e a minha passagem por laboratórios internacionais de investigação, como o *Ubiquitous Knowledge Processing Lab* da Universidade Técnica de Darmstadt e o *Seminar für Sprachwissenschaft* da Universidade de Tuebingen. Estes períodos de intensa e valiosa aprendizagem constituem um particular privilégio no meu percurso profissional e académico, os quais não posso deixar de agradecer à Doutora Iryna Gurevych, ao Doutor Erhard Hinrichs, ao Niklas Jakob, e a todos aqueles que me acolheram de forma tão generosa.

Estendo ainda os meus agradecimentos à Linguateca, e em particular aos organizadores do HAREM II, evento que me permitiu aprender, explorar, aplicar e discutir alguns dos métodos que me propunha utilizar.

Aos meus mais recentes colegas da Associação Fraunhofer Portugal quero agradecer o imediato acolhimento e a forma simpática com que me permitiram integrar no ambiente do trabalho, situação que muito ajudou nos momentos finais deste trabalho.

Dedico um agradecimento muito especial à família mais próxima que esteve presente durante todo este processo. À minha mãe, agradeço ter-me permitido chegar até aqui, bem como todo o esforço e a determinação que demonstrou e demonstra em todos os momentos da vida, e a força que teve para ultrapassar e vencer os menos fáceis. Ao meu pai, agradeço ter-me transmitido a paixão pela ciência, pela engenharia, pela matemática em particular, e pela descoberta em geral. As nossas longas conversas e desafios, desde a mais tenra idade, ensinaram-me o pensamento lógico e pragmático que direcciona todos os campos da minha vida. Ao Klaus não posso deixar de agradecer na língua que adoptámos (e adaptamos) como nossa. *To Klaus, who truly knows me, with all my temper, crazy working hours, and still give me unconditional love and care, I thank all the patience and understanding.* Por último, quero agradecer ao meu filho Bernardo, por me ter ensinado uma nova e poderosa definição de amor incondicional, e por ter tornado os últimos meses deste trabalho na recordação mais bonita de sempre!

palavras-chave

Tecnologias da Linguagem Humana, Processamento de Linguagem Médica, Extração de Informação, Representação de Conhecimento Médico, Ontologias Médicas, Codificação Automática, Sistemas de Apoio à Decisão.

resumo

O armazenamento electrónico dos dados médicos do paciente é uma prática cada vez mais comum nos hospitais e clínicas médicas de todo o mundo. No entanto, a maior parte destes dados são disponibilizados sob a forma de texto livre, uma forma conveniente de expressar conceitos e termos mas particularmente desafiante quando se pretende realizar procuras, sumarização ou análise estatística de uma forma automática. As tecnologias de extração automática de informação podem ajudar a solucionar alguns destes problemas através da interpretação semântica e da abstracção do conteúdo dos textos. O sistema de Extração de Informação Médica apresentado neste documento, o MedInX, é o primeiro sistema desenvolvido para o processamento de cartas de alta hospitalar escritas em Português. O principal objectivo do sistema é a melhoria do acesso à informação *trancada* nos textos e, consequentemente, a melhoria da eficiência dos cuidados de saúde, através do acesso rápido e confiável à informação, quer relativa ao doente, quer aos profissionais de saúde.

O MedInX utiliza diversas componentes, baseadas em princípios de processamento de linguagem natural, para a análise dos textos clínicos, e contém vários mecanismos para ler, processar e utilizar recursos externos, como terminologias e ontologias. Este recursos são utilizados, em particular, no mapeamento automático do texto livre para uma representação estruturada. No entanto, a arquitectura flexível e escalável do sistema permitiu, também, a sua aplicação na tarefa de Reconhecimento de Entidades Nomeadas numa avaliação conjunta relativa ao processamento de textos de domínio geral, escritos em Português.

A avaliação do sistema num conjunto de cartas de alta hospitalar reais, indica que o sistema realiza a tarefa de extração de informação com uma medida F de 95% e a tarefa de extração de relações com uma precisão de 95%.

A utilidade do sistema em aplicações reais é demonstrada através do desenvolvimento de um conjunto de projectos exemplificativos, que pretendem responder a problemas concretos e comuns em ambiente hospitalar. Estes problemas estão relacionados com a codificação automática de diagnósticos e de outras condições relacionadas com o estado de saúde do doente, seguindo as classificações internacionais, ICD-9-CM e ICF. A revisão automática do conteúdo dos documentos é outro exemplo das possíveis aplicações práticas do sistema. Esta última aplicação é representada pelo o sistema de auditoria do MedInX.

keywords

Human Language Technologies, Medical Language Processing, Information Extraction, Clinical Knowledge Representation, Medical Ontologies, Automatic Coding, Decision Support Systems.

abstract

The electronic storage of medical patient data is becoming a daily experience in most of the practices and hospitals worldwide. However, much of the data available is in free-form text, a convenient way of expressing concepts and events, but especially challenging if one wants to perform automatic searches, summarization or statistical analysis. Information Extraction can relieve some of these problems by offering a semantically informed interpretation and abstraction of the texts.

MedInX, the Medical Information eXtraction system presented in this document, is the first information extraction system developed to process textual clinical discharge records written in Portuguese. The main goal of the system is to improve access to the information locked up in unstructured text, and, consequently, the efficiency of the health care process, by allowing faster and reliable access to quality information on health, for both patient and health professionals.

MedInX components are based on Natural Language Processing principles, and provide several mechanisms to read, process and utilize external resources, such as terminologies and ontologies, in the process of automatic mapping of free text reports onto a structured representation.

However, the flexible and scalable architecture of the system, also allowed its application to the task of Named Entity Recognition on a shared evaluation contest focused on Portuguese general domain free-form texts.

The evaluation of the system on a set of authentic hospital discharge letters indicates that the system performs with 95% F-measure, on the task of entity recognition, and 95% precision on the task of relation extraction.

Example applications, demonstrating the use of MedInX capabilities in real applications in the hospital setting, are also presented in this document. These applications were designed to answer common clinical problems related with the automatic coding of diagnoses and other health-related conditions described in the documents, according to the international classification systems ICD-9-CM and ICF. The automatic review of the content and completeness of the documents is an example of another developed application, denominated MedInX Clinical Audit system.

“When I say the word ‘umbrella’, you see the object in your mind. You see a kind of stick, with collapsible metal spokes on top that form an armature for a waterproof material which, when opened, will protect you from the rain. This last detail is important. Not only is the umbrella a thing, it is a thing that performs a function - in other words, expresses the will of man... my question is this.

What happens when a thing no longer performs its function?

Is it still the thing, or has it become something else?

When you rip the cloth off the umbrella, is the umbrella still an umbrella?

You open the spokes, put them over your head, walk out into the rain, and get drenched. Is it possible to go on calling this object an umbrella?”

--Paul Auster, in The City of Glass

Contents

1	Introduction	1
1.1	Human Language Technology and Clinical Text	1
1.2	Research objectives	4
1.3	Contributions	5
1.4	Published results	6
1.5	Outline	8
2	Background and Related Work	9
2.1	eHealth	9
2.2	Information Extraction	12
2.2.1	Information Extraction in the Clinical Domain	14
2.2.2	Named Entity Recognition	20
2.3	Clinical Knowledge Representation	25
2.3.1	Knowledge Representation and Ontologies	26
2.3.2	Terminological and Ontological Resources in the Medical Domain	28
2.4	Summary	33
3	Resources and Tools	35
3.1	Resources	35
3.1.1	Wikipedia	35
3.1.2	Unified Medical Language System	39
3.1.3	Corpus of Patient Discharge Letters	41
3.2	Processing Tools	55
3.2.1	UIMA	56
3.2.2	Protégé	58
3.2.3	Knowtator	58
3.2.4	Jena	59

3.2.5	Lucene	60
3.3	Summary	60
4	Medical ontologies	61
4.1	Method for Ontology Development	62
4.2	ICD-9-CM ontology	62
4.3	ICF ontology	65
4.4	Drugs ontology	69
4.5	MEDINX Patient Discharge Letters Knowledge Representation	70
4.6	Summary	78
5	MEDical Information eXtraction system	81
5.1	MEDINX architecture	81
5.2	Document Reader	83
5.3	General natural language processing	84
5.4	REMMIX, the Named Entity Recognizer of MEDINX	86
5.4.1	REMMIX in the Second HAREM contest	86
5.4.2	REMMIX in Clinical Text	93
5.5	Consumers	100
5.6	Summary	101
6	Evaluation	103
6.1	Evaluation Metrics	103
6.2	MEDINX results in Second HAREM contest	104
6.2.1	Results	105
6.2.2	Discussion	109
6.3	MEDINX results in Patient Discharge Letters	110
6.3.1	System setting	110
6.3.2	Evaluation Setup	111
6.3.3	Judges Expertise	114
6.3.4	Evaluation	114
6.3.5	Interjudge agreement	117
6.3.6	Evaluation Results	119
6.3.7	Intrajudge Agreement	128
6.3.8	Discussion	130
6.4	Summary	131

7	Example Applications	133
7.1	Automatic Code Assignment support	133
7.1.1	Automatic ICD-9-CM code assignment	134
7.1.2	Automatic ICF code assignment	134
7.2	Audit system	136
7.3	Summary	140
8	Conclusions	141
8.1	Future work	142
	Bibliography	143
A	Annotation Guidelines	169
B	MEDINX type system	213
C	Evaluation Manual	215
D	MEDINX Results	231
E	ICF Summary	235

List of Figures

2.1	RTS Regional Electronic Patient File.	12
2.2	Generic architecture of an Information Extraction system.	13
2.3	Tree of the categories and types considered in the Second HAREM contest (adapted from (Carvalho <i>et al.</i> , 2008)).	22
2.4	RDF triple.	27
3.1	Partial view of the UMLS Semantic Network.	40
3.2	Clinical Information Flow	43
3.3	Style-preserving illustration of a Patient Discharge Letter. An original document written in Portuguese is presented on the left side of the figure, and on the right side, its English translation.	46
3.4	Patient Discharge Summaries Knowledge Representation Model	54
3.5	Gold Standard annotation representation in Knowtator.	55
3.6	UIMA architecture (adapted from (Ferrucci and Lally, 2004)).	57
3.7	Protégé’s default interface	59
4.1	MedInX ICD-9-CM ontology model.	64
4.2	ICD-9-CM ontology in Protégé	65
4.3	Example of the structure of an ICF code.	66
4.4	MedInX ICF ontology model.	67
4.5	ICF Model ontology in Protégé	68
4.6	MedInX Drugs ontology model.	69
4.7	Drugs ontology in Protégé	70
4.8	Tree of categories, types and subtypes considered in the Patient Discharge Letters (PDLs) knowledge representation.	72
4.9	Relationships between the class <i>Episode</i> and the classes <i>Condition</i> , <i>Anatomical Site</i> , <i>Evolution</i> and <i>Procedure</i> in the MEDINX ontology.	75

4.10 Relationships between the class <i>Procedure</i> and the classes <i>Condition</i> , <i>Anatomical Site</i> , <i>Modifier</i> , <i>Negation</i> . <i>Time</i> and <i>Value</i> in the MEDINX ontology.	76
4.11 Properties of the class <i>Procedure</i> and its subclasses.	77
4.12 Properties and subclasses of <i>Condition</i> .	77
4.13 Detail of the class <i>Time</i> and <i>Value</i> .	78
4.14 <i>Negation</i> , <i>Lateralization</i> and <i>Modifier</i> classes.	78
4.15 <i>Administration Route</i> , <i>Evolution</i> , <i>Anatomical Site</i> and <i>Episode</i> classes.	78
4.16 MEDINX ontology in Protégé	79
5.1 System Architecture	82
5.2 Part of MEDINX UIMA type system inheritance diagram.	83
5.3 REMMIX pipeline in HAREM II	87
5.4 REMMIX pipeline in the clinical setting	94
5.5 Part of the MEDINX type system inheritance diagram.	99
6.1 REMMIX runs in HAREM II	105
6.2 Precision and Recall in the task of semantic classification for each run and category.	107
6.3 Normalized comparison according to the minimum and maximum values.	108
6.4 REMMIX runs in PDLs processing.	111
6.5 Evaluation interface.	112
6.6 Detail of the precision assessment: the entities and respective relations extracted from the sentence <i>Ischemic stroke - recent ischemic stroke of the posterior cerebral artery</i> .	113
6.7 Evaluating entities and relations.	113
6.8 Recall assessment.	114
6.9 Frequency of each category for <i>Ontology+Resources</i> and <i>Resources</i> runs	118
6.10 Evaluation of each reviewer of the <i>Resources</i> and <i>Ontology+Resources</i> runs	118
6.11 Distribution of the Precision and Recall of MEDINX in the task of semantic classification, grouped by <i>Reports</i> and <i>Reviewers</i> .	121
6.12 F-measure of MEDINX	122
6.13 Precision for each category in the <i>Resources</i> and <i>Ontology+Resources</i> runs	123
6.14 Recall for each category in the <i>Resources</i> and <i>Ontology+Resources</i> runs	124
6.15 F-measure for each category in the <i>Resources</i> and <i>Ontology+Resources</i> runs	124
6.16 Precision x Recall in the <i>Resources</i> run	125
6.17 Precision x Recall in the <i>Ontology+Resources</i> run	126
6.18 Precision for each relationship in the <i>Resources</i> and <i>Ontology+Resources</i> run	127
6.19 Precision on the task of Relation Extraction.	128

7.1	Automatic ICD-9-CM code assignment with MEDINX.	134
7.2	REMMIX pipeline with the ICF annotator.	135
7.3	Automatically created ICF summary of the PDL presented in Figure 7.4. . . .	136
7.4	Style-preserving illustration of the Patient Discharge Letter used to create the ICF Summary presented in Figure 7.3. The underlined text corresponds to the ICF entities automatically coded by the system.	137
7.5	Audit System Rule Editor.	139
7.6	SPARQL query <i>Episodes with less than 17 conditions and more than 17 medications.</i>	139
7.7	Audit System Results interface.	140
B.1	Entity type system inheritance diagram.	213
B.2	MedInXConcept type system inheritance diagram.	214
E.1	Automatically created ICF summary.	236

List of Tables

2.1	Main features of the general Clinical Information Extraction systems.	17
2.2	Main characteristics of the presented ontologies and terminologies	32
3.1	Examples of Wikipedia titles, aliases and categories from the disambiguation page of the acronym “EHR”.	37
3.2	MRCONSO columns used in MEDINX.	41
3.3	Principal diagnoses for cerebrovascular diseases.	45
3.4	Number of non-empty default segments of the corpus of 915 PDLs.	46
3.5	Number of tokens and sentences of the MEDINX corpus.	47
3.6	Descriptive statistics for the amount of tokens of discharge documents and its segments.	47
3.7	Descriptive statistics for the amount of sentences of discharge documents and its segments.	48
3.8	Structural and contextual features of MEDINX discharge documents. The original examples are added in ().	49
3.9	Number of documents, tokens and sentences in the MEDINX corpus and its subsets.	50
3.10	Entities of the gold standard annotation schema.	52
3.11	Relationships used on the gold standard annotation schema.	53
4.1	Classes and Subclasses of the MEDINX Knowledge Representation	73
4.2	Relations of the MEDINX Knowledge Representation.	74
5.1	Acronyms and abbreviations commonly used in the MEDINX development set.	85
5.2	Examples and total number of signal words used in the definition of contextual rules.	88

5.3	Examples and total number of signal words used in the extraction of a category from an Wikipedia's article first sentence.	91
5.4	UMLS semantic types used in MEDINX.	93
6.1	MEDINX results in the task of semantic classification.	105
6.2	MEDINX results for each category in the task of semantic classification. . . .	106
6.3	Number of PDLs evaluated by each judge.	115
6.4	Descriptive statistics for the amount of entities extracted from the 30 reports of the <i>Ontology+Resources</i> run.	115
6.5	Descriptive statistics for the amount of entities extracted from the 20 reports of the <i>Resources</i> run.	116
6.6	Descriptive statistics for the amount of relations extracted from the 30 reports of the <i>Ontology+Resources</i> run.	116
6.7	Descriptive statistics for the amount of relations extracted from the 20 reports of the <i>Resources</i> run.	117
6.8	Inter-judge agreement.	119
6.9	Overall results of MEDINX in the task of semantic classification.	120
6.10	MEDINX results for each category.	123
6.11	MEDINX results in the relation extraction task.	127
6.12	MEDINX results for each relationship.	127
6.13	Intrajudge agreement.	129
6.14	Intrajudge agreement for each of MEDINX categories.	129
7.1	Five most common Diseases, Procedures and Drugs mentioned in the PDLs. .	138
D.1	Descriptive statistics for the amount of entities extracted from the 30 reports of the <i>Ontology+Resources</i> run.	231
D.2	Descriptive statistics for the amount of entities extracted from the 20 reports of the <i>Resources</i> run.	232
D.3	Descriptive statistics for the amount of relations extracted from the 30 reports of the <i>Ontology+Resources</i> run.	232
D.4	Descriptive statistics for the amount of relations extracted from the 20 reports of the <i>Resources</i> run.	233

Glossary

API Application Programming Interface.

CAS Common Analysis Structure.

CC Complicating Condition.

CUI Concept Unique Identifier.

DARPA Defense Advanced Research Projects Agency.

DDS Decision Support Systems.

DeCS Descritores em Ciencias da Saúde.

DKPro Darmstadt Knowledge Processing Repository.

DRG Diagnosis Related Group.

EHR Electronic Health Records.

EMR Electronic Medical Records.

GUI Graphical User Interface.

HAREM Avaliação de Reconhecimento de Entidades Mencionadas.

HIP Hospital Infante D. Pedro.

HLT Human Language Technology.

ICD International Classification of Diseases.

ICD-9 International Classification of Diseases, Ninth Revision.

ICD-9-CM International Classification of Diseases, Ninth Revision, Clinical Modification.

ICF International Classification of Functioning, Disability and Health.

ICT Information and Communication Technology.

IE Information Extraction.

IR Information Retrieval.

IHTSDO International Health Terminology Standards Development Organization.

KR Knowledge Representation.

LAT Language of Terms.

LSP-MLP Linguistic String Project - Medical Language Processor.

MedLEE Medical Language Extraction and Encoding system.

MeSH Medical Subject Headings.

MLP Medical Language Processing.

MUC Message Understanding Conferences.

NE Named Entity.

NER Named Entity Recognition.

NLP Natural Language Processing.

OBO Open Biomedical Ontologies.

OWL Web Ontology Language.

PDL Patient Discharge Letter.

POS Part-of-speech.

REMMIX Reconhecimento de Entidades Mencionadas do MedInX.

RDF Resource Description Framework.

RTS Rede Telemática da Saúde.

SAB Source Abbreviation.

SNOMED-CT Systematized Nomenclature of Medicine - Clinical Terms.

SPARQL SPARQL Protocol an RDF Query Language

SPRUS Special Purpose Radiology Understanding System.

SW Semantic Web.

TTY Term Type.

UIMA Unstructured Information Management Architecture.

UKP Ubiquitous Knowledge Processing.

UML Unified Modeling Language.

UMLS Unified Medical Language System.

URI Uniform Resource Identifier.

USNLM United States National Library of Medicine.

WHO World Health Organization.

WSD Word Sense Disambiguation.

W3C World Wide Web Consortium.

Introduction

The recent growth of information technology and the wide prevalence of electronic medical records in health care institutions contributed to the increasingly available amount of electronic clinical data, created as part of routine patient care. However, despite the emphasis on collecting information in structured fields of electronic records, the key clinical information still resides in unstructured (free) text. Natural language text is a convenient form of information entry, specially, in the clinical setting, but it is also a challenge for automatic computerized access as it is extensive, ambiguous and contains personal differences in vocabulary and style.

Human language technologies, the computational tools that analyse and generate natural human language, can relieve some of these problems by offering a semantically informed interpretation and abstraction of the texts (Allvin *et al.*, 2010). Semantically processed text can support abstraction, visualization and query tools for the benefit of clinicians, patients, quality managers and researchers.

These techniques compose the field of Natural Language Processing (NLP). NLP is an interdisciplinary field concerned with the processing of language by computers (Mitkov, 2003). Research in NLP involves concepts in computer science, linguistics and logic. However, NLP has a special role in computer science because it seeks to model language computationally and many aspects of the field are concerned with linguistic features of computation (Joshi, 1991).

1.1 Human Language Technology and Clinical Text

Electronic access to health information is becoming more frequent worldwide through numerous electronic health information systems, which generate millions of gigabytes of health information annually, more than in many other domains (Cios and Moore, 2002). In fact, this type of access to information is critical for improved health care and to reduce medical errors.

In Portugal, with a population of, approximately, ten million inhabitants, the latest available statistics (National Institute of Statistics, 2009) report that every year there is almost a million of hospital admissions and forty six million specialized care outpatient visits in the National Health System. European Commission recent studies indicate that 73.6% of the individual administrative patient data originated in Portuguese health institutions is electronically recorded and stored (European Commission, 2008).

Several efforts are currently being undertaken to generalize the use of Electronic Medical Records (EMR) in the Portuguese health institutions. However, most of the information available in these systems is in textual form and, even if electronically available, remains locked up within text.

The content of a patient health record is extremely complex. These texts describe a sequence of events, narratives, reflecting the need for a precise and complete explanation when describing the health status of a patient. This type of expressive description inherently bears a substantial ambiguity and personal differences in vocabulary and style (Lovis *et al.*, 2000; Suominen, 2009). Frequently, specialists from the same discipline cannot agree on unambiguous terms to be used while describing a patient's condition. Not only they use different names (synonyms) to describe the same disease, but they render the task even more daunting by using different grammatical constructions to describe relationships among medical entities (Cios and Moore, 2002).

In general, clinical texts are written by clinicians in the clinical setting. These texts describe patients, their pathologies, their personal, social and medical histories, and findings made during interviews or during procedures (Meystre *et al.*, 2008).

The main purpose of this type of text is to serve as a summary or hand-over note of patient care, but it can also be used for legal requirements, care continuity, reimbursement, management and research. Clinical text covers every care phase and, depending on the purpose the documents differ. Clinical text may have been entered in “real time” or in retrospect, as a summary, by the bedside or elsewhere (Thoroddsen *et al.*, 2009).

This type of text contrasts with biomedical text, which is the kind of text that appears in books, articles, literature abstracts, posters, and so forth. But what makes clinical text different from biomedical text, and why does it pose a special challenge to NLP?

1. Some clinical texts do not conform to the rules of grammar, are short and composed of telegraphic phrases;
2. Clinical narratives are full of abbreviations, acronyms and other shorthand phrases. Also, these shorthand lexical units are often overloaded, i.e., the same set of letters has multiple interpretations (Liu *et al.*, 2001);
3. Misspellings are frequent in clinical text, as it is often produced without any spelling support;
4. Clinical narratives contain, frequently, pasted sets of laboratory values or vital signs with embedded non-text strings, complicating otherwise straightforward NLP tasks like sentence splitting, since they are usually filled with periods;
5. Templates and pseudotables as, for instance, plain text made to look tabular by the use of white space or lists, are common;

Information search from this type of narrative text is difficult and time consuming. Evidences have been given that electronic health information systems support information gathering but not its active utilization (Snyder-Halpern *et al.*, 2001). Standardization and structuring have been proposed as solutions for this situation. However, these approaches have also some problems. Converting narratives to numerical and structured data is laborious and leads easily to differences and errors in coding. Moreover, if these tasks are performed manually, which is the most common approach currently, text ambiguity and personal differences may cause inconsistencies (Suominen, 2009). Also, converting narratives into structured data may lead to significant information losses, as it limits the expressive power of free-text (Lovis *et al.*, 2000; Walsh, 2004).

Research should, therefore, aim to develop automated systems capable of standardizing and structuring the clinical narratives, in a systematic way, and retain the benefits of this type of free-form text as its expressiveness and ease of production.

Human Language Technology (HLT) and, particularly, NLP, is increasingly gaining the interest of both health care practitioners and academic researchers. Instead of leaving the health professional with the problem of having to read several tens of thousands of documents, HLT offers the possibility of extracting precise facts from a document set and of finding interesting associations among disparate facts, leading to the discovery of new or unsuspected knowledge (Ananiadou and Mcnaught, 2005).

The increased interest in HLT in the clinical domain is evident through numerous dedicated workshops and tutorials in major conferences in bioinformatics and NLP. Examples of these conferences include the International Conference on Computational Linguistics, the

Association for Computational Linguistics Annual Meeting and their associated workshops, as the Workshop on Current Trends in Biomedical Natural Language Processing, held since 2002, and the Louhi Workshop on Text and Data Mining of Health Documents, introduced in 2008. Several international journals dedicate their pages to the applications of HLT in the clinical area, as the International Journal of Medical Informatics and the Journal of the American Medical Informatics Association. The recent interest of IBM and Nuance, two leading Information and Communication Technology (ICT) companies, in directing their *Jeopardy!* winning system, Watson (IBM, 2011), into the health-care industry, particularly to process clinical reports, can also help understanding the emergent need for technology that can process large amounts of health care data.

At the moment, HLT is still rarely used in EMR, in particular in Portuguese and many other European languages, although promising results were already accomplished in very specific tasks, as this thesis reports (see chapter 2). More work is needed to connect these components and actors together to develop more comprehensive solutions that establish themselves in Electronic Health Records (EHR) systems.

1.2 Research objectives

The main goal of the system presented in this document is to improve the access to the information locked up in free-form text. The system concentrates in the processing of clinical discharge records and has as main objectives the improvement of the efficiency of the health care process by allowing faster and reliable access to quality information on health, for both patient and health professionals. By enabling the creation of faster and more accurate structured clinical data, the system aims to increase safety and reduce medical errors, while the standardization of the information included in the discharge reports can help improving patient mobility and borderless access to health care.

Moreover, the system aims to reach the previous goals while allowing physicians to continue to practice using their current descriptive language in free-text reports without a requirement to enter structured data in a complex, time consuming computer-based system.

The research objectives of this thesis are mainly related with the study and application of HLT to the processing of textual health records.

The first category of objectives addresses the problem of developing a system capable of extracting information from free text discharge letters and use this information to gather ontological clinical knowledge to be used in practical applications capable of improving health care efficiency. The use of authentic health records was a prerequisite for developing a medical language processing system to be used in real clinical applications. Given so, one of the

objectives is the development of a corpus of authentic discharge letters written in Portuguese and originated in Portuguese health care institutions.

The creation of medical knowledge sources and methods for the automatic processing clinical of text written in Portuguese is another objective of this work.

Being the work intended to be used in the health area, evaluation plays an important role. Safety requirements and need to guarantee valid information transmission is a main concern in clinical text processing. Given so, the last category of objectives refers to the performance of reliable evaluation of the system results.

1.3 Contributions

This thesis presents a novel system for the automatic processing of hospital Patient Discharge Letters (PDLs) written in Portuguese, the MEDINX system. It allows, not only the automatic extraction of information from free text clinical reports, but also automatically instantiating a new knowledge representation model for the discharge reports. The system developed is the first Information Extraction (IE) system for the automatic processing of Portuguese clinical narratives.

The full implementation of the system brought some other innovative contributions to the field of Medical Language Processing (MLP) in Portuguese. These can be summarized as follows:

1. Knowledge resources in Portuguese:

- (a) a new clinical corpus of authentic PDL of the area of hypertension and cerebrovascular diseases;
- (b) new knowledge representation model of the discharge letters being produced in the Hospital Infante D. Pedro (HIP);
- (c) a new computational model and detailed guidelines for the creation of the gold standard corpus of discharge letters;

Following the recent recommendations of the European Commission on the development of realistic and accessible solutions, with a modest approach to terminology development of focused and incremental steps (European Commission, 2009), four new ontologies were created:

- (a) two new ontologies reflecting widely used international classification systems, the International Classification of Diseases, Ninth Revision, Clinical Modification

- (ICD-9-CM) and International Classification of Functioning, Disability and Health (ICF);
 - (b) one ontology of the drugs currently being used in Portugal;
 - (c) a conceptualization of the structure and content of the PDLs used in this work.
2. Medical language processing applications:
- (a) Clinical Named Entity Recognizer for Portuguese discharge reports, capable of reading, processing and using external resources, such as terminologies and ontologies;
 - (b) three example practical applications of the system in real clinical tasks: support to Automatic Code Assignment following the international classifications ICD-9-CM and ICF, and a Clinical Audit system.
3. Evaluation
- (a) Creation and implementation of new evaluation tools.

1.4 Published results

The medical language processing system and its components have been already subject for several publications over the last years. MEDINX development approach was established to be extensible and scalable, in a system capable of using several external knowledge resources, from several domains, in the task of IE in European Portuguese. This flexible architecture allowed the application of the system in several domains and tasks, as the following list of publications indicates:

- Recently a book chapter proposal was submitted for publication in the *Handbook of Research on Information and Communications Technology for Healthcare and Social Services*, I. M. Miranda and M. M. Cruz-Cunha (eds.). The book chapter focuses on the final version of MEDINX and its example applications.
 - An article describing the first results obtained in the application MEDINX to IE from PDLs using an ontological approach was presented in the VI *Jornadas en Tecnología del Habla*, Speech and Language Technologies for Iberian Languages.
- L. Ferreira, A. Teixeira and J. P. S. Cunha (2010). *Information Extraction from Portuguese Discharge Letters*. In C. G. Mateo, F. C. Díaz and F. M. Pazó (eds.), FALA 2010: VI Jornadas en Tecnología del Habla, Speech and Language Technologies for Iberian Languages, pages 39-42, Vigo, Spain.

- The first application of MEDINX in authentic hospital discharge reports written in Portuguese is described in *Linguamática*, the International Journal of Automatic Processing of Iberian Languages:

L. Ferreira, C. Oliveira, A. Teixeira and J. P. S. Cunha (2009). *Extracção de Informação de Relatórios Médicos*. In J. J. Almeida, A. M. B. Simões and X. G. Guinovart (eds.), *Linguamática, Revista para o Processamento Automático das Línguas Ibéricas*, pages 89-101, Volume 1, Number 1.

- Integrated in the 10th anniversary of Linguateca, a review of the technologies used in MEDINX was published in:

L. Ferreira, A. Teixeira and J. P. S. Cunha (2008). *Linguatca e Processamento de Linguagem Natural na Área da Saúde: Alguns Comentários e Sugestões*. In L. Costa, D. Santos and N. Cardoso (eds.), *Perspectivas sobre a Linguatca / Actas do encontro Linguatca : 10 anos*, pages 13-23, Linguatca.

- The participation of MEDINX in the Second HAREM contest, a Named Entity Recognition (NER) contest for Portuguese, is presented in the book chapter:

L. Ferreira, A. Teixeira and J. P. S. Cunha (2008). *REMMA - Reconhecimento de Entidades Mencionadas do MedAlert*. In C. Mota and D. Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 213-229, Linguatca.

- An experiment in the development and automatic population of a vaccination ontology and the system that automatically recognizes vaccination information on medical Portuguese texts was presented in the 5th International Workshop on Natural Language Processing and Cognitive Science, in 2008. The paper also describes a first approach to automatic information extraction regarding inter-class relations using Association Rule Mining.

L. Ferreira, A. Teixeira and J. P. S. Cunha (2008). *Ontology-driven Vaccination Information Extraction*, In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, pages 94-103, Barcelona, Spain.

Integrated in the studies developed in this doctoral project has been a short visit to the Ubiquitous Knowledge Processing (UKP) lab in the Technical University of Darmstadt, Germany. The area of Opinion Mining and, particularly, the detection of product features in customer reviews was the main research topic of the period spent in the UKP lab. Although several learned techniques have been, implicitly or explicitly, applied in this work, the main outcome of this cooperation has been published in :

L. Ferreira, N. Jakob and I. Gurevych (2008). *A Comparative Study of Feature Extraction Algorithms in Customer Reviews*. In Proceedings of the 2nd IEEE International Conference on Semantic Computing, ICSC, pages 144-151, Santa Clara, USA.

1.5 Outline

In this thesis, several aspects of medical IE in European Portuguese are investigated.

Chapter 2 introduces the areas and the existing work related with the thesis. It starts by describing eHealth and how it is currently applied in Portuguese health care institutions. The main fields of Information Extraction and Knowledge Representation are also introduced in this chapter, with a particular focus on the applications in the clinical subdomain.

Chapter 3 presents the knowledge sources and the tools used in the development of the MLP system. The corpus of authentic hospital discharge letters used in this work, is also introduced in this chapter.

Chapter 4 presents the knowledge representation models developed and their current content. This chapter describes the ontologies created in the context of this work, namely, two formalizations of the international classification systems ICD-9-CM and ICF, a drugs ontology and a conceptualization of the structure and content of the discharge reports, the MEDINX ontology.

Chapter 5 presents the architecture of the system, the components and algorithms used to extract knowledge from clinical text and to automatically populate the MEDINX ontology.

Chapter 6 discusses the methods used to evaluate the performance of the system in two distinct phases and applications of the system. The chapter starts by presenting the results obtained by the system in the Second HAREM contest, a general domain NER contest developed by Linguateca. The results on the task of clinical entities and relations extraction, on a set of authentic PDLs, conclude the chapter.

Chapter 7 presents some examples of practical applications of the system, namely the Automatic Code Assignment system and the Clinical Audit system developed to exemplify some of the several possible applications of the work in real clinical settings.

Chapter 8 concludes the document and presents topics for future work.

Background and Related Work

This chapter analyzes recent research literature on the application of HLT to the clinical domain. It starts by introducing the area of eHealth and its current application in Portuguese health institutions, and gives particular attention to the EMR system which motivated the development of MEDINX, the *Rede Telemática da Saúde (RTS)* project. Information extraction and its applications in the clinical area are also presented, and the chapter ends with a brief overview of the Clinical Knowledge Representation area.

Research on information extraction from biomedical literature is not in the range of this work and, thus, is not discussed in this document, but is widely described in reviews by Cohen and Hersh (2005) and Zweigenbaum *et al.* (2007).

2.1 eHealth

The term eHealth has been used to describe activities that employ information and telecommunication technologies to deliver health care (Eysenbach, 2001). The term is frequently used to describe a range of services or systems that are at the edge of medicine/health care and information technology. These systems are used to assist and enhance the prevention, diagnosis, treatment, monitoring and management of health and lifestyle (Oh *et al.*, 2005).

Over the past years, a rapid development has taken place in the eHealth area in Europe. A basic ICT infrastructure, consisting of computers and internet connections, is today available in most of the general practitioners or primary care physicians practices in Europe. The electronic storage of administrative and medical patient data, the use of a computer during consultation, and other uses of ICT in the health area, are becoming a daily experience in

the practices. European Commission recent studies (European Commission, 2008) show that 87% of all primary care physician practices use a computer. In Portugal, this value ascends to 88%. The studies also show that, approximately, 74% of the administrative patient data originated in Portuguese health care institutions is currently stored in electronic format.

The worldwide interest in eHealth technologies is also shown by the recent efforts undertaken by several ICT companies in applying their tools to the health care arena.

An example of this interest is Microsoft's HealthVault (Microsoft, 2007), first presented in October 2007 as a platform to store and maintain health and fitness information. HealthVault addresses both individuals and health care professionals, and stores health information from an individual. It also allows health and fitness data to be transferred from devices, such as heart rate watches and blood pressure monitors, into an individual's HealthVault record.

Another example is given by Google Health service (Google, 2011b). First unveiled in February 2008, Google Health is a personal health information centralization service that allows Google users to volunteer their health records into the system and thereby merging potentially separate health records into a single Google Health profile. Volunteered information can include health conditions, medications, allergies, and lab results. Once entered, Google Health uses the information to provide the user a merged health record containing information about the diseases and possible interactions between drugs, patient's conditions, and allergies.

More relevant to the work presented in this document, is the significant effort invested in the last years in the development of systems for electronic access to clinical information within hospitals. This kind of access to clinical information is becoming a generalized reality in Europe and all over the world through numerous enterprise-wide EHR systems (European Commission, 2008; Stroetmann *et al.*, 2007).

An EHR system can be described as a systematic collection of electronic information about the individual patients or populations health. The Health Information Management Systems Society (Health Information Management Systems Society, 2010), a not-for-profit organization dedicated to promoting a better understanding of health care information and management systems, defines the EHR as a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports.

It is important, however, to clarify the differences between EHR and EMR. The EMR is the legal record created in hospitals and ambulatory environments that is the source of data for the EHR. An EHR automates and streamlines the clinician's workflow. The EHR has the ability to generate a complete record of a clinical patient encounter, as well as supporting

other care-related activities, directly or indirectly, including evidence-based decision support, quality management, and outcomes reporting (Habib, 2010). However, before we can move to effective EHR environments, provider organizations must implement complete EMR solutions.

In Portugal, efforts are being followed to generalize the use of EMR in hospitals and other health facilities. Currently, several EMR are already in use within several Portuguese health institutions, as the latest state of the art report developed by the Portuguese Ministry of Health and the Central Administration of the Health System (Ministério da Saúde and Administração Central do Sistema de Saúde, 2009), shows. The document lists the following examples of Portuguese EMR systems:

- *Processo Clínico Electrónico Único do Sistema Regional de Saúde da Região Autónoma da Madeira*, the individual EMR system in development in the Autonomous Region of Madeira (Ministério da Saúde and Administração Central do Sistema de Saúde, 2009, p. 96);
- *Rede Telemática da Saúde* is an integrated health telematics network, developed within the *Aveiro Digital* project and in use in the health institutions of Aveiro's region. This system is described in more detail in the following paragraphs;
- *Urgência Pediátrica Integrada do Porto (UPIP)*, an integrated network of pediatric emergency care in use in the region of Porto since June 2008 (Ministério da Saúde and Administração Central do Sistema de Saúde, 2009, p. 103);
- *Processo Clínico Electrónico*, the EMR system of Hospital Geral de Santo António in Porto, developed in cooperation with the University of Minho (Ministério da Saúde and Administração Central do Sistema de Saúde, 2009, p. 108);
- *HSJ.ICU (Informação Clínica do Utente)*, the EHR system of Hospital de São João, also in Porto (Ministério da Saúde and Administração Central do Sistema de Saúde, 2009, p. 111);
- *Unidade de Hematologia do Centro Hospitalar de Coimbra*, a pilot project from Coimbra Hospital Center which aims the sharing and secure access to information from clinical records of patients with hematologic (hemato-oncology and general hematology) pathologies (Ministério da Saúde and Administração Central do Sistema de Saúde, 2009, p. 113).

The *Rede Telemática da Saúde* project (Oliveira and Cunha, 2011), the EMR system in use in the region of Aveiro, and the source of the clinical reports used in this work, is a telematic infrastructure to progressively support multiple clinical communication services at a regional level. The project aims at promoting the secure electronic access to clinical information

stored in the different health care providers to all credentialed professionals. The RTS provides a summarized *Regional Electronic Patient File* (Figure 2.1) that combines “on the fly” information from the clinical electronic files existing in all the institutions that compose the network, smoothing the clinical communication between different health institutions. The RTS allows information, such as release forms, clinical tests bulletins and vaccination records, to be accessible to the professionals from other institutions through a simple web browser.

Sumário do episódio

Carta de Alta Administrativa	Não disponível
Carta de alta	
Diagnósticos	
Diário	
Prescrição	
Transferências	
Sessões de Hospital de Dia	

© 2004-2008 Rede Telemática da Saúde

Figure 2.1: RTS Regional Electronic Patient File.

At the same time, citizens can manage their health issues over the RTS telematic platform. Registered users can interact with their family doctor, request appointments and manage their health agenda through the restricted area *MyRTS*. A public area is also available with general interest health related contents, such as health guidelines and health issues.

RTS is currently being used in Aveiro’s region, in Portugal, and provides access to more than 11,000,000 clinical episodes from more than 350,000 patients. In particular, RTS is the EMR system being used in Hospital Infante D. Pedro (HIP), and where the PDLs used in this work are created.

Several benefits have already been experienced with the use of RTS in the region of Aveiro. These include avoiding duplication of diagnosis auxiliary exams and the reduction of unnecessary patients’ visits (Oliveira and Cunha, 2011).

2.2 Information Extraction

Information Extraction (IE) is a sub-area of NLP which goal is, as the term implies, to extract information from unstructured and/or semi-structured machine-readable documents, but to do so without requiring the end user of the information to read the data. It contrasts with Information Retrieval (IR) which focuses on finding documents. IR returns documents whereas IE returns information or facts. Some very popular examples of IR systems are Google (Google, 2011a) or PubMed (USNLM, 2011) search engines. Having used a search engine, the user must read each document to know the facts reported in it. However, when the goal is to be able to tabulate the facts reported in a large quantity of documents, IE becomes a more relevant technology (McNaught and Black, 2005).

The remainder of this section presents a short introduction to the most common methods used for IE. It also introduces some recent published research in the area of clinical IE and in its subtask Named Entity Recognition (NER).

IE has become a popular research topic since the late eighties with the promotion of Message Understanding Conferences (MUC) supported by Defense Advanced Research Projects Agency (DARPA). The MUC conferences fostered much of the work in the IE domain and consisted of competitive evaluations of systems developed for extraction of specific information such as named entities (*people, organizations, locations*), events, and relations (e.g. `employee_of`, `location_of`, `manufacture_of`).

A typical IE system may have two main subtasks: entity recognition and relation extraction. Entity recognition seeks to locate and classify atomic elements in natural language texts into predefined categories, while relation extraction aims at identifying the relations between the entities in order to fill predefined templates.

Hobbs (1993, 2002) describe a typical IE system as cascade of transducers or modules, being each system characterized by its own set, but generally containing a tokenizer, sentence boundary detector, part-of-speech tagger, morphological analyzer, shallow parser, gazetteer, named entity recognizer, discourse module, template extractor, and template combiner. Figure 2.2 shows the combination of some of these modules in a simple pipeline architecture for an IE system which takes raw text of a document as its input and generates a list of `<entity, relation, entity>` tuples.

Several approaches to IE have been followed over the last years. One common approach is based on *pattern-matching* and exploits basic patterns over a variety of structures - text strings, part-of-speech tags, semantic pairs, and dictionary entries (Pakhomov *et al.*, 2005). However, this type of approach has a lack of generalizability, which limits its extension to new domains.

Figure 2.2: Generic architecture of an Information Extraction system.

The need for IE systems that can be easily adapted from one domain to another lead to the development of different approaches based on adaptive IE, starting with the Alembic Workbench (Aberdeen *et al.*, 1995) and followed by the Melita system (Ciravegna *et al.*, 2003). The idea behind these approaches was to use various kinds of *machine learning* algorithms to allow IE systems to be easily targeted into new problems. The effort required to redesign a new system was replaced with that of generating a new batch of training data and applying the learning algorithm (Tablan, 2009). Given that the work for generating training data comprises mainly manual annotation of text, which **usually** requires less skilled personnel, this brings down the overall cost for re-targeting a system to a new domain.

A more recent approach is the ontology-driven IE which aims at using ontology to guide free-form text processing (Hahn *et al.*, 2002). Here, ontologies play a crucial role as they provide formal and explicit specifications of conceptualizations (Wimalasuriya and Dou, 2010). Since Berners-Lee (Berners-Lee *et al.*, 1994; Berners-Lee and Fischetti, 1999) proposed and started to endorse ontologies as the backbone of the Semantic Web (SW) in the nineties, a whole research field evolved around the fundamental engineering aspects of ontologies, such as the generation, evaluation and management of ontologies. Several IE groups focused on the development of extraction methods that use the content and predefined semantics of an ontology to perform the extraction task without any human intervention and dependency on other knowledge resources (Yildiz, 2007; Buitelaar and Siegel, 2006; Embley *et al.*, 1998; Maedche *et al.*, 2002).

2.2.1 Information Extraction in the Clinical Domain

The previous examples illustrate the efforts carried by the general domain IE communities towards the development of sophisticated methods for deeper and more comprehensive information extraction from text. However, in the clinical and medical domain, the constraints

that protect patient confidentiality and impose limited access to shareable clinical data, complicate NLP on clinical text, making it especially challenging. In spite of the challenges, several research has been described in the literature on the topic of IE from clinical text.

This section focuses on recent published research involving extraction of information from clinical narratives. The work presented was mainly selected from bibliographic queries in PubMed (USNLM, 2011)¹, and from the reviews of recent research presented by Meystre *et al.* (2008) and Ananiadou and Mcnaught (2005).

A major challenge to IE in the clinical domain is the creation of shared data, tasks, annotation guidelines, annotations, and evaluation techniques (Meystre *et al.*, 2008). So far, there have been three clinical shared tasks and comparative evaluations on IE from clinical texts:

1. Automatic assignment of ICD-9-CM codes to clinical free text (Computational Medicine Center, 2007). The shared task occurred in 2007 and involved mapping ICD-9-CM codes to radiology reports. Pestian *et al.* (2007b) describe the task, its organization, and results.
2. De-identification of discharge summaries within the i2b2 (i2b2, 2006) initiative held in November 2006. The task is described in (Uzuner *et al.*, 2007) and pretended to remove all personal health information from the clinical documents. Top systems achieved F-measure results in the high 90%.
3. Patient smoking status discovery from discharge summaries within the i2b2 initiative held in November 2006 (Uzuner *et al.*, 2008).

2.2.1.1 General Clinic Information Extraction

In the clinical domain, IE was initially approached with complete systems, i.e. systems featuring all functions required to fully analyze free-form text.

The first of these large-scale projects was the Linguistic String Project - Medical Language Processor (LSP-MLP) from the New York University. The LSP-MLP aims at enabling extraction and summarization of signs/symptoms, drug information, and identification of possible medication side effects (Sager *et al.*, 1986). The system is based on a subset of natural language, so called *sublanguage*. In their evaluation test set, LSP-MLP achieved a performance of 82.1% recall and a precision value of 82.5%.

The Specialist system (McGray *et al.*, 1987) was developed at the United States National Library of Medicine (USNLM) and is intended to function as a set of NLP tools for biomedical

¹for “information extraction”, “text mining”, “natural language processing”, “medical language processing”, and “natural language understanding”

knowledge bases, and the Medline abstracts. The Specialist is part of the Unified Medical Language System (UMLS) project, which includes the Specialist Lexicon, the Semantic Network and the UMLS Metathesaurus.

Inspired by this work, Friedman *et al.* (1995) developed the Medical Language Extraction and Encoding system (MedLEE). MedLEE, created at the New York Presbyterian Hospital and at Columbia University, is mainly semantically driven and is used not only to extract information from clinical narrative reports, but also to participate in an automated decision-support system, and to allow natural language queries. MedLEE was the first biomedical NLP system to be applied to an institution different than the one where it was developed. This system has been adapted and used for several different tasks. Besides being progressively extended to most of the documents present in the EHR (Friedman, 2000) and tested for its transferability in another institution (Hripcsak *et al.*, 1998), it has been used to detect findings evocative of breast cancer (Jain and Friedman, 1997), to analyze modifications to data entry templates (Wilcox *et al.*, 2002) and even combined with machine translation to detect abnormal findings and devices in Portuguese radiology reports (Castilla *et al.*, 2007). The initial evaluation of this rule based system resulted in 85% recall and 87% precision results. Later, Hripcsak evaluated MedLEE to use the coded data for automated decision-support. The system was tested for identification of six medical conditions from radiology reports. Recall and precision were found to be 81% and 98%, respectively.

Special Purpose Radiology Understanding System (SPRUS) is an NLP application developed by the Medical Informatics group at the University of Utah (Haug *et al.*, 1990). SPRUS is a purely semantically-oriented system, and is not able to exploit syntactic information. The evaluation report for the prototype system presents values of 87% recall and 95% precision. The developers recognized the need for syntactic processing and created the Symbolic Text processor (SymText) (Haug *et al.*, 1995). SymText has syntactic and probabilistic semantic analysis and relies on Bayesian networks. The evaluation of the system achieves 92% recall and 88% precision values for producing diagnosis in chest radiology reports.

Several methods were used to perform clinical IE within these and other systems. IE has been attempted with the use of basic pattern matching techniques such as regular expressions. Dictionaries of variable size were also typically used. For example, Long (2005) has extracted diagnoses from discharge summaries using regular expressions and a punctuation marks dictionary, as well as, the UMLS Metathesaurus (McCray *et al.*, 1993). To extract blood pressure and antihypertensive treatment intensification information, Turchin *et al.* (2006) also used regular expressions. The REgenstrief eXtraction tool (Friedlin and McDonald, 2006) uses pattern matching to extract diagnoses and radiological findings related to congestive heart failure. Finally, a module extracting UMLS concepts and based on pattern matching, was

developed by Bashyam *et al.* (2007).

Approaches combining syntactic and semantic analysis constitute the majority of the systems. MedLEE is an example of such a system. MetaMap, and its Java version called MMTx (MetaMap Transfer), were also developed to extract information from clinical documents using such approaches. Some examples are given by Schadow and McDonald (2003), who used this method to extract information from pathology reports, Chung and Murphy (2005), who used it with echocardiography reports, and Meystre and Haug (2006), who used it to extract medical problems.

Other systems combining syntactic and semantic analysis have recently been developed and evaluated. The cancer Text Information Extraction System caTIES (National Cancer Institute, 2010) provides tools for de-identification and automated coding of free-form text structured pathology reports. It also has a client that can be used to search these coded reports.

A summary of the main features of the complete clinical IE systems described in the previous paragraphs is presented in Table 2.1.

Table 2.1: Main features of the general Clinical Information Extraction systems.

NAME	ORIGIN	USE CASES	METHODS
LSP-MLP (Sager <i>et al.</i> , 1986)	New York University	extracting and summarizing sign/symptoms, drugs information, identification of possible medication side effects	sublanguage
Specialist (McGray <i>et al.</i> , 1987)	USNLM	set of NLP tools for biomedical knowledge bases as Medline abstracts	lexicon
MedLEE (Friedman <i>et al.</i> , 1995)	New York Presbyterian Hospital and Columbia University	detect findings evocative of breast cancer, automatic code assignment, decision support	syntactic + semantic, rule-based
SPRUS (Haug <i>et al.</i> , 1990)	University of Utah	extract and encode both the findings and the radiologists' interpretations	semantic oriented
caTIES (National Cancer Institute, 2010)	University of Pittsburgh	de-identification and automated coding of free-text structured pathology reports	semantic + syntactic

2.2.1.2 Extracting Codes from Clinical Text

To develop and implement the complete MLP systems referred in the previous paragraphs, significant resources were required. In order to overcome this issue, several authors progressively experimented with more simple systems focused on specific IE tasks and on a limited number of different types of information to extract.

These more focused systems demonstrated good performance and are now the majority of systems used for IE. Some examples of this type of systems are the ones developed to automatically code clinical text, particularly, to extract codes of diagnoses using the International Classification of Diseases (ICD) versions 9 and 10 (the ICD classification is further introduced in Section 2.3.2.).

The year of the Medical NLP challenge, 2007, was a particularly interesting year for this type of systems. The Medical NLP challenge was a shared task exercise that provided test and training corpus of radiological reports and their corresponding human-coded ICD-9-CM codes. The challenge and its project are described in (Pestian *et al.*, 2007a).

Ultimately, 44 different research teams participated. Two participating systems are described by Aronson *et al.* (2007) and Crammer *et al.* (2007). Aronson *et al.* (2007) combined several existing technologies (e.g., USNLM’s Medical Text Indexer, a support vector machine classifier, a k-nearest neighbor classifier, etc.) and arranged them in a stack-like architecture to evaluate their fused performance. This architecture placed the system at 11th in the challenge, with a mean F-measure of 85%, considerably higher than the average score for all participants (mean score was 77%, while the best score was 89%). Crammer *et al.* (2007) also described a multi-component coding system; it used machine learning, a rule-based system, and an automatic coding system based on human coding policies. The system placed fourth in the challenge.

Outside of the Medical NLP challenge, MedLEE has been used to perform automatic code assignment in many contexts. Friedman *et al.* (1999) describes an automated pneumonia severity score coding system using the system, and Elkins *et al.* (2000) describe an adaptation of its use for neuroradiology standard concept extraction.

Another example is presented by Kukafka *et al.* (2006), who used MedLEE to encode rehabilitation discharge summaries according to the ICF (World Health Organization, 2010b) standards for health, and health-related states.

2.2.1.3 Clinical Text Mining

Text mining is the discovery and extraction of new knowledge from unstructured data (Hearst, 1999; Ananiadou and McNaught, 2005). It contrasts with data mining, which finds patterns from structured data, and with information extraction, which extracts the known facts from text and presents them in a structured form.

Typical text mining applications cluster documents in sets that have a common semantic basis. The semantic basis can be queried beforehand by the user, or discovered automatically by the software using statistical techniques. Other typical text mining applications try to

summarize documents, or conduct the user to parts of documents that contain information that the user probably did not see before. Swanson (1999) present the pioneer work on text mining, in which it is demonstrated that chaining facts from distinct literature sources could lead to the generation of new scientific hypotheses.

Clinical text mining has been mainly investigated for finding association patterns. An example, is the work developed by Chen *et al.* (2008). They employ text mining and statistical techniques to identify disease-drug associations in the biomedical literature and discharge summaries and conclude that there are distinct patterns in the drug usage as reported in the literature and as recorded in the patient record.

Other studies are presented by Cao *et al.* (2005) and Rindflesch *et al.* (2005). The first, shows that statistical methods are successful in finding strong disease-finding relations using a knowledge base construction for the patient problem list generation. Rindflesch *et al.* (2005) use statistical methods to construct a database of drug-disorder co-occurrences from a large collection of clinical notes from the Mayo Clinic.

2.2.1.4 Clinical Text Corpora and Their Annotation

The use of automatic IE tools depends heavily on the quality of the annotated corpora available for their training and testing. Given so, several work on guidelines development for clinical corpus annotation is being undertaken, particularly, on how to identify relevant features to annotate, and on the characterization of what makes a particular corpus usable.

Chapman and Dowling (2006) present an annotation schema to manually annotate clinical conditions, developed based on 40 emergency department reports and tested on 20 of such reports. The corpus was annotated by the two authors and achieved a high agreement and an F-measure of 93%. The study points out the lack of standard guidelines to determine which words to include in the annotation of clinical texts. Their proposal focuses on which semantic categories and words are important to include in such annotations.

In a follow-up investigation, Chapman *et al.* (2008) examined the improvement in agreement among annotators after they were trained with the annotation schema. For this investigation, three physicians and three lay people were used as annotators. The study concludes that lay people performed, almost, as well as physicians when trained on the schema. The results suggest that good annotation guidelines are essential to good annotation quality, especially when the annotators are not domain experts.

Roberts *et al.* (2007) describe the work carried out in the CLEF project for the semantic annotation of a corpus to assist in the development and evaluation of an IE system as part of a larger framework for the capture, integration and presentation of clinical information.

They created two annotation schema and used them to annotate a set of clinical entities, the relations between them, the time expressions and their temporal relations with the clinical entities in the text.

Liu *et al.* (2005) study and review the types of error a system that automatically extracts information from pathology reports makes. The information extracted was compared to a manually annotated gold standard. The authors classified the errors into system errors and semantic disagreement between the report and the annotation. Over half of the errors were not related to the information extraction system, but rather reflected disagreement between the reporting pathologist and the expert annotator. This second point showed that, even when gold standard annotations are available, they may still be difficult to interpret.

2.2.1.5 Clinical IE in Portuguese

The review of the literature on the topic of NLP of clinical text written in Portuguese reveals the lack of studies developed to process the Portuguese language, in any of its variants. In fact, one of the most relevant studies was developed in the University of São Paulo, Brazil, and focuses on the application of the medical language processor, MedLEE, developed for the English language, on a set of Portuguese medical texts, X-ray reports, machine translated into English. The results presented by Castilla *et al.* (2007) show that the successful application of systems developed for one language to another is not straightforward.

Some other studies were developed for the Brazilian variant of Portuguese, although only one has used a set of discharge summaries in Portuguese. Oleynik *et al.* (2010) present a study on the performance of a Part-of-speech (POS) tagger trained with a corpus discharge summaries in Portuguese, from the Clinical Hospital of Porto Alegre, Brazil which obtained an accuracy of 91.5%.

In European Portuguese, however, almost no efforts have been devoted to the analysis of this type of text. In fact, and despite the work presented in this thesis, we were only able to find information regarding a recent work on the development of ontologies to model neurovascular reports in Portuguese. Borrego and Quaresma (2010) developed an ontology to represent the information of the reports and a set of NLP based tools for its automatic population.

Also available, is an experiment developed by us in the automatic population of a vaccination ontology (Ferreira *et al.*, 2008c). The study concentrated on the methods used to model and populate the vaccination ontology and on the 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 (Agrawal

and Srikant, 1994).

The difficulty in obtaining access to clinical text, due mainly to confidentiality and data protection, is actually the highest obstacle in the development of further research on IE from clinical Portuguese texts.

2.2.2 Named Entity Recognition

To extract useful knowledge from the clinical and biomedical literature, it is essential to be able to recognize names of clinical entities, such as diseases, drugs, procedures and anatomical sites.

NER is a subtask of information extraction and refers to the task of recognizing expressions denoting entities, the Named Entity (NE), in natural language documents and usually the classification of these NEs into particular categories.

In recent years, the recognition of NEs in the biomedical scientific literature has become the focus of many research. A large number of systems have been built to recognise, classify and map biomedical entities to ontologies. On the other side, only some work have been reported in clinical NER (Jimeno *et al.*, 2008; Ogren *et al.*, 2008; Roberts *et al.*, 2007). NER from scientific articles and newswire text has achieved high performance, however, applying such methods to clinical notes, written by clinicians, pose additional challenges, as this type of text is often less structured, ungrammatical and pregnant with idiosyncratic abbreviated symbolic expressions.

In general, several approaches have been followed to NER in the clinical literature. These can be roughly divided into the following four groups:

- Dictionary-based approaches that try to find names of the well-known nomenclatures in texts;
- Rule-based approaches that manually or automatically construct rules and patterns to directly match them to candidate NE in the texts;
- Machine learning approaches using techniques, such as Hidden Markov Models and Support Vector Machines, to develop models for NER;
- Hybrid approaches that merge two or more of the above approaches, mostly in a sequential way, to deal with different aspects of NER.

Clinical NER systems are, generally, rule or pattern based and the rules or patterns applied by the systems may not be generalisable due to the specific writing style of individual

clinicians or of the area in analysis. Machine learning approaches are not fully advanced in clinical NER due to a lack of available training data.

The remaining of this section is mostly concerned with NER from Portuguese documents and will, therefore, concentrate on the efforts undertaken to create and evaluate NER systems for Portuguese and, particularly, on the shared evaluation contest on NER in Portuguese general domain texts, developed by Linguatca.

2.2.2.1 HAREM II

Avaliação de Reconhecimento de Entidades Mencionadas (HAREM), is the name of a NER evaluation contest for Portuguese, organized by Linguatca (Linguatca, 2008), a project devoted to the promotion of the computational processing of Portuguese. HAREM has had two editions. The *First HAREM* (Santos and Cardoso, 2007) started in September 2004, and comprised two evaluation events. It officially ended at the time of the First HAREM Workshop in Porto in July, 2006. The second edition of HAREM, *Second HAREM* (Mota and Santos, 2008a), took place between September 2007 and September 2008, and officially ended in the Second HAREM Workshop, at the University of Aveiro.

REMMIX, the named entity recognizer of the system presented in this thesis, participated in this second edition², as described in Chapter 5.

Some of HAREM's distinguishing features are described by Santos *et al.* (2006) and include:

- Separation between properly identifying and classifying NEs;
- Taking into account vagueness and indeterminacy, both while building the golden resource and when evaluating the systems;
- Allowing the choice of a subset of semantic categories for evaluation;
- Providing meta-information associated with the texts, to allow investigation on genre and Portuguese variety.

The Second HAREM started in September 2007 and had the submission period between the 14th and the 28th of April, 2008. The 10 participant systems had 48 hours to tag a collection of 1,040 documents (ca. 670,610 words), from which a smaller part, the Golden Collection of 129 documents, had been manually annotated by the organization, according to detailed guidelines discussed with the participants.

²At the time of its participation the component was named REMMA.

As for training material, six fully annotated texts, ca. 1,500 words were made available to participants, together with the golden collections from the First HAREM (ca. 140,000 words) corresponding to 9,000 NEs from 257 different documents.

HAREM's identification task aimed at evaluating systems with respect to their ability to correctly limit the bounds of a NE, irrespectively of its semantic classification. The semantic classification task pretended the assignment of a semantic category and type to every identified NE from a predefined set of 10 categories, 42 and types and 21 subtypes. Figure 2.3 summarizes HAREM's categories and types.

Figure 2.3: Tree of the categories and types considered in the Second HAREM contest (adapted from (Carvalho *et al.*, 2008)).

One of the distinguishing features of HAREM is its semantic model. In HAREM a NE is classified based on the context, i.e., going beyond its dictionary meaning. This feature is exemplified in the following cases expounded in (Freitas *et al.*, 2010):

- (1) A morte é reportada no Diário de Notícias do dia
(*'The death is annouced in Diário de Notícias'*)
- (2) A diferença entre o “Jornal de Notícias” e o “Diário de Notícias”

(*'The difference between Jornal de Notícias and Diário de Notícias'*)

- (3) O seu pai era funcionário público do Ministério da Justiça e crítico musical do 'Diário de Notícias'
(*'His father was an employee of the Ministry of the Justice and a music reviewer for diário de Notícias'*)
- (4) ...foi fotografado pelo Diário de Notícias (DN) a fumar uma cigarrilha ...
(*'had a picture taken by Diário de Notícias smoking a cigarette'*)

The examples 1 to 4 show that the reference to a name such as *Diário de Notícias* or *Jornal de Notícias* can be understood as a place, as an object, as a (private) organization or as a person or group of people standing for their role as interviewers or recipients of information.

So, instead of classifying the instances of the NE as newspaper or mass media, its dictionary meaning, HAREM required their meaning in context.

HAREM guidelines defined a formal criteria in the identification of NEs: the requirement of a capitalized first letter, except for the entities of category TEMPO. The guidelines also required that lowercase expressions that clearly belong to the NE should be identified. The incorrect identification of a NE could compromise its classification. The following example illustrate the described situation:

- (5) ...ministro da Administração Interna ...
(*...minister of Internal Administration ...*)
- CORRECT:** <EM ID=ex CATEG='PESSOA' TIPO='CARGO'> ministro da
Administração Interna
- INCORRECT:** ministro da <EM ID=ex CATEG='ABSTRACÇÃO'
TIPO='DISCIPLINA'>Administração Interna

Another feature of HAREM's annotation syntax is the classification of ranges of values, which should be classified as an unique NE, as in the following example. The first part of the example presents the original text and annotations in Portuguese. The second part its English translation.

- (6) Ele saltou <EM ID= CATEG="VALOR" TIPO="QUANTIDADE">entre 7 a 10 metros na sua fuga.
- He jumped <EM ID= CATEG="VALUE" TIPO="QUANTITY">from 7 to 10 meters in his escape.

The Second HAREM guidelines also allowed NEs to receive more than one tag, whenever the context where it occurs does not allow deciding for only one category. This vagueness is illustrated in the example 7:

- (7) A Administração Bush identifica-se com a Justiça Divina
(*'Bush Administration takes the role of Divine Providence'*)

where the entity *Bush Administration* can be interpreted as both a group of people (PERSON GROUP OF MEMBERS) and as an organization (ORGANIZATION ADMINISTRATION).

Second HAREM added also the possibility of systematic annotation of embedded NEs that take part of larger entities, through the alternatives (ALT) mechanism. For example, the analysis of the sentence:

- (8) Quantos atletas participaram nos Jogos Olímpicos de Barcelona?
(*'How many athletes participated in the Barcelona Olympic Games?'*)

originates two alternative annotations:

- the whole entity *Barcelona Olympic Games*, an EVENT,
- the embedded entities:
 1. *Barcelona* (a LOCAL);
 2. *Olympic Games* (an EVENT).

So, instead of deciding arbitrarily for the widest possible NE, HAREM allowed for the possibility of classifying both as possible correct analyses, using ALT in the output. Example 9 shows the desired output:

- (9) <ALT><Jogos Olímpicos de Barcelona>
|
<Jogos Olímpicos> de <Barcelona></ALT>

To handle all the described features, an innovative evaluation setup had to be defined. Participant systems were allowed to choose a subset of categories in which they wanted to be evaluated. Given so, for each submitted run, HAREM analyzed two different scenarios:

total taking into account all categories defined in HAREM;

selective taking into account a subset of the categories, namely those which the system is tuned to recognize.

To deal with alternative annotations (ALT) two different evaluation modes were also offered:

strict taking into account all the possible alternatives on a text segment, having each alternative a weight equal to the inverse of the number of alternatives in that segment;

relaxed taking into account the element of ALT which maximizes the system classification.

Ten systems participated in the main track of the contest, from which only two, Priberam (Amaral *et al.*, 2008) and REMBRANDT (Cardoso, 2008), recognized the complete set of categories, types and subtypes; all other systems opted for different subsets of the classification tree.

Of the ten systems, only one (R3M (Mota, 2008)) adopted a machine learning approach (specifically, co-training). The others relied on hand-coded rules in combination with dictionaries, gazetteers, and ontologies. Two of them (REMBRANDT and REMMIX³ (Ferreira *et al.*, 2008b)) made use of the Portuguese Wikipedia, in different ways. This shows that the community dedicated to NER in Portuguese hasn't embraced machine learning techniques, contrarily to the situation for English.

The best performing system was a commercial product (from Priberam), which had a very close performance to REMBRANDT's best run. Both achieved an F-measure of approximately 60%. The comparison of the results of the remaining systems is not as straightforward due to the different selective scenarios they participated on.

2.3 Clinical Knowledge Representation

The fast growing amount of information in biology and medicine makes it difficult for researchers to remain up to date with the current biomedical knowledge and to make sense of the massive amounts of online information. The task of managing terabytes of semantic content created the need of more sophisticated techniques of intelligent information and knowledge management. An outcome of this effort is the increasing collection of semantic reference systems often characterized as vocabularies, thesauri, terminologies and ontologies (Freitas *et al.*, 2009; Rubin *et al.*, 2008).

Lexical, terminological and ontological resources are frequently described as enabling resources in MLP systems. These resources are used to support tasks such as NER and Relation Extraction.

³named REMMA at the time of the contest

2.3.1 Knowledge Representation and Ontologies

Knowledge Representation (KR) is an old field of Artificial Intelligence whose main goal is to study how knowledge about the world can be represented and what kinds of reasoning can be done with that knowledge. KR can be viewed as a set of commitments, i.e., as an answer to the question “*In what terms should I think about the world?*” (Davis *et al.*, 1993).

Knowledge-based systems have a computational model of some domain of interest in which symbols serve as surrogates for real world domain artifacts, such as physical objects, events, relationships, etc. (Sowa, 2000). The domain of interest can cover any part of the real world or any hypothetical system about which one desires to represent knowledge for computational purposes (Grimm *et al.*, 2007).

Knowledge representation has several formalisms and appears in different forms, the most prevalent of which are based on semantic networks, rules and logic (Grimm *et al.*, 2007). The last is used to realise a precise semantic interpretation for both of the other forms. By providing formal semantics for knowledge representation languages, logic-based knowledge formalisms as, for instance, Description Logics (Baader *et al.*, 2003), lay the basis for automated deduction.

2.3.1.1 Ontologies

Motivated by the vision of the Semantic Web (Berners-Lee *et al.*, 2001), the term *ontology* has become one of the most fashionable terms in Computer Science in the last years. Ontologies are used to describe domains in detail and to employ these descriptions in many types of applications, ranging from NLP to logic reasoning and decision support systems.

The word ontology was taken from Philosophy, where it means a systematic explanation of being. In the last decade, this word has become extremely relevant for the Knowledge Engineering community.

There are several definitions for ontology. Neches *et al.* (1991, page 40); Gruber (1993b, page 199); Borst (1997, page 12); Studer *et al.* (1998, page 185); Bernaras *et al.* (1996, page 298); Uschold *et al.* (1999, page 11-2) present some of these definitions and how they have changed and evolved over the years. Despite the variety, these different definitions provide complementary points of view of the same reality. Some authors provide definitions that are independent of the processes followed to build the ontology and of its use in applications, while other definitions are influenced by its development process.

A definition that reflects the consensus among the ontology community is that ontologies aim to capture *consensual* knowledge in a generic way, and that they may be reused and shared across software applications and by groups of people. They are usually built cooperatively by

different groups of people in different locations (Gómez-Pérez *et al.*, 2004).

This thesis follows the the distinction proposed by Guarino and Giaretta (1995) to use the words “Ontology” (with capital “o”) and “ontology” to refer to the Philosophical and Knowledge Engineering senses, respectively.

These knowledge resources are widely used in applications related to knowledge management and NLP. Several benefits have been already related with the use of ontologies in such applications, as the “higher” level of reuse than is usually the case in software engineering. It also enables the developer to reuse and share application domain knowledge using a common vocabulary across heterogeneous software platforms (Guarino, 1998; Gómez-Pérez *et al.*, 2004).

Ontology Representation Formalisms: A formal language used for an ontology is called an *Ontology Representation Model*. RDFS/OWL is one of the most common ontology representation model, accepted and standardized by the World Wide Web Consortium (W3C) (W3C, 2010). It is based on a model called Resource Description Framework (RDF) (Lassila and Swick, 1998) recommended by W3C for representing information about resources in the World Wide Web.

Figure 2.4: RDF triple.

The RDF model is often called a triple because it has three parts: subject, predicate and object (Figure 2.4). Each triple is an RDF statement. The RDF Schema (Brickley and Guha, 2004) provides the vocabulary for the RDF statements, and allows, for instance, the creation of classes of data. A class is a group of things with common characteristics. For example, after specifying the classes *Episode*, *Patient* and *Physician*, an RDF statement can specify, for example, that the physician **Telmo Oliveira** is an instance of the class *Physician*, and by another statement that the episode number **12345** is an instance of the class *Episode*.

RDF Schema is, however, a weak ontology language in the sense that it offers only the modeling concepts class, subclass, relations, property, domain and range restrictions. There are many modeling primitives that are useful in modelling documents in the health care sector and other areas which are missing in the RDF Schema. For example, by using RDF Schema we can not specify that the classes (e.g., *Physician* and *Patient*) are not necessary disjoint, or even to build new classes by setting operations as, for example, when declaring that the class *Doctor* is the union of the classes *Physician* and *Dentist*. The Web Ontology Language (OWL)

was developed to overcome this and other limitations.

OWL (Patel-Schneider *et al.*, 2004) has been standardised by the W3C as a language for semantic annotation of web content and is widely accepted within the Semantic Web community. An important issue for the design of OWL was the trade-off between the expressivity of the language and the scalability of reasoning. To this purpose, OWL comes with three different levels, namely OWL Lite, OWL DL and OWL Full, reflecting different degrees of expressiveness (Patel-Schneider *et al.*, 2004):

OWL Lite supports those users primarily needing a classification hierarchy and simple constraints. For example, it supports cardinality constraints but only in the values 0 or 1. Lite is not widely used;

OWL DL is designed to provide the maximum expressiveness possible while retaining computational completeness and the availability of practical reasoning algorithms. OWL DL is named as so due to its correspondence with Description Logic, a field of research that has studied the logics that form the formal foundation of OWL;

OWL Full OWL Full is based on a different semantics from OWL Lite or OWL DL and was designed to preserve some compatibility with RDF Schema. For example, in OWL Full a class can be treated simultaneously as a collection of individuals and as an individual in its own right, this is not permitted in OWL DL. However, due to this level of expressiveness, it is unlikely that any reasoning software will be able to support complete reasoning for OWL Full.

The OWL working group has recently produced a W3C Recommendation for a new version of OWL which adds features to this 2004 version, while remaining compatible. More information about this new version is presented by W3C OWL Working Group (2009).

SPARQL is an RDF query language standardized by the RDF Data Access working group of the W3C and is considered a key semantic web technology. The syntax of SPARQL resembles SQL, but SPARQL is far more powerful enabling queries spanning multiple data sources containing heterogeneous semi-structured data. SPARQL Protocol an RDF Query Language (SPARQL) is further described by Prud'hommeaux and Seaborne (2008).

2.3.2 Terminological and Ontological Resources in the Medical Domain

Most current development in Medical Informatics aims at systems in which information is shared for a variety of purposes. For the last years, problems of standardising medical language and terminology have been a major concern in the area. The main purpose of clinical

terminologies and ontologies is to support clinical software as EHR systems, computer assisted decision support, quality assurance and information management systems.

2.3.2.1 Terminologies vs. ontologies

Terminology is a relatively young, interdisciplinary research field which has its roots in linguistics and cognitive science. Sager (1994) defines Terminology as a theory concerned with the aspects of the nature and the functions of language which permit the efficient representation and transmission of items of knowledge in all its complexity of concepts and conceptual relationships.

Medicine is characterized by a wealth of, so called, terminologies. These terminologies are generally built to serve well-defined purposes like document retrieval, resource annotation or health services billing. Biomedical terminologies do not use formal and well-defined descriptions; they rather define the terms (if ever) by human language expressions, and express the associations between terms by informal relations. Words are generally the basic blocks of terminologies, often organized in hierarchies that relate their meanings in terms of synonymy (same meaning), hyperonymy (broader meaning) and hyponymy (narrower meaning).

Although terminologies successfully represent abstract meaning they are not precise and expressive enough for more knowledge-intensive applications. In the presence of a use case that demands precise relations between terms, a language-centered resource is not expressive enough.

Ontologies, however, can represent a domain in a form that computers can interpret according to the semantics of the definitions instead of employing only terms or semantic identifiers. A system can, in this case, check whether some interpretation is correct or if a given statement is true according to some ontology.

Clinical ontologies are developed to solve problems such as the demand for the reuse and share patient data, the transmission of these data, or the need of semantic-based criteria for statistical purposes. The unambiguous communication of complex and detailed medical concepts is a crucial feature in current medical systems.

In practice, however, since most terminologies have some degree of organization, and many ontologies also collect names for their entities, the distinction between ontological and terminological resources is somewhat arbitrary.

Several terminological and ontological resources have been developed in the last years in the medical area. The following paragraphs summarize their main features.

2.3.2.2 The International Classification of Diseases

The International Classification of Diseases (ICD) (World Health Organization, 2010a) has its origins in the 1850s. The first edition, known as the International List of Causes of Death, was adopted by the International Statistical Institute in 1893. The World Health Organization (WHO) took over the responsibility for ICD at its creation in 1948 when the sixth revision was published. This was, for many years, the only medical terminological resource. Several editions were created and maintained by WHO until its current edition, the tenth edition. ICD-10 is currently translated into 42 languages and provides about 13,000 classes for the encoding of diseases and reasons of encounter.

The ICD is the international standard diagnostic classification for all general epidemiological states and many health management purposes and clinical use. These include the analysis of the general health situation of population groups and monitoring of the incidence, prevalence of diseases, reimbursement, resource allocation, quality and guidelines.

ICD classes are hierarchically arranged with the *is-a* (subclass) relation, expressing that each member of a class is also member of any parent class, like in a tree-like hierarchy. ICD's scope extends the realm of diseases as it also includes injuries and external causes of health problems, signs and symptoms, and any kind of conditions that justify the encounter with health professionals.

In many countries, including Portugal, the ICD is also employed as a basis for Diagnosis Related Groups (DRGs) used for billing. DRGs group patients that are clinically similar and are therefore expected to use the same health care resources.

2.3.2.3 The Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) (Nelson and Schulman, 2007; Medical Subject Headings, 2010), edited and maintained by the USNLM, is a controlled vocabulary used for indexing, cataloging and searching biomedical and health-related information and documents. Its current version (2011) consists of 26,142 descriptors (main headings), organized in 16 hierarchies (Anatomy, Organisms and Diseases, among others). In contrast to the tree-like hierarchy of ICD, MeSH headings are placed in multiple hierarchies. The hierarchical order of MeSH is based on the principle that all documents indexed by a given heading are also relevant for any parent descriptor.

MeSH is available in 41 languages, including Portuguese in its Brazilian variant. *Descritores em Ciencias da Saúde (DeCS)*, the health sciences descriptors, were developed from MeSH and contain a total 30,369 descriptors, of which 25,671 come from MeSH, and 4,698 are exclusively in DeCS. This resource provides the translated terms in three different

languages: Portuguese, Spanish and English.

2.3.2.4 SNOMED-CT

The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT), developed by the College of American Pathologists, is a comprehensive biomedical terminology created to cover the whole patient record. It comprises Body Structures, Procedures and relevant health-related aspects, including also social aspects (Spackman, 2004).

Since April 2007, SNOMED CT is owned, maintained, and distributed by the International Health Terminology Standards Development Organization (IHTSDO) (International Health Terminology Standards Development Organization, 2010), a not-for-profit association based in Denmark.

SNOMED-CT products and services are open for researchers but its use for clinical coding or other commercial usages is restricted to its licensees. SNOMED-CT is officially available in English and Spanish, while other translations (e.g. Dutch, Danish, Swedish) are currently being created.

This resource is based on multiple *is-a* hierarchies containing about 310,000 nodes. SNOMED-CT nodes, referred to as concepts, denote mostly classes of individual entities (such as Diseases, Procedures, Lab results, Drugs, Geographic entities, etc). SNOMED-CT concepts are uniquely identified by numeric keys, together with their fully specified names. Most SNOMED CT concepts include several synonyms (referred to as “descriptions”) and, in a few cases, also free-form text definitions. Additional attributes are provided, as the SNOMED qualifiers, which define optional refinements for concepts as, for example, laterality for Anatomy or severity for Diseases (Freitas *et al.*, 2009).

One of the most important distinctive criterion of SNOMED-CT is the use of a rich ontology representation language compatible with the Semantic Web standard OWL-DL.

Given the numerous and largely overlapping amount of medical knowledge resources currently available, considerable efforts have been devoted to align these terminologies and ontologies. The most important example of such effort is the UMLS project and, particularly, the Metathesaurus knowledge source.

2.3.2.5 The Unified Medical Language System Metathesaurus

The richest source of biomedical terminologies, thesauri, classification systems and ontologies is constituted by the UMLS Metathesaurus. It has been initiated in 1986 by the USNLM, with the purpose of integrating information from a variety of distinct terminological

sources.

The UMLS Metathesaurus is one of the three knowledge sources developed and distributed by the USNLM, as part of the UMLS project (Nelson *et al.*, 2006). Version 2011AA, the latest available of the Metathesaurus, contains more than 2.2 million concepts and 8.2 million unique concept names, from over 150 controlled vocabularies and classifications (some in multiple languages) used in patient records, administrative health data, bibliographic and full-text databases, and expert systems.

The Metathesaurus also assigns a unique identifier to each concept and establishes new relations between terms from different source vocabularies. Each concept is also categorized with at least one semantic type from the UMLS Semantic Network, also an UMLS knowledge source. Because of its utilization in the work presented in this document, the UMLS project and its knowledge sources are further explored in Section 3.1.2.

2.3.2.6 The Open Biomedical Ontologies Foundry

To prevent the anarchic proliferation of biomedical terminologies and ontologies, by establishing principles for the coordinated development of interoperable resources, the Open Biomedical Ontologies (OBO) Foundry created in 2003 a platform that evolved as a library of online and public-domain biomedical ontologies (Open Biomedical Ontologies, 2010). The OBO Foundry initiative developed a set of shared principles regulating the development of biomedical ontologies (Smith *et al.*, 2007). Currently, more than 50 ontologies are listed in the OBO Foundry.

Table 2.2 resumes some of the key features of the described resources by showing their scope and usage.

The creation of systems capable of accessing texts and databases containing detailed data, able to perform reasoning and, particularly, the development of decision support systems, is heavily dependent on the existence of precise knowledge in a computational mean. However, the field of Knowledge Representation has still to overcome several challenges.

A key issue resides in integration. As the number of biomedical ontologies increases, many applications need to employ more than one ontology, which leads to a series of significant consequences due to knowledge heterogeneity (Freitas *et al.*, 2009). A relevant challenge to integration is how to handle existing biomedical ontologies that contain overlapping information, providing different views on a certain subdomain or covering different domains. To enable ontology integration, plenty of research is taking place. This issue is further discussed in (Freitas *et al.*, 2006).

Table 2.2: Main characteristics of the presented ontologies and terminologies

NAME	SCOPE	APPLICATIONS	URL
ICD	Diseases	Health Statistics, Epidemiology, Health Reporting, Billing	www.who.int/classifications/apps/icd
MeSH	Medicine, Nursing, Dentistry, Veterinary Medicine, Health Care Systems, Preclinical Sciences	Indexing articles from 4.800 of the world's leading biomedical journals for the MEDLINE/PubMED database	www.pubmed.org
SNOMED	Everything encoded in the electronic health record	Information about a patient's medical history, illness, treatments and laboratory results	www.ihtsdo.org
UMLS	Biomedical and health related concepts	scientific literature, guidelines and public health data, natural language processing	www.nlm.nih.gov/research/umls/
OBO	Bioinformatics and molecular Biology	Used as a repository and an unified schema to interoperate biomedical projects	www.obofoundry.org

2.4 Summary

The current chapter starts by introducing the eHealth area and how it is currently applied in Portuguese health institutions and, particularly, in the region of Aveiro. It continues by reviewing the advances in clinical IE from free-text documents. IE is still a relatively new field of research in the clinical domain and presents, currently, a wide range of results. Some reasons are proposed for this difference as the need for more experience, the lack of annotated text corpora, and also the fact that clinical text is simply harder to analyze than biomedical literature or newswires. During the last several years, performance has gradually improved. However, the creation of annotated clinical text corpora is still one of the main challenges for the future of this field, as the effort required is significant and patient data confidentiality issues hamper access to data.

It is consensual that the potential uses of IE from clinical text are numerous and far-reaching. However, current applications are rarely applied outside of the laboratories they have been developed in. Improvements in system performance will subsequently enhance the acceptance and usage of IE in concrete clinical and biomedical research contexts.

The current chapter also described a set of medical terminologies and ontologies which represent a variety of semantic standards in medicine. It presented an overview of the substantial efforts being carried to describe the entities they denote, in order to support querying and knowledge processing, in general and specific applications.

Resources and Tools

The creation of systems capable of accessing texts and databases containing detailed data, capable of performing reasoning and, particularly, the development of decision support systems is heavily dependent on the existence of precise knowledge in a machine-readable mean.

The present chapter describes the resources and tools used in the development of the medical information extraction system. The knowledge sources used in the task of NER, namely Wikipedia and UMLS, are first presented and followed by a detailed characterization of the corpus of discharge summaries used in the development and evaluation of the system. The chapter concludes with an introduction to the processing tools used, ranging from the NLP infrastructure chosen for the development of the MLP system, to the tools used for manual annotation.

3.1 Resources

It is often claimed that IE, particularly NER systems need extensive gazetteers - lists of names of people, organisations, locations, and other named entities. The compilation of such gazetteers is sometimes mentioned as a bottleneck to the design of NER systems due to the time and human effort needed to create and maintain high-quality resources as these (Mikheev *et al.*, 1999). This limitation was overcome in MEDINX with the use of external knowledge sources, such as Wikipedia and UMLS, in order to improve semantic category label extraction.

3.1.1 Wikipedia

We have recently seen a rapid and successful growth of Wikipedia (Wikipedia, 2011), a multi-lingual web-based, free-content encyclopedia which is updated continuously in a

collaborative way.

Wikipedia has now more than 3,2500,000 articles in the English version and more than 677,000 articles in the current Portuguese version (March 2011). Since Wikipedia aims to be an encyclopedia, most articles are about NEs and are more structured than raw texts. Although it cannot be used as a gazetteer directly, since it is not intended as a machine readable resource, extracting knowledge from Wikipedia is easier than from raw texts or from usual Web texts because of its structure.

Other reasons that determined the use this encyclopedia are:

- Its content has free license, i.e., it will always be available for research without restrictions and without needing to acquire any license.
- It is a general knowledge resource. Thus, it can be used to extract information for open domain systems.
- Its data has some degree of formality and structure (e.g. categories) which helps to process it.
- It is continuously updated.

3.1.1.1 Basic Structure

An article in Wikipedia is identified by a unique name which can be obtained by concatenating the words in the article title with an underscore “_”, being the first word always capitalized. For example, the unique name for the article *University of Aveiro* is `University_of_Aveiro`.

Typically, the title is the most common name for the entity described in the article. When the name is ambiguous, it is further qualified with a parenthetical expression. For instance, the article on *Stroke*, the set of actions to propel the boat while rowing, has the title `Stroke_(rowing)`.

In general, there is a many-to-many correspondence between names and entities. This relation is captured in Wikipedia through *redirect* and *disambiguation* pages. However, Wikipedia articles have many other useful structures for knowledge extraction such as headings, lists, internal links, categories and tables. Some of those, the most relevant to the system presented, are described next. Table 3.1 summarizes and exemplifies the utility of such structures. The examples presented in this table will be further examined in the following paragraphs.

Table 3.1: Examples of Wikipedia titles, aliases and categories from the disambiguation page of the acronym “EHR”.

TITLE	REDIRECT	DISAMBIG	CATEGORIES
Electronic Health Record	<i>none</i>	EHR	Medical informatics; Health standards; International standards
The English Historical Review	<i>none</i>	EHR	History journals; Publications established in 1886; English-language journals; Oxford University Press academic journals; Humanities journal stubs
European Hit Radio	<i>none</i>	EHR	Radio stations in Latvia; Radio stations in Lithuania; Latvia stubs; Lithuania stubs; European radio station stubs
Ehr, Germany	<i>none</i>	EHR	Municipalities in Rhineland-Palatinate; Rhein-Lahn-Kreis geography stubs
VistA	VistA EHR	EHR	Health software; United States Department of Veterans Affairs; Public health; Medical informatics; Health care informatics; Open source health care resources; Public domain software
THIRRA (EHR)	THIRRA EHR	EHR	Medical informatics; Open source health care resources; Medical equipment stubs
List of The Adventures of Tintin characters	Bab El Ehr	EHR	Tintin characters

3.1.1.2 Redirection

A *redirect page* exists for each alternative name that can be used to refer to an entity in Wikipedia. Redirections are marked up as **#REDIRECT** `[[A B C]]` in source files, where `[[...]]` is a syntax for a link to another article in Wikipedia (internal links). If the source file has such description, users are automatically redirected to the article specified by the entity name in the brackets (`A_B_C` for the above example). Redirections are used for several purposes regarding ambiguity. For example, they are used for acronym resolution such as from *EHR* to *Electronic Health Record*. They are also used in the context of more difficult disambiguations.

3.1.1.3 Disambiguation pages

Some authors make a *disambiguation* page for an ambiguous entity name¹. A disambiguation page typically enumerates possible articles for that name. For example, the disambiguation page for the name *EHR* lists 7 associated entities. Therefore, besides the non-ambiguous names that come from redirect pages, additional aliases can be found by looking

¹a name that can be used to refer to several entities (i.e., articles in Wikipedia)

for all disambiguation pages that list a particular Wikipedia entity.

3.1.1.4 Categories

Every article in Wikipedia is required to have at least one category. As shown in Table 3.1, *Electronic Health Record* is associated with three categories: *Medical informatics*, *Health standards* and *International standards*. Categories allow articles to be placed in one or more topics. These topics can be further categorized by associating them with one or more parent categories.

3.1.1.5 Hyperlinks

Articles in Wikipedia often contain mentions of entities that already have a corresponding article. When contributing authors mention an existing Wikipedia entity inside an article, they are required to link at least its first mention to the corresponding article, by using *links* or *piped links*. Both types of links are exemplified in the following source code of a sentence from the article on Ehr, a place in Germany:

```
(10) Ehr is a municipality in the [[Rhein-Lahn-Kreis|district of
      Rhein-Lahn]] , in [[Rhineland-Palatinate]], in western [[Germany]].
```

The string from the second link (*Rhineland-Palatinate*) denotes the title of the reference article. The same string is used in the display version. If the author wants another string displayed (e.g., *district of Rhein-Lahn* instead of *Rhein-Lahn-Kreis*), then the alternative string is included in a piped link, after the title string. Consequently, the display string for the aforementioned example is:

```
(11) Ehr is a municipality in the district of Rhein-Lahn, in Rhineland-Palatinate, in western
      Germany.
```

3.1.1.6 Data used in the NER task

Snapshots of the entire contents of Wikipedia are provided in XML format for each language version. The Portuguese version of February 2008, which includes 1,290,836 pages, was used as a knowledge source to the NER task. The data was exported into an SQL database and used in the work described in the next chapters.

3.1.2 Unified Medical Language System

In 1986, the USNLM, initiated a long-term research and development effort known as the UMLS. Anticipating increasing amounts of biomedical information available in electronic form, they foresaw the need of facilitating the development of advanced information systems that could retrieve and integrate information from a variety of disparate information sources, including bibliographic databases, patient record systems, factual databanks, and knowledge bases (Nelson *et al.*, 2006).

A major barrier to effective retrieval and integration of information from multiple sources was identified: the *naming problem* or the variety of different ways that the same concepts are expressed in different information sources, and by different information seekers.

The UMLS (USNLM, 2008) is currently a compendium of many controlled vocabularies in biomedical sciences. It provides a mapping structure among these vocabularies and, thus, allows one to translate among the various terminology systems. It may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts.

There are three major UMLS knowledge sources:

1. Metathesaurus, the core database of the UMLS is a collection of concepts and terms from the various controlled vocabularies and their relationships;
2. Semantic Network, a set of categories and relations that are being used to classify and relate the entries in the Metathesaurus;
3. SPECIALIST Lexicon, a database of lexicographic information which contains syntactic, morphological and orthographic information for the biomedical area, as well as, for common words in the English language.

The **Metathesaurus** is a large, multi-purpose and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names and the relations among them. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and on basic, clinical and health services research. The Metathesaurus is organized by concept or meaning. Its purpose is to link alternative names and views of the same concept together and to identify useful relations between different concepts.

Four of the 150 UMLS source vocabularies contain a Portuguese translation. These are the *Medical Dictionary for Regulatory Activities Terminology*, the *International Classification of Primary Care*, the *WHO Adverse Drug Reaction Terminology* and DeCS, the Portuguese

translation of MeSH. On its 2010 version, more than 74,000 concepts and 150,000 unique concept names have been translated to Portuguese.

The UMLS **Semantic Network** is a top-level ontology for biomedical concepts and relations among these concepts. Figure 3.1 shows a partial view of this top-level ontology. The Semantic Network was not derived from the biomedical sources integrated in UMLS but created as a part of UMLS with the aim of providing a consistent structure or categorization, in which the Metathesaurus concepts are included. Each Metathesaurus concept is attached to a concept or concepts of the Semantic Network. The Semantic Network contains currently 134 top-level concepts and 54 relationships among them.

Figure 3.1: Partial view of the UMLS Semantic Network.

The **SPECIALIST Lexicon** has been developed to provide the lexical information needed for the SPECIALIST Natural Language Processing System. It is intended to be a general English lexicon that includes many biomedical terms. Coverage includes, both, commonly occurring English words and biomedical vocabulary. The lexicon entry for each word or term records the syntactic, morphological and orthographic information needed by the SPECIALIST NLP System.

Because it integrates a large number of terminologies, the UMLS Metathesaurus is the

terminological system most frequently used in the analysis of biomedical and clinical texts. Today more than 1,000 individuals and institutions worldwide license the UMLS resources which are free-of-charge (USNLM, 2008). The majority of the licensees use one or more of the UMLS components in information systems, often in creative and innovative undertakings. The USNLM itself uses UMLS components to enhance retrieval from several of its information services, including the MEDLINE database available via PubMed (USNLM, 2011).

3.1.2.1 Data used in the NER task

MetamorphoSys is the UMLS installation wizard and Metathesaurus customization tool included in each UMLS release. This tool allows to install one or more of the UMLS Knowledge Sources. In MEDINX, the UMLS 2010 release was used as knowledge source to the several IE related tasks. The four UMLS vocabularies containing Portuguese translated terms were installed. This subset of the UMLS Metathesaurus was then loaded into a MySQL database. The UMLS MySQL database created in MEDINX has a total of 46 tables and 580Mb of information. The UMLS database table containing the main Metathesaurus information is MRCONSO which contains all the concept names and the sources of this resource, i.e., every string or concept name included in the Metathesaurus also appears in this table together with the connections to its language, source vocabulary and its concept identifier.

The content of this table was one of the main sources of knowledge used in MedInX. The most relevant columns of the table are described in Table 3.2.

Table 3.2: MRCONSO columns used in MEDINX.

COLUMN	DESCRIPTION
CUI	Unique identifier for a concept
LAT	Language of a term
SAB	Abbreviated source name
TTY	Abbreviation for term type in the source vocabulary, for example, PT (Preferred Term) or HT (Hierarchical Term)

3.1.3 Corpus of Patient Discharge Letters

The use of automatic IE tools depends heavily on the availability of corpora for development and testing. Also, the use of authentic health records is a prerequisite for developing MEDINX, a MLP system intended to be used in real applications. However, the narratives developed during clinical and hospital encounter are rarely made available outside the hospital setting due mainly to legal and ethical aspects. Even when referring to documents

written in English, the clinical IE community hasn't yet developed a gold-standard corpus of annotated clinical reports which can be used as a shared standard for evaluating automatic knowledge extraction systems. The same is valid when referring to Portuguese discharge letters.

Therefore, the development of a corpus of authentic discharge letters, written in Portuguese and originated in Portuguese health care institutions, was of extreme importance for the development, validation and evaluation of MEDINX.

The methods used to create the MEDINX corpus and its content are the focus of the current section. Some legal and ethical considerations regarding storage and usage of such documents are first discussed, followed by the analysis of the structure and content of the discharge letters. The creation of the gold-standard collection of the corpus and the development and test sets, conclude the current section.

3.1.3.1 Legal and ethical aspects

Authentic health records are a prerequisite for this thesis; without them developing intelligent, domain-oriented HLT solutions and evaluating their quality is impossible. Inherent in such research, however, is the use of data without the patient's consent. Thus, storing and using this data requires careful consideration and compliance with legal and ethical principles.

In this work we followed the security procedures designed by Berman (2002) for accessing electronic health records for health care. Particularly, and in order to ensure patient confidentiality, the following practices were adopted:

Permits The PDLs used in the development and evaluation of MEDINX have been collected without patients' consent, as they were obtained through the access to the EMR system being used in HIP, the RTS (see Section 2.1). Instead, we obtained proper permits from the HIP and University of Aveiro authorities having the right to acknowledge the permissions. Also, the RTS has already been given a formal consent regarding the use of the data obtained with its utilization in HIP in the development of research and development projects in the University of Aveiro. This consent has also been approved by the National Committee for Data Protection;

Research material Personal identifying data is not required for the research or for the verification of the results achieved. Actually, the RTS separates all personal identifying data, as the name, address, birthday, national health system and social security numbers, in their structured files and so it does not need to be added in the discharge summary. Given so, the documents used in this project do not contain any personal data and are therefore anonymized;

Risks Despite deliberately conforming to good scientific practice, the risk of violating patient confidentiality still remains. The main risks in this thesis arise from the possibility of deducing the patient identity by combining gender, date and location and other similar details, often present in the documents. However, no data is used in this thesis to recognize the identity of individual people and special care has been paid to data protection.

3.1.3.2 Patient Discharge Letters selection

During the hospitalization of a patient several data is produced in textual form, as for example, the admission document, the examination reports and the daily notes. Figure 3.2 presents the common patient information flow during its admission period. The focus of MEDINX is on narratives, specifically on the summaries produced on the moment of discharge, the PDLs. The aim of these documents is to transfer summarized information from the hospital setting to other wards, normally, to the general practitioner, in order to assure the continuity of care. This type of narratives was selected because they cover the whole inpatient period and summarize the main occurrences during that period, as well as, the clinical history of the patient.

Figure 3.2: Clinical Information Flow

MEDINX focuses its analysis on the Hypertension domain. Hypertension, also called High Blood Pressure, is a condition in which the blood pressure, in either arteries or veins, is abnormally high. High blood pressure is not only a serious condition by itself, but it is also the leading cause of cerebrovascular diseases, the most common cause of death in Europe and, particularly, in Portugal (National Institute of Statistics, 2009).

MEDINX can contribute to the worldwide efforts on the reduction of hypertension

incidence and to the consequent reduction of cerebrovascular diseases through the extraction of detailed knowledge from the discharge letters and consequently improve the access to quality information about the patient health.

The first step in creating the MEDINX corpus of PDLs was gathering a list of hospital episodes to which had been assigned a DRG code for the diagnoses of cerebrovascular disease.

The DRGs (see Section 2.3.2) are a patient classification schema which provides a manner of relating the type of patients a hospital treats to the costs incurred by the hospital. Currently, the DRGs are the main classification system used in Portuguese hospitals.

The corpus consists, thus, of 915 discharge letters of patients admitted with at least one of the following DRG codes:

- 14 - Specific Cerebrovascular Disorders except Transient Ischemic attack;
- 15 - Transient Ischemic Attack & Precerebral occlusions;
- 16 - Nonspecific Cerebrovascular disorders with Complicating Condition (CC);
- 17 - Nonspecific Cerebrovascular disorders without CC;
- 532 - Transient Ischemic attack, Cerebral occlusions, Convulsions and Headache.

Particularly, the corpus contains **825 PDLs** to which has been assigned the **DRG code 14**, **66 PDLs** with **DRG 15**, **two** with **DRG 16**, **six** with **DRG 17** and **16 PDLs** with **DRG 536**.

Table 3.3 presents some of the ICD-9-CM codes for the principal diagnoses related with the considered DRGs.

3.1.3.3 Patient Discharge Letters

The structure and content of the discharge letters are two of the main aspects of the documents. Figure 3.3 presents a style preserving illustration of a PDL, which shows the discharge summaries default headings. The headings direct the use of a specific **structure**. Each component of this structure will be henceforth referred to as a *Segment*.

The default segments in discharge documents are:

- Admission Reason;
- Clinical History;
- Physical Examination;

Table 3.3: Principal diagnoses for cerebrovascular diseases.

ICD-9-CM	DESCRIPTION
430	Subarachnoid hemorrhage
431	Intracerebral hemorrhage
432.0	Nontraumatic extradural hemorrhage
432.1	Subdural hemorrhage
432.9	Unspecified intracranial hemorrhage
434.0	Cerebral thrombosis
434.1	Cerebral embolism
434.9	Cerebral artery occlusion, unspecified
436	Acute, but ill-defined, cerebrovascular disease
437.3	Cerebral aneurysm, nonruptured
784.3	Aphasia
433.0	Basilar artery occlusion
433.1	Carotid artery occlusion
433.2	Vertebral artery occlusion
433.3	Multiple and bilateral precerebral occlusion
433.8	Other specified precerebral artery occlusion
433.9	Unspecified precerebral artery occlusion
435.0	Basilar artery syndrome
435.1	Vertebral artery syndrome
435.2	Subclavian steal syndrome
435.8	Other specified transient cerebral ischemia
435.9	Unspecified transient cerebral ischemia
348.3	Unspecified encephalopathy
348.8	Other specified brain conditions
348.9	Unspecified brain conditions
349.8	Other specified central nervous system disorder
349.9	Unspecified central nervous system disorder
437.0	Cerebral atherosclerosis
437.1	Unspecified acute cerebrovascular insufficiency
437.7	Transient global amnesia
437.8	Other specified cerebrovascular disease
437.9	Unspecified cerebrovascular disease

- Therapeutics;
- Evolution;
- Destination.

However, the use of this structure is not mandatory and the attending physician can decide on whether to use the proposed structure and on which content to add in the segment. As a result, the content of each segment varies a lot from episode to episode.

Table 3.4 presents the number of segments of the MEDINX corpus of PDLs that were filled with text during the document production, i.e., that were used by the physician while

<p>Motivo Internamento AVC isquémico de repetição; HTA</p> <p>História Clínica Doente de 83 anos, com antec. de HTA e depressão nervosa. Faz uso regular de fármacos que desconhece nomes. Hoje veio transferida do Hospital do Visconde de Salreu por desvio da comissura labial para esquerda. Medicada regular/ com Ogasto; Tenoretic; Micardis e Motillium. A doente refere que cerca das 09H00 teve episódio de disartria acompanhado de desvio da comissura labial a dta, sem alterações da FM ou da sensibilidade. Recorreu ao Hosp de Estarreja tendo sido encaminhada a esta Urgência por hipótese de AVC. Desde a entrada no HVS que refere melhoria progressiva das alterações da fala, sem défices neurológicos de novo. Sem queixas sugestivas de síndrome infecciosa ou de febre.</p> <p>Exame Físico COC, eupneica sem SDR. Apiretica. Chorosa. Corada e hidratada. TA- 193/68 mmHg; spO2 (AA)- 99%; PR- 60/min. AC- irregular, por ES. AP- Mv+ sem RA valorizáveis. Sem edemas dos MI's. ENS: EG- 15; sem lateralização motora; FM preservada; sem alteração da linguagem; pupilas I/R; olhos na linha média; esboço de paresia facial central a Dta.</p> <p>Terapêutica Efectuada Terapêutica efectuada: -asa 100, enoxaparina, esomeprazol, insulina sos, nitratos transdérmicos, soros.</p> <p>Destino Hos. de Salreu</p> <p>Evolução Transferida para o Hosp. de Salreu</p>	<p>Admission Reason Recurrent ischemic stroke, HTA</p> <p>Clinical History 83 years old patient, with history of HTA and clinical depression. Uses drugs regularly of which does not know names. Was transferred today from the Hospital Visconde Salreu by deviation of the left lip. Regularly medicated with Ogasto; Tenoretic; Motillium and Micardis. The patient states that at approximately 09:00 had an episode of dysarthria with deviation of the right lip without changes in UL or sensation. Appealed to the Hospital of Estarreja and was forwarded to this Urgency due to stroke suspicion. Refers progressive improvement of speech changes since he entered the HVS, no new neurological deficits. No complaints suggestive of infectious syndrome or fever.</p> <p>Physical Examination COC, eupneic without RDS. Afebrile. Tearful. Healthy coloring and hydrated. BP-193 / 68 mmHg; spO2 (AA) - 99%, PR-60/min. CA- irregular, due to ES. PA-Bs + without considerable rales. No edema in the LLs. ENS: EG-15, no motor lateralization; MS preserved, without changes in language, PERRLA; eyes in the midline; outline of central facial palsy at the right.</p> <p>Therapeutics Therapeutic: -asa 100, enoxaparin, esomeprazole, sos insulin, transdermal nitrates, salines.</p> <p>Destination Salreu Hos.</p> <p>Evolution Transferred to Salreu Hosp.</p>
---	---

Figure 3.3: Style-preserving illustration of a Patient Discharge Letter. An original document written in Portuguese is presented on the left side of the figure, and on the right side, its English translation.

describing the patient admission period and were not left empty.

Table 3.4: Number of non-empty default segments of the corpus of 915 PDLs.

SEGMENT	NON EMPTY SEGMENTS
Admission Reason	749
Clinical History	862
Physical Examination	632
Therapeutic	727
Evolution	717
Destiny	571

Considering the 915 documents of the corpus, it shows that 18% of the MEDINX discharge documents do not contain information regarding the *Admission Reason*. The same thought can be followed for the remaining segments of the documents, showing that 6% do not

have information regarding the patient *Clinical History*; 31% do not describe the *Physical Examinations* performed during the admission period and 21% do not present the *Therapeutics* carried out. Finally, 38% of the documents do not contain information about the *Destination* of the patient after discharge. One can conclude that the physicians often do not follow the defined normative structure.

Another important aspect of these documents is its content distribution, i.e., the amount of text in the document and its segments. Tables 3.5 to 3.7 present some descriptive statistics of the size, in tokens and sentences, of the corpus and its segments.

Table 3.5: Number of tokens and sentences of the MEDINX corpus.

Segment	Tokens	Tokens/Segment	Sentences	Sentences/Segment
Admission Motive	7,963	11	960	1
Clinical History	100,489	116	4,860	6
Physical Examination	41,964	66	3,548	6
Therapeutic	29,315	40	1,030	1
Evolution	42,291	59	2,485	3
Destiny	21,729	38	1,321	2
Total	243,751	260	14,204	16

Table 3.6: Descriptive statistics for the amount of tokens of discharge documents and its segments.

Segment	Tokens	Tokens/Document	Standard Deviation	Min	Max
Admission Motive	7,963	11	32	2	414
Clinical History	100,489	116	150	3	810
Physical Examination	41,964	66	71	5	563
Therapeutic	29,315	40	42	7	212
Evolution	42,291	59	83	2	539
Destiny	21,729	38	41	2	196
Total	243,751	260	331	2	1,462

The tables help to better understand the distribution of content of the discharge letters. It is possible to observe how data is most frequently distributed in the document. The patient *Clinical History* and the *Physical Examination* are the longest segments of the document. However, and despite the equal average number of sentences, the *Clinical History* is the segment with more content, with a mean value of 116 tokens per document, in contrast with 66 tokens per document in the *Physical Examination* segment. These values seem to indicate the presence of longer sentences while described the past conditions and procedures of the patient, and the use of shorter and direct sentences when listing the examinations performed during the inpatient period. This will be further explored in the next paragraphs, while

analysing the contextual features of the documents.

Table 3.7: Descriptive statistics for the amount of sentences of discharge documents and its segments.

Segment	Sentences	Sentences/Document	Std. Deviation	Min	Max
Admission Motive	960	1	2	1	31
Clinical History	4,860	6	6	1	50
Physical Examination	3,548	6	4	1	25
Therapeutic	1,030	1	1	1	9
Evolution	2,485	3	3	1	27
Destiny	1,321	2	1	1	13
Total	14,204	16	14	1	64

The analysis of the structure and contents of the longest, smallest and average discharge documents can also help understanding the type of data present in these documents.

For instance, in the longest discharge document (1,462 tokens and 64 sentences) only the segments *Admission Reason*, *Clinical History* and *Physical Examination* were used. The writer added, immediately, in the starting segment *Admission Reason* all the significant information about the patient history and its inpatient period. The information present in this segment refers to age, past conditions, current medication, reason for admission and physical examinations results, as body temperature and blood pressure at the time of admission. The segment contains also the transcription of the results of other procedures as the X-ray computed tomography result transcription. The status of the patient at the time of the discharge document production, and the main diagnoses associated with the episode, end the segment. As all the main information concerning the patient hospitalization has been already presented in this segment, the two other filled segments, *Clinical History* and *Physical Examination*, are just an exact copy of the text introduced in the *Admission Reason* segment. Thus, despite the existence of textual data in the referred segments, no new content was added to the discharge letter.

The average size discharge document contains 260 tokens and 16 sentences and all the normative segments were used. In this document the segments were correctly filled; the document summarizes the information concerning the inpatient admission reason, clinical history, examinations performed during the referred period, the status of the patient at the time of discharge, the therapeutics carried out and the destiny of the patient after discharge.

Finally, the shortest discharge document contains only two tokens. The single used segment was *Admission Reason*, filled with the sentence “*AVC isquémico*” (*ischemic stroke*).

The discharge documents have also interesting **contextual features**. In general, it is

evident that the narratives are written from a professional to a professional in order to support information transfer, remind important facts and supplement numerical data.

The texts are normally intelligible² and the meaning becomes evident from the context even in the presence of numerous linguistic and grammatical mistakes, several word abbreviations and signs (e.g., $\pm$, $<$, $--$ $>$). Pronouns are used infrequently and almost all the sentences are incomplete and do not contain any verb. This is particularly visible in the *Physical Examination* and *Therapeutics* segments. The use of acronyms and abbreviations is very common. Almost every document included several acronyms, being these frequently made after the English form (for instance, HDL for High Density Lipoprotein instead of the equivalent Portuguese acronym for *Lipoproteína de Alta Densidade*).

A succinct qualitative analysis of the discharge documents is given in table 3.8.

Table 3.8: Structural and contextual features of MEDINX discharge documents. The original examples are added in ().

Structure	Examples
Incomplete sentences and missing verbs	<i>No complaints suggestive of infectious syndrome or fever.</i> (<i>Sem queixas sugestivas de síndrome infeccioso ou de febre.</i>) <i>COC, eupneic without RDS.</i> (<i>COC, eupneica sem SDR.</i>)
Misspellings are found but most of the content or meaning is still clear.	<i>Dysathric with coherent speech.</i> (<i>Disátrico, com discurso coerente.</i>)
Content	Examples
The word <i>patient</i> as a subject or object is infrequently mentioned.	<i>Uses drugs regularly.</i> (<i>Faz uso regular de fármacos.</i>)
Signs are typically used	<i>PA-Vm + without considerable rales.</i> (<i>AP- Mv+ sem RA valorizaveis.</i>)
Abbreviations and acronyms are frequent and vary between physicians	<i>BP-193 / 68 mmHg; spO2 (AA) - 99%, PR-60/min. CA-irregular, due to ES.</i> (<i>TA- 193/68 mmHg; spO2 (AA)- 99%; PR- 60/min. AC-irregular, por ES.)</i>

3.1.3.4 Gold Standard selection and modeling process

The main goal of MEDINX is the identification of semantic entities and relations between the last in the discharge documents. Typical approaches require human annotated documents to provide both evaluation standards and material for the system development.

In MEDINX several efforts were undertaken to follow this approach.

²When being analyzed by a non-medical expert, as myself.

Development and Test Sets: Given the expense of human annotation, the gold standard portion of the corpus had to be a relatively small subset of the whole corpus of 915 documents.

Given so, the MEDINX corpus of discharge letters was divided into two subsets correspondent to the *Development Set*, intended to be used in the definition, development and refinement of the systems components, and the *Test Set*, to be used in the evaluation of the performance of the system, and consequently in the gold standard.

Approximately 10% of the MEDINX corpus, exactly **86** discharge documents, were randomly selected to be part of the **gold standard**, while the remaining **829** documents compose the MEDINX **development set**.

Table 3.9 give some statistics for the amount of documents, sentences and tokens existing in the development and test set.

Table 3.9: Number of documents, tokens and sentences in the MEDINX corpus and its subsets.

NUMBER OF	DEVELOPMENT SET	TEST SET	TOTAL
Documents	829	86	915
Tokens	215,730	21,788	237,518
Tokens/Document	260	253	260
Sentences	12,974	1,346	14,320
Sentences/Document	16	16	16

Gold Standard Annotation Schema: Evaluating and developing automatic IE systems is usually performed through comparison between automatically extracted terms against manual reference standard annotations, i.e., gold standard collections. However, there are no standard guidelines for determining which words from a textual document to include in manual annotations and the undefined task can result in substantial variation among the manual annotators.

For example, in the sentence:

- (12) Doente de 59 anos, com sequelas de AVC há 3 anos.
 ('59 years old patient with sequelae of stroke occurred 3 years ago.')

several annotations of the clinical condition are possible: *stroke*, *sequelae of stroke*, *sequelae of stroke occurred 3 years ago*.

To overcome this and other similar difficulties, a detailed annotation schema was developed in cooperation with a qualified physician. This cooperation was essential to obtain the medical expertise required for the development of a useful and valid annotation schema. This schema

includes variables representing medical concepts (e.g., conditions, drugs), modifier types (e.g., negation, modifier) and the relations between them.

The standardized annotation schema was induced from the PDLs through the manual identification of relevant clinical entities and their relations in the texts. *Entity* is meant as some real-world concept referred to in the text, such as the drugs that are mentioned, the tests that were carried out, etc. The *relations* between entities are, in this case, the condition indicated by a drug, the result of an examination, etc. This annotation schema enabled the annotators to individually annotate the clinical entities and their relations in the same way.

The annotation schema was created using a multi-staged approach. First, the annotation goal was defined:

- to annotate the most specific, atomic clinical entities, including temporal information and modifiers, as well as the relations between them.

It was decided to not include demographic information, as age and sex, since this type of information exists already in a structured form in the RTS. Next, several documents from the development set were analyzed in order to enumerate the specific entities to be considered while annotating clinical entities and relations.

As an example, consider the following sentence retrieved from a PDL, and its English translation:

- (13) Doente de 59 anos, com antecedentes de diabetes não insulino-tratados, HTA, dislipidemia, sequelas de AVC há 3 anos com hemiparésia esquerda discreta e cirurgia da carótida direita há 6 meses.

59 years old patient, with history of non-insulin-treated diabetes, hypertension, dyslipidemia, sequelae of stroke three years ago with left-sided mild hemiparesis and surgery of the right carotid 6 months ago.

The sentence allows to identify several clinical entities, as, for example, the **CONDITIONS** as *diabetes, hypertension, dyslipidemia, stroke, hemiparesis*. The sentence also contains information about the **SURGICAL** procedure *surgery, targeted at the ANATOMICAL SITE carotid, modified* in location by the **LATERALIZER** *right*.

Iteratively, several discharge documents were read, the entities occurring in the text and its relations were defined and grouped into related concepts, determining which variables should or should not be annotated.

The developed annotation schema is illustrated in Figure 3.4, and a list of all entities and its relations is given in Tables 3.10 and 3.11, along with descriptions. The annotation schema

was modeled using the Unified Modeling Language (UML) (Object Management Group, 2011; Rumbaugh *et al.*, 1999).

Table 3.10: Entities of the gold standard annotation schema.

ENTITIES	DESCRIPTION
Condition	Complications, conditions and other problems manifested by a patient
Anatomical Site	Anatomical structure or location, normally the locus of a <i>Condition</i>
Evolution	The clinical evolution of the patient or <i>Condition</i> after a given <i>Therapeutics</i>
Examination	Interaction between doctor and patient or <i>Anatomical Site</i> with the purpose of measuring or studying some aspect of a <i>Condition</i>
Finding	The numeric or qualitative finding of an <i>Examination</i> , excluding <i>Condition</i>
Location	Geographically defined location, normally where an <i>Examination</i> or <i>Therapeutic</i> is performed
Therapeutic	Action performed by a clinician targeted at a patient, <i>Anatomical Site</i> or <i>Condition</i> with the purpose of changing or treating a <i>Condition</i>
DateTime	Temporal expressions, including dates and times (absolute or relative), duration and frequencies
Value	Absolute or relative quantifications or classifications
Negation	Expressions that negate other entities
Characterization	Expressions that characterize other entities

In order to ensure the consistency, and consequently the quality, of the gold standard all documents have to be annotated to the same standard. Several questions regularly arise when annotating, as, for instance, *should multi-word expressions be split?* or *should accidente vascular cerebral (cerebrovascular accident) be annotated as a CONDITION or as a CONDITION and an ANATOMICAL SITE?*

To deal with this situation, a set of guidelines was developed and provided to annotators. These are presented in Appendix A and describe in detail what should and should not be annotated, how to decide if two entities are related, and a number of special cases. The guidelines also provide a sequence of steps which the annotator should follow when working on a document.

The annotation schema was implemented within Knowtator (see Section 3.2.3), a Protégé plugin (presented in Section 3.2.2). Figure 3.5 presents the Knowtator interface and illustrates the gold standard annotation process. The left part of the image shows the representation model created, reflecting the annotation schema entities and relations, while the center and right side of the figure show the annotations process, particularly the selection of the text

Table 3.11: Relationships used on the gold standard annotation schema.

RELATIONS	1ST ARGUMENT	2ND ARGUMENT	DESCRIPTION
has_target	Condition, Examination, Therapeutic	Anatomical Site	Describes the bodily location of a specific condition, examination or therapeutic;
has_finding	Examination	Condition, Finding	Relates an examination to the condition demonstrated or to the produced result;
has_location	Examination, Therapeutic	Location	Relates the examination or therapeutic to the geographic location it takes place;
has_indication	Therapeutic	Condition	Relates a therapeutic with the condition it was prescribed to.
has_evolution	Condition	Evolution	Relates a condition with its clinical evolution
has_value	Condition, Finding, Therapeutic	Value	Relates a condition, finding or therapeutic with the associated numeric expression;
has_date	Condition, Evolution, Finding, Therapeutic, Examination	DateTime	Relates the classes to the temporal expression it is associated with;
modifies	Negation, Characterization Signal	Condition, Finding, Evolution, Therapeutic, Anatomical Site	Relates the classes to its negation or other characterizing information;
co-refers	Any	Any	Relates two spans of text referring to the same class in the real world.

Figure 3.5: Gold Standard annotation representation in Knowtator.

MEDINX.

3.2 Processing Tools

An important challenge faced during the development of this work concerns the decision of which tools to use in the implementation of the several components of MEDINX. Most current tools for NLP are complex and usually created in a stand-alone manner, intended to solve specific problems and fit specific systems. In the last couple of years, several repositories have been created including, for example, the Linguistic Data Consortium (Linguistic Data Consortium, 2010), the Natural Language Software Registry (Natural Language Software Registry, 2010), the European Language Resources Association (European Language Resources Association, 2010) and the distributed language resource center for Portuguese, Linguatca (Linguatca, 2008). However, the attempts to reuse this software or create composite NLP systems from modules selected from these repositories are often extremely challenging if one is interested in some level of interoperability.

Also, the flow of knowledge to and from the several parties involved in the development of such system, as the computer scientists, clinicians and linguistics, is not always easy or straightforward to manage. A good synergy and circulation of knowledge between these experts is fundamental to the development of a successful tool.

The tools used to implement the different components of the system, along with some of the reasons that determined their specific choice, are shortly presented in the next sections. In general, the choice was based on the requirement of using modular and versatile platforms, that ensure the information flow to and from the involved actors (computer scientists, physicians and linguists) and that represented the state of the art in the area they were developed for.

3.2.1 UIMA

The Unstructured Information Management Architecture (UIMA) is an open, industrial-strength, scalable and extensible platform for creating, integrating and deploying unstructured information management solutions from combinations of semantic analysis and search components (Ferrucci and Lally, 2004).

Although UIMA has been originated at IBM, it has moved on to be an open source project which is currently incubating at the Apache Software Foundation (UIMA, 2011).

In recent years, the UIMA framework has been proposed as a middleware platform which offers integration by design through common type systems and standardized communication methods for components analysing streams of unstructured information, such as natural language. The UIMA framework offers a solid processing infrastructure that allows developers to concentrate on the implementation of the actual analytics components.

An increasing number of members of the NLP community have adopted UIMA (Hahn *et al.*, 2008; Kunze and Rösner, 2008; Müller *et al.*, 2008; Savova *et al.*, 2008) as a platform facilitating the creation of reusable NLP components that can be assembled to address different NLP tasks depending on their order, combination and configuration. Carnegie Mellon University's Language Technology Institute (Carnegie Mellon University, 2010) and Jena University's Language & Information Engineering Lab (JULIE, 2010) have both created UIMA component repositories, where developers can post information about their components and anyone can find out more about free and commercially available UIMA-compliant tools. The UKP lab (Ubiquitous Knowledge Processing Lab, 2011) from the Technical University of Darmstadt has also produced a collection of software components for NLP based on UIMA, the Darmstadt Knowledge Processing Repository (DKPro). The available components cover the whole range of NLP related processing tasks, providing wrappers for, not only original NLP components, but also for third party tools. DKPro is available under the Apache Software License (Apache Software Foundation, 2011).

The modular and versatile architecture of UIMA provides a framework to address the challenges inherent to the development of a MLP system. UIMA is a component architecture and software framework implementation (including a UIMA Software Development Kit) to develop applications that analyse large volumes of unstructured information, and has been

increasingly adopted by a significant part of the bio and medical NLP community that needs industrial-grade and robust applications to exploit the whole bibliome.

An UIMA application may be thought of as comprising two phases: analysis and delivery. In the analysis phase, collections of documents are acquired and analyzed. The results are stored in one or more forms as needed for the delivery phase. This phase may include tokenization and semantic class detection on input documents, and may use structured sources such as dictionaries or ontologies, to find and annotate classes of entities such as organizations, persons, locations, and events.

In the delivery phase, the analysis results, possibly together with the original documents or other structured information, are made accessible to the application's end user through application-appropriate access methods and interfaces. In this phase, the application might present a query interface and allow the user to search for documents that contain some combination of tokens, entities, and relationships through a semantic search engine.

Figure 3.6: UIMA architecture (adapted from (Ferrucci and Lally, 2004)).

At the heart of UIMA are two key features: the UIMA type system and a common representation system called the Common Analysis Structure (CAS). The UIMA type system is a declarative definition of an object model (Verspoor *et al.*, 2009), and serves two main purposes:

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

The CAS is used to provide analysis engines with read access to the artifact being analyzed (e.g., document, image, video, etc) and read/write access to the analysis results or annotations

associated with defined regions of the artifact. Regions may correspond to words, sentences or paragraphs in text frames and parts of frames in video, among others.

Analysis engines process CASs. They look at the subject of analysis and any results produced by previous analysis engines and 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.

In addition to the analysis engines, UIMA introduces two other primary components. These are the Collection Reader and the CAS Consumer. The Collection Reader's job is to connect to and iterate through a source collection, acquiring documents and initializing the CAS for analysis.

CAS Consumers, as the name suggests, function at the end of the flow. Their job is to do the final CAS processing. A CAS Consumer may be implemented, for example, to index CAS contents in a search engine, extract elements of interest and populate a relational database or serialize and store analysis results to disk for subsequent and further analysis.

The complete flow from source, through document analysis, and to CAS Consumers supported by UIMA is illustrated in Figure 3.6.

3.2.2 Protégé

Protégé is the ontology development tool chosen to support ontology development in this work. Protégé is an extensible and customizable toolset for constructing knowledge bases and for developing applications that use these knowledge bases, developed by the Stanford Medical Informatics group at Stanford University. Protégé began as a small application designed for a medical domain (protocol based therapy planning), but has evolved into a general-purpose set of tools (Gennari *et al.*, 2002).

Protégé is characterized by its open architecture, which allows an easy integration with other applications. It is knowledge representation language independent, contrarily to many other ontology development tools, as for instance, the Ontolingua Server (Farquhar *et al.*, 1997), the OntoSaurus (Swartout *et al.*, 1997) and WebOnto (Domingue, 1998) tools. Also, being written in Java, runs under a wide variety of operating systems.

Protégé assists users in the construction of large electronic knowledge bases. It has an intuitive user interface that enables developers to create and edit domain ontologies. Figure 3.7 shows a snapshot of the ontology editor presenting the class **CONDITION** of the gold standard annotation schema. Protégé's numerous plugins provide alternative visualization mechanisms, enable management of multiple ontologies, allowing the use of inference engines and problem solvers with Protégé ontologies, along with other functionalities.

Figure 3.7: Protégé’s default interface

3.2.3 Knowtator

Several text annotation tools, that can be employed to accomplish a wide variety of annotation tasks, have emerged in the last few years. Some of the general-purpose annotation tools include Callisto (MITRE, 2010), WordFreak (WordFreak, 2010), MMAX2 (MMAX2, 2010) and GATE (Cunningham *et al.*, 2002), a software architecture for NLP that has, as one of its many components, manual text annotation functionality. Each of these tools is distributed with a limited number of annotation tasks that can be used “out of the box”. However, the need to create a complex annotation schema, with hierarchical and constrained relations between annotation types, to be used by several actors from different expertise areas, lead us to the use of Knowtator, a general-purpose text annotation tool.

Knowtator (Ogren, 2006a) has been implemented as a Protégé plug-in and runs in the Protégé environment, leveraging Protégé’s knowledge representation capabilities to specify annotation schemas. In Knowtator, an annotation schema is defined using Protégé’s classes, instances, slots and facet definitions and with the help of Protégé’s knowledge-base editing functionalities.

Knowtator relies on the fact that the defined annotation schema can be applied to a text annotation task without having to write any task specific software or edit specialized configuration files. Also, the annotation schemes in Knowtator can model both semantic (e.g.

disease-drug interactions) and linguistic phenomena (e.g. coreference resolution).

3.2.4 Jena

Jena (JENA, 2010) is a Java framework for building Semantic Web applications based on W3C recommendations for RDF and OWL. It provides a programmatic environment for RDF, RDFS and OWL, 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) was used in this work. This API defines object classes to represent RDF graphs, resources, properties and literals. More specifically, Jena was used to develop a programmatic representation of the ontology created to model the structure and content of the discharge letters, as well as, to provide access to its content and to ensure its correct automatic population.

3.2.5 Lucene

Lucene is a high performance, scalable Information Retrieval (IR) library which can be added to an application to allow indexing and searching capabilities. Lucene is a free, open-source project implemented in Java and currently, the most used free Java IR library (Hatcher and Gospodnetic, 2004).

Lucene was used in MEDINX in order to ensure the maintainability, performance, efficiency and scalability of the system. Lucene's indexing capabilities were used to index the information contained in the MEDINX knowledge sources (as the UMLS content and the ontologies created to support information extraction, presented in chapter 4) in order to achieve a faster and more efficient system performance.

3.3 Summary

This chapter presented the resources and tools used to develop MEDINX. The creation of systems capable of performing intelligent information and knowledge processing is heavily dependent on the use of domain-oriented external knowledge sources. UMLS, one of external knowledge MEDINX resources used in MEDINX, integrates a variety of biomedical information sources and is intended to facilitate the development of advanced information systems.

The content and structure of the corpus of hypertension related PDLs is also characterized. This analysis allowed to understand how physicians typically use the structured PDL and to conclude, for instance, that physicians often do not follow the defined normative structure, which complicates both the information access and intelligibility.

In the development of clinical IE systems it is crucial to ensure the flow of knowledge to and from the several involved actors, more specifically, between the computer scientists, physicians and linguists involved in this work. This flow is facilitated by the use of appropriate processing tools. The main processing tools used for the development of the several components of MEDINX were introduced. This selection is considered a major issue in order to ensure the components interoperability.

Chapter 4

Medical ontologies

Clinical documents such as discharge letters contain mostly technical terms, from which only a few are found in common language use (written and spoken) often enough to worth inclusion in common dictionaries. However, most IE systems rely on finding information on textual forms, partly through dictionary look-up and, thus, rely on finding information about terms as they occur in texts. Moreover, IE systems capable of building up representations of clinical facts and events need to access structured representations of the domain in order to be able to assign domain-specific meanings to terms and use these meanings in their operations.

Ontologies are crucial for knowledge discovery in the clinical domain. They form the link between terms in texts and their meaning, and are, therefore, an important resource for semantic annotation of texts. They are also essential to perform inference based on ontological relations. Given so, sophisticated IE, of the kind required to handle clinical narratives, require access to such models.

Knowledge resources such as the UMLS contain an abundance of nomenclatures, controlled vocabularies and ontologies. These knowledge sources exist in several languages, including some translations to Portuguese, as it was seen in Section 3.1.2. However, the information existing in the Portuguese vocabularies of UMLS is not enough to extract relevant clinical entities from discharge summaries written in Portuguese, as are the ones processed with MEDINX. For example, the information regarding the drugs used in Portugal and their active ingredients is not available in UMLS.

This chapter describes the ontologies created within MEDINX. More specifically, two formalizations of the international classification systems ICD-9-CM and ICF were created, plus a drugs ontology and a conceptualization of the structure and content of the discharge reports, designated by MEDINX ontology.

4.1 Method for Ontology Development

Ontologies must be designed to be internally consistent, offer wide coverage of the domain while supporting interoperability. In this work, special attention was given to standard principles while designing ontologies, namely *consistency*, *completeness* and *conciseness*. The ontologies developed in MEDINX were designed to not allow contradictory conclusions from any definition and axiom, and the completeness of each definition was verified to guarantee the ontology completeness. Special care was given to ensure that no unnecessary or useless definitions were stored in the ontologies and that no redundancies could be inferred from its definitions and axioms.

The modeling process of an ontology depends on what the ontology is for. In MEDINX four different ontologies, intended to be used in different tasks, were developed. The methods used to build each of the MedInX ontologies vary according to its purpose and content. Nevertheless, some basic steps, based on the guidelines developed by Noy and McGuinness (2001), were performed while creating all the ontologies used in MEDINX:

1. The purpose and scope of the ontology was identified. This step pretends to clarify why the ontology is being built, what are its intended uses (to be reused, shared, used as a part of a information extraction system, etc) and what are the relevant terms in the domain.
2. The knowledge of the domain was captured, according to the following steps:
 - (a) identification of key concepts and relationships in the domain;
 - (b) production of precise and unambiguous textual definitions for such concepts and relationships; and
 - (c) automatically identify the terms that refer to such concepts and relationships.

This process will be further explored in the next sections, while describing the methods used to develop each one of the created ontologies.

4.2 ICD-9-CM ontology

A formalization in OWL-DL of the ICD-9-CM was developed in MEDINX in order to support, not only the IE task, but also the automatic code assignment to the PDLs.

Although the tenth version of ICD is already available, the ninth version, clinical modification of WHO's ICD is still the classification used in assigning codes to diagnoses associated with inpatient episodes in several countries, and particularly in Portugal.

ICD-9-CM is structured in a core classification of three digit codes and a fourth digit to further specify the category. Its hierarchy is organised into four levels:

- **17 chapters** group together diseases according to major categories (*location* - e.g. infectious and parasitic diseases (001-139) - and *pathology* - e.g. neoplasms (140-239));
- Chapters contain **sections**, a group of three-digit categories that represent a single disease entity, or a group of similar or closely related conditions (like 001-009 - Intestinal Infectious Diseases);
- Sections contain **categories**. A three-digit category represents a single disease entity, or a group of similar or closely related conditions (as 001 Cholera);
- Categories contain **subcategories** of four digits, providing more specificity or information regarding the etiology (cause of a disease or illness), site (location), or manifestation (display of characteristic signs, symptoms, or secondary processes of a disease or illness). Four-digit subcategories are collapsible to the three-digit level (e.g., 001.0 Cholera due to *Vibrio cholerae*).

ICD-9-CM axiomatically assumes that sibling classes do not overlap, i.e., a single disease cannot be classified under two different categories. This warrants that no class has more than one parent class and that there is exactly one terminal class for each entity to be classified, hence its characterization as a *classification*. This way one disease is never counted twice.

Figure 4.1 illustrates the hierarchy of ICD-9-CM as it was implemented in MEDINX. The ontological position taken in this work was to view this coding scheme as an ontology of medical conditions, referred as categories, with a set of associated terms having these categories as their denotata. The ontology contains, thus, concepts (classes) representing each of ICD-9-CM categories and subcategories, named by the correspondent ICD code, and is populated with instances of these classes. The super class ICD-9-CM attributes allow for the introduction of the associated terms. The attributes of the ICD-9-CM classes are the following:

- *Preferred Name*, the name used to define the category;
- *Other Names*, synonyms of the *Preferred Name*;
- *CUI*, the UMLS Metathesaurus Concept Unique Identifier (CUI); and
- *Code*, the ICD-9-CM code.

The MEDINX ICD-9-CM model was developed programmatically, using Jena's ontology API (see Section 3.2.4) and automatically instantiated, using UMLS knowledge sources, as follows:

Figure 4.1: MedInX ICD-9-CM ontology model.

ICD-9-CM terms in UMLS: To obtain the information needed to define ICD-9-CM subclasses, the UMLS Metathesaurus was used. MRCONSO, the source table of UMLS Metathesaurus database (presented in Section 3.1.2), was examined for the CUI of records containing *ICD9CM* in the Source Abbreviation (SAB) fields and *PT*, Preferred Term, or *HT*, Hierarchical Term, in the Term Type (TTY) field.

Since ICD-9-CM records in UMLS do not exist in Portuguese, the information retrieved with the former process was used to extract the terms with the same CUI, i.e., representing the same concept, but with the Language of Terms (LAT) defined as *POR* (Portuguese).

An illustrative example of this process is given with the code *002.0*, corresponding to the disease *Typhoid fever* and identified in UMLS with the CUI *C0041466*. Querying the UMLS database for this CUI and language (LAT) defined as *Portuguese* will return the term *Febre tifóide (typhoid fever)*.

For each ICD-9-CM code a new class was defined. If the code corresponded to an ICD-9-CM category or subcategory, i.e., a three or four digits code, a new instance of the class was defined. The properties of the instance were then filled with the information obtained in UMLS, namely, the *Preferred Name*, *CUI* and ICD-9-CM *Code*.

The coverage of the ontology was expanded by looking in UMLS for other terms in Portuguese with the same CUI. For instance, in the case referred in the previous example, the terms *Febre entérica (Enterica Typhoid fever)*, *Tifo Abdominal (Abdominal Typhoid)*, *Tifóide (Typhoid)* and *Infecção por Salmonella typhi (Salmonella typhi infection)* were found as synonyms of the instance 002.0, and used to fill the property *OtherNames* of the instance.

Setting the disjoints between classes, in order to guarantee that single disease cannot be classified under two different categories, concluded the creation of the ontology.

This ontology is illustrated in Figure 4.2 in Protégé’s interface. The right side of the image shows the properties of the instance of 002.0, *typhoid fever*, used in the previous example.

Figure 4.2: ICD-9-CM ontology in Protégé

The current ICD-9-CM ontology contains 5,488 instances, defined by 5,830 terms in Portuguese. This ontology can be further populated and reused in different applications and systems.

4.3 ICF ontology

International Classification of Functioning, Disability and Health (ICF) is a classification of health and health-related domains (World Health Organization, 2010b). This classification was first created in 1980 by the WHO, as the International Classification of Impairments, Disabilities, and Handicaps, intended to provide an unifying framework for classifying the consequences of diseases. The ICF is based on an integration of the medical and social models of disability, and addresses the biological, individual, and societal perspectives of health in a biopsychosocial approach (Peterson, 2005). The ICF classifies functioning and disability by considering the body functions and structures, activities and participation components, and addresses contextual influences through environmental and personal factors.

The ICF contains classifications and codes for each of the following components:

- **Body functions**, the physiological functions of body systems. It includes, for example, seeing, hearing, feeling pain or remembering;
- **Body structures**, the anatomical parts of the body, as for example the eye, the mouth, the spinal cord;
- **Activities and Participation** where an activity is defined as the execution of a task or action by an individual and participation as the involvement in a life situation;
- **Environmental factors** comprise the physical, social and attitudinal environment in which people live, and which are external to the individual.

ICF and ICD both belong to the WHO family of international classifications. However, ICD uses an etiological framework for the classification, by diagnoses, of diseases, disorders and other health conditions, while ICF classifies the functioning and disability associated with health conditions (World Health Organization, 2010a). The ICD and ICF are therefore complementary and their use together allows the creation of a broader and more meaningful picture of the experience of health of individuals and populations.

Figure 4.3: Example of the structure of an ICF code.

The general format of an ICF code is illustrated in the example given in Figure 4.3. Consider the example code s250.28 recorded for a person experiencing a moderate problem with the structure of the middle ear. The code can be divided into several elements, where:

- 's' denotes the component, in this case Body structures. Body functions are denoted with 'b', activities and participation with 'd' and the environmental factors with 'e';
- the first digit (2) denotes the domain, in the example the ear and related structures;
- the second and third digits (50) denote the second-level category, the structure of middle ear in Figure 4.3;

- the first digit after the decimal point (2) denotes a generic qualifier, indicating, in this case, a moderate impairment with the middle ear. The qualifiers indicate the level of a certain parameter, such as the level of facilitation or impairment conferred by a health condition.
- the second digit after the decimal point (8) indicates, in this case, that the nature of the impairment is not specified.

ICF terms in UMLS MEDINX ICF ontology structure is similar to the structure of the ICD-9-CM ontology; it models the information related with the components of ICF and contains concepts (classes) representing the ICF components. Each of these classes is instantiated, and the attributes of the instances contain the terms used to refer to these concepts. Figure 4.4 outlines the hierarchy of classification used to model the ICF ontology.

Figure 4.4: MedInX ICF ontology model.

The information regarding the qualifiers of a code was not included in the ontology. These qualifiers were defined in order to record the extent of the “problem” in relation to the impairment, activity limitation, participation restriction and environmental barrier, and have specific scales determining its usability. Since it is intended to model the components of ICF and not its application, this information was not reflected in the ontology.

The method used to build the ICF ontology is similar to the one used in the ICD-9-CM ontology; it was programmatically defined and automatically populated through the examination of the UMLS Metathesaurus. More specifically, the UMLS source table MRCONSO was examined for the CUI of records containing the Source Abbreviation *ICF*.

For each concept a new class was created. If the code of the class corresponded to a second-level category (one character and three digits code) a new instance of the class was defined. The properties *Preferred Name*, *CUI* and *Code* of the MEDINX ICF ontology classes were automatically filled with the information retrieved from UMLS. In particular, the property *Preferred Name* was filled with the English definition of the equivalent code.

However, one additional difficulty was found in this case. ICF describes specific health related states and the terms used in the definition of each code, as for example, *Watching (d110)*, *Consciousness functions (b110)*, *Structure of brain (s110)*, *Products or substances for personal consumption (e110)* do not have any associated synonyms. Also, at the time of writing, no Portuguese translation of the ICF definitions was available in the UMLS Metathesaurus.

Given so, and in order to process clinical documents written in Portuguese, the property *Other Names* was manually filled with the associated Portuguese terms. For example, the instance *b110* correspondent to the Body Function *Consciousness functions* was associated with the terms *consciente (conscious)* and *inconsciente (unconscious)* frequently used in the PDLs while referring to this Body function. The example is illustrated in Figure 4.5, representing MEDINX ICF model in Protégé's Graphical User Interface (GUI).

Figure 4.5: ICF Model ontology in Protégé

The ontology contains currently 397 instances and 67 terms in Portuguese. The ontology can be further populated in order to be used in IE or other related applications.

4.4 Drugs ontology

One of the most common clinical entities referred in MEDINX PDLs is the patient **medication**, including the medication given to the patient during his admission period, as well as, the drugs already used by the patient, or even the medication recommended after discharge.

However, no formal conceptualization of the pharmaceutical drugs used in Portugal, being them used in hospitals or any other health institution, is available for research or similar purpose. Infarmed, the National Authority of Medicines and Health Products (Informed, 2011), provides detailed information concerning all the human medication used in Portugal and for which a market authorization has been given. However, no OWL/RDFS representation of such knowledge is available currently. The MEDINX *drugs* ontology was build to amend this issue.

For the required purposes, a simple knowledge model, containing only two concepts, *Medication* and *Active ingredient*, and the composition relationship between them (Figure 4.6), was idealized.

Figure 4.6: MedInX Drugs ontology model.

The model was programmatically defined using Jena's API and its population was automatically performed with the help of the information publicly available in Infarmed's website (Informed, 2011).

A wide search for all authorized drugs, including generics, was performed and its retrieval, containing information about the name of the medication, its active ingredients and hierarchy type in a tabular form, was used to automatically populate the ontology. For each retrieved active ingredient, a new instance of the class *Active Ingredient* was created and its hierarchical type (the group the drug belongs to as, for example, *Antiepileptic*, *Antipyretic*, etc.) filled. Similarly, for each medication listed in the file, a new instance of the class *Medication* was created, and the object property *Active Ingredient* filled with the corresponding instance. The hierarchical type of the medication and the boolean property *generic*, indicating if the product corresponds to a generic drug, were also appropriately completed.

The current ontology contains information concerning 3,876 medications and 1,780 active

Figure 4.7: Drugs ontology in Protégé

ingredients, organized into 214 hierarchical types. Figure 4.7 presents the structure of the ontology. It is possible to observe in the image the detail of the instance *Topiramato toLife 100mg Comprimidos*, a generic medication of the type *Antiepiléticos e Anticonvulsionantes* (*Antiepileptic and Anticonvulsant*), containing as active ingredient *Topiramato*.

New medications are frequently given authorization to the market or have their authorization cancelled. Considering this situation, a request was sent to Infarmed in order to obtain the complete content of their database. Only a frequent update of the ontology, with the information provided by Infarmed, would ensure the completeness and the validity of the data existing in it.

4.5 MEDINX Patient Discharge Letters Knowledge Representation

The ontologies described in the previous sections were developed in order to be used as knowledge sources in MEDINX tasks, such as IE and automatic code assignment. This section describes the extensible knowledge model used for storing and structuring the hypertension discharge summaries entities and their relations, including temporal information and modifiers. This model will be referred in the remaining document as the MEDINX

ontology.

The primary goal of this ontology was to model the PDLs and the concepts they describe, as the conditions described in the texts or the values observed in the procedures carried out during patient's admission. For instance, the MEDINX ontology describes the fact that a *Sign* or *Symptom* is a *Condition*, a prescribed *Medication* is a *Therapeutic*, a *Procedure* can be targeted to an *Anatomical Site*, and so forth.

Such knowledge can be used in various practical applications, as, for example, in clinical decision support, in exchanging information among different EHR systems, or, for example, in the identification of groups of patients with similar disease progression and in the summarization of the progress of the conditions of a single patient.

To identify the concepts in the ontology a middle-out strategy was used, i.e., first were identified the core of basic terms, and then the terms were specified and generalized as required. For that the gold standard schema was used (see Section 3.1.3.4). In particular, the previous analysis and the annotation schema developed to build the MEDINX golden collection was used as an initial base for the ontology. Several evolutions were done to better reflect the terms used by the clinicians in the clinical setting, as described in the following paragraphs.

Table 4.1 lists and describes the classes of the model, while the relations between the classes are listed in Table 4.2. Figure 4.8 summarizes the classes and subclasses used in the MEDINX ontology. Each class can be thought of as a category of NEs with associated specific attributes, while each subclass represents the type of the category.

Figure 4.8: Tree of categories, types and subtypes considered in the PDLs knowledge representation.

Table 4.1: Classes and Subclasses of the MEDINX Knowledge Representation

CATEGORY	TYPE	DESCRIPTION
Anatomical Site		All references to an anatomical structure or location, normally the locus of a <i>Condition</i>
Condition	Disease Sign or Symptom Finding Pathology	All the references to health states, complications, conditions, sign and symptoms, pathologies and other problems manifested by a patient
Evolution		All references to the supervision and clinical evolution of the inpatient or of a <i>Condition</i>
Procedure	Therapeutic or Preventive Laboratory Diagnostic Chemical	Interaction between a physician and a patient or <i>Anatomical Site</i> with the purpose of measuring or treating some aspect of a <i>Condition</i>
Modifier		All expressions that categorize an entity
Negation		Expressions that negate an entity. These expressions should be considered separately from the entity being categorized
Lateralization		All the references to the lateralization of a <i>Condition</i> , <i>Procedure</i> or <i>Anatomical Site</i> (e.g., left, right, upper, etc.)
Administration Route		The route of administration of a <i>Chemical</i> <i>Procedure</i>
Time	CalendarTime Duration Frequency	Temporal expressions, including dates and times (absolute or relative), duration and frequencies
Value	Classification Dosage Quantity	Absolute or relative quantifications or classifications. The units that characterize the numerical expression should be included.
Episode		General content referred in the PDL

Table 4.2: Relations of the MEDINX Knowledge Representation.

OBJECT PROPERTY	FIRST ARGUMENT	SECOND ARGUMENT	DESCRIPTION
includes	Condition, Procedure	Condition, Procedure	Relates two or more <i>Conditions</i> (or <i>Procedures</i>) with each other;
hasValue	Condition, Procedure	Value	Relates a <i>Condition</i> or <i>Procedure</i> with the associated numerical expression;
hasTime	Condition, Evolution, Procedure	Time	Relates the classes with the temporal expressions they are associated with;
hasTarget	Procedure, Condition	Anatomical Site	Describes the <i>Anatomical Site</i> of a specific <i>Condition</i> or <i>Procedure</i> ;
hasSubstance	Condition	Substance	Describes the <i>Substance</i> related with a given <i>Condition</i> ;
hasQuantity	Chemical Procedure	Quantity	Described the administrated quantity of a given <i>Chemical Procedure</i> ;
hasNegation	Condition, Procedure, Modifier, Evolution	Negation	Characterizes the negated classes;
hasModifier	Condition, Procedure Evolution, Anatomical Site	Modifier	Provides further characterization to the classes;
hasLateralization	Condition, Anatomical Site	Lateralization	Further characterizes an <i>Anatomical Site</i> or the location of a given <i>Condition</i> ;
hasIndication	Procedure	Condition	Relates a given <i>Procedure</i> with the <i>Condition</i> that it indicates;
hasFrequency	Chemical Procedure	Frequency	Describes the <i>Frequency</i> of a given <i>Chemical Procedure</i> ;
hasFinding	Procedure	Finding	Relates a given <i>Procedure</i> with the exact <i>Finding</i> it indicates;
hasEvolution	Condition,	Evolution	Relates a <i>Condition</i> with its clinical <i>Evolution</i> ;
hasDuration	Chemical Procedure	Duration	Describes the <i>Duration</i> of a given <i>Chemical Procedure</i> ;
hasDosage	Chemical Procedure	Dosage	Relates a <i>Chemical Procedure</i> with the strength it is applied;
hasAdministrationRoute	Chemical Procedure	Administration Route	Describes the <i>Administration Route</i> of a given <i>Chemical Procedure</i> ;

To further characterize each of the classes and its relations, several partial views of the total model were developed. Figures 4.9 and 4.10 concentrate, respectively, on the classes *Episode* and *Procedure*, and their relations with other classes. In particular, the class *Episode* captures the content of each analyzed PDL. The attributes of this class, presented in Figure 4.15, specify the header data of each report, as its identification, the codes associated with the DRG code of the document, the ICD-9-CM code of the main diagnosis and the textual description of each of these codes. The class contains also attributes that specify its relations with the classes *Condition*, *Anatomical Site*, *Evolution*, *Procedure*.

Figure 4.9: Relationships between the class *Episode* and the classes *Condition*, *Anatomical Site*, *Evolution* and *Procedure* in the MEDINX ontology.

The class *Procedure* (Figure 4.11) describes the several initiatives developed to measure or study some aspect of a given *Condition*. The class is subdivided in further subclasses, representing the different types of *Procedures* often included in the PDLs. This includes *Therapeutic or Preventive* procedures, as well as, *Diagnostic*, *Laboratory* and *Chemical* procedures. This last subclass conceptualizes the drugs administrated to the inpatient. These drugs can be subcategorized into *Medication*, *Active Substance* and *Substance*. Figure 4.11 shows the details of *Procedure* class, in particular, its attributes and relations.

Subsequent figures describe other classes in more detail. Particularly, Figure 4.12 presents the details of the *Condition* class. The attributes of this class describe the CUI and ICD code of the instance extracted from the text. Other attributes identify the lateralization, negation and modifiers of the *Condition* in analysis.

The classes *Time* and *Value* are shown in Figure 4.13. Instances of the class *Value* can describe a measurement, as for instance, the measure value of the patient's heart rate. Over the course of time, a patient can have multiple *Condition* episodes or perform several *Procedures*, and have different *Evolution*s. The class *Time* captures the temporal information of the discharge letter. The relations between this class and the others allow to specify the exact

Figure 4.10: Relationships between the class *Procedure* and the classes *Condition*, *Anatomical Site*, *Modifier*, *Negation*. *Time* and *Value* in the MEDINX ontology.

time period in which the event occurred, adding a time stamp to the associated information.

The classes *Negation*, *Lateralization* and *Modifier* are presented in Figure 4.14. These classes provide further characterization to the instances of the other classes. They describe, for instance, that an *Evolution* can be *Negated*, a *Condition* has a *Lateralization* (e.g., right, left, upper and so forth) or a *Modifier* (e.g., acute, small, distinct, etc.).

Figure 4.15 describes the classes *Anatomical Site*, *Administration Route*, *Evolution* and the class *Episode*. The class *Evolution* characterizes the patient progression by reflecting the changes of the inpatient's health condition. This class refers to the references in text of the supervision and clinical evolution of the inpatient or its specific *Condition*. The references to the route of administration of a *Chemical Procedure* are captured in the class *Administration Route*. Finally, the class *Anatomical Site* intends to characterize the anatomical structures or body locations where, normally, is the locus of a *Condition*.

MEDINX ontology was created using Protégé. The achieved hierarchy is presented in Figure 4.16. The figure shows the attributes of the class *Episode*, as the **diagnosis code description** or the **episode text** and also the relationships of this class with the classes *Anatomical Site*, *Condition*, *Evolution* and *Procedure*.

Figure 4.11: Properties of the class *Procedure* and its subclasses. Figure 4.12: Properties and subclasses of *Condition*.

The population of the ontology was performed in a semi-automatic form, using the information extracted by MEDINX. This process is described in more detail in Section 6.3.2.

Figure 4.13: Detail of the class *Time* and *Value*. Figure 4.14: *Negation*, *Lateralization* and *Modifier* classes. Figure 4.15: *Administration Route*, *Evolution*, *Anatomical Site* and *Episode* classes.

4.6 Summary

The success of a system closely correlates with the coverage of the required domain knowledge, which is made available to the system as knowledge sources. The UMLS Metathesaurus contains an abundance of nomenclatures, controlled vocabularies and ontologies, available in several languages, including some translations to Portuguese. However, the information existing in such resources is not enough to extract relevant clinical entities

Figure 4.16: MEDINX ontology in Protégé

from discharge summaries written in Portuguese, as are the ones processed in MEDINX.

MEDINX needs to build up representations of clinical facts and events, to assign domain-specific meanings to terms and use these meanings in their operations. For this purpose four new ontologies were developed in this work: two formalizations of the international classification systems ICD-9-CM and ICF, one ontology of the drugs currently in use in Portugal, and a conceptualization of the structure and content of the PDLs used in this work.

The current chapter presented the methods used to create and populate the ontologies. The resultant conceptualizations are easily extensible by adding additional concepts and relationships, and are, therefore, reusable in different situations. In particular, the MEDINX ontology has multiple use cases, as the identification of patients with similar disease progression or the summarization of the inpatient condition evolution.

Chapter 5

MEDical INformation eXtraction system

Architecture and Components

The analysis of natural language documents is a complex task, usually, not carried out in a single, large step. Rather, it is normally deal at several steps of processing. This approach has also been followed in MEDINX, the medical information extraction system described in this thesis.

MEDINX is tailored to the clinical domain and contains components to extract hypertension-specific characteristics from unstructured PDLs. However, the system was also used to recognize and categorize named entities in general domain newspaper texts. In order to be adaptable to several domains, MEDINX architecture was designed to be flexible and include different components in its processing. The next sections describe the architecture of MEDINX and decompose the grand picture of MEDINX in its subprocesses.

5.1 MEDINX architecture

The main goal of MEDINX is to extract clinical entities and their relations from free-form text discharge narratives written in Portuguese. Given so, MEDINX components are based on NLP principles, and provide several mechanisms to read, process and utilize external resources, such as terminologies and ontologies.

The system architecture is based on a set of annotators that mark up unstructured textual documents, inserting *annotations* that can be associated with a particular piece of text. A subsequent annotator can read and process all previously created annotations. These annotators were aggregated into a set of components, that represent the main processing

steps of the system.

Figure 5.1: System Architecture

Figure 5.1 illustrates MEDINX architecture, in which is possible to identify the following components:

1. **Document Reader:** a component which converts an XML file into plain text and extracts implicit meaning from the structure of the document, by converting the embedded tags of the input document into annotations;
2. **General natural language processing:** components for sentence discovery, tokenization and part-of-speech tagging;
3. **REMMIX:** the NER component of MEDINX, composed of three other components:
 - **Context dependent annotator**, an annotator that creates annotations from one or more tokens, using regular expressions and surrounding tokens as clues;
 - **Concept finding**, a component which extracts concepts based on specified terminologies and ontologies, and determines negation, lateralization and modifiers;
 - **Relation extraction**, a component which extracts relations between concepts using contextual information;
4. **CAS Consumers:** responsible for ending the process and creating the desired output. The three main consumers of MEDINX are:
 - **XML consumer**, which produces an XML file with the annotations of the previous annotators;
 - **Ontology population**, that populates the MEDINX ontology; this component produces an OWL file with the information contained in the CAS;

- **Template filling**, which outputs the knowledge extracted from the narratives to a template containing information about the patient and its health related state, with the correspondent ICF codes.

MEDINX components run within the UIMA framework (see section 3.2.1) and make use of its *type system* definition to define the kinds of meta-data that can be identified by the system, as well as, the types each annotator expects to input and the types it inserts in the CAS, as output.

Figure 5.2 presents part of the UIMA type system inheritance diagram of MEDINX. This diagram, fully presented in Appendix B, presents the hierarchy of UIMA types used in the system, as well as their attributes. The type system hierarchy reflects the information (e.g., sentences, tokens, entities, etc.) transmitted through the components of the system and will be referred in the next sections, while present the system.

Figure 5.2: Part of MEDINX UIMA type system inheritance diagram.

5.2 Document Reader

The first step required to extract information from free-form text documents is to connect to and iterate through the document collection, acquiring the documents and initializing the

CAS. This task is performed by the *Document Reader* component, a collection reader which converts embedded tags of an input document into annotations.

In the case of the PDLs, these tags correspond to the *Segments* of the narratives (see PDL structure in 3.1.3). As part of this conversion, *Segment* annotations are added to the CAS. Simultaneously, the *Document Reader* adds document level information, such as the unique identification of the document and other available header information to the CAS, as the DRG and ICD-9-CM codes and the respective descriptions.

This component is also responsible for reading the MEDINX ontology and add its information to the CAS in order to be used in subsequent processing. Jena's ontology API was used to create the ontology model and map its content to each of the UIMA *Ontology Concept* types (see Figure 5.2).

5.3 General natural language processing

Before any analysis can take place, the basic sentences and tokens existing in the texts (e.g., words, acronyms, abbreviations, numbers, punctuation symbols and so forth) have to be identified. These operations, together with the POS tagging, are performed in the MEDINX general NLP component.

Most of the currently available tokenizers and sentence splitters accomplish their task by implementing simple tools based on white spaces and punctuation symbols detection. However, such simple heuristics easily run into several problems:

- Abbreviations - words are not always separated from other tokens by white space. A period may also signal an abbreviation (e.g., *Vol.* or *etc.*) in which case it has to be distinguished from a sentence-delimiting period. This issue is even more relevant if such items appear at the end of a sentence, as in, for instance, **Sem edemas no M.I.** (*Without edemas in the L.L.*) as the full stop is here performing double duty. This situation can interfere with sentence boundary detection.
- Multiple Formats - numbers may occur in multiple formats containing ambiguous separators, for example, 123.456,78 or 123,456.78 or even 123 456.78. Dates and procedure results are also frequently written under different formats.
- Sentence boundary detection - sentences are usually demarcated by the typical sentence delimiting punctuation marks (e.g., period, exclamation mark, or question mark). Sometimes, however, other punctuation symbols, or even no punctuation symbols, are used to indicate sentence boundaries.

These issues were dealt in MEDINX in the several stages of its pipeline. MEDINX uses a simple white space tokenizer and punctuation sentence splitter, based on the one provided in the UIMA examples package. However, and in order to ensure the transmission of correct information to the next stages, an *Abbreviation annotator* and a new sentence splitter, an *Abbreviation aware Sentence Splitter*, were included in the general NLP component.

Abbreviation Annotator: The MEDINX *abbreviation annotator* uses a manually compiled list of the most common abbreviations present in the MEDINX development set (approximately 130 abbreviations with their expanded form). This list, partially presented in Table 5.1, is used to identify the abbreviations present in text, which are then added to the CAS.

Table 5.1: Acronyms and abbreviations commonly used in the MEDINX development set.

ABBREVIATION	EXPANSION	<i>English Gloss</i>
AVC	Acidente Vascular Cerebral	<i>Stroke</i>
ausc.	auscultação	<i>auscultation</i>
sist.	sistólico	<i>systolic</i>
Amox.	Amoxicilina	<i>Amoxicillin</i>
BM	Boehringer Mannheim	<i>Boehringer Mannheim</i>
MIIs	Membros Inferiores	<i>Lower Limbs</i>
Abd	Abdominal	<i>Abdominal</i>
AC	Auscultação Cardíaca	<i>Cardiac Auscultation</i>

Abbreviation aware Sentence Splitter: Having recognized this shorthand units, there is still the need to develop a new sentence splitter capable of processing the previously identified annotations. This sentence splitter uses the information added to the CAS by the *abbreviation annotator* to help distinguish sentence-marking full stops from other kinds, being this punctuation at the end of the sentence or not. In particular, it analyses the tokens that appear before the sentence full stop and determines if it is an abbreviation. It also considers previous information added to the CAS by the *Document Reader* component as the *Segment* limits in order to determine, more accurately, the end of a sentence.

The last module of the general NLP component is the POS tagger. The use of standard POS taggers, parameterized through the use of newspaper corpora, i.e., general language domain corpora, originates a noticeable loss of performance when applied in sublanguage domains, as it is the case of clinical text data (Campbell and Johnson, 2001). In fact, several corpora were already developed to train POS taggers with clinical narratives written in English (Liu *et al.*, 2007; Hahn and Wermter, 2004). However, such resources are not

available for Portuguese and their development is not in the scope of this work. Given so, the information retrieved by the POS tagger was used *carefully* within MEDINX components, being reserved for the situations where no other reliable source could be used or for punctuation detection.

Several POS taggers, following different approaches and developed for several languages, are currently available (Brill, 1992; Almeida and Pinto, 1995; Barreto *et al.*, 2006). MEDINX uses a POS tagger to which an UIMA implementation was already developed and tested: the TreeTagger (Schmid, 1995), a language independent tool for annotating text with part-of-speech and lemma information developed at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger is a probabilistic tool based on decision trees and is currently available for several languages, including Portuguese (Gamallo, 2011). The UIMA implementation of the TreeTagger is based on a wrapper of the tool, developed by the UKP lab of the Technical University of Darmstadt, and available in the DKPro repository (Gurevych and Müller, 2008). The necessary alterations were performed in order to deal with text in Portuguese, as for instance, the adaptation of the tag set.

5.4 REMMIX, the Named Entity Recognizer of MEDINX

Reconhecimento de Entidades Mencionadas do MedInX (REMMIX) is the NER component of MEDINX. REMMIX concentrates on the later stages of IE, i.e., takes the linguistic objects as input and finds domain-dependent classifications and relations among them.

REMMIX is first presented in the context of its participation in the Second HAREM contest, i.e., as a **general domain** NER. This first experiment and contact to the task of NER in Portuguese allowed us to direct REMMIX to the medical area in a better and more informed manner, as described in section 5.4.2.

5.4.1 REMMIX in the Second HAREM contest

The second HAREM contest distinguishing features (introduced in Section 2.2.2.1), particularly the separation between properly identifying and classifying NEs, were taken in consideration while developing the REMMIX annotators. Consequently, the HAREM's version of REMMIX is composed of two different main modules: the *Candidates Generator* and the *Named Entities Classifier*. Figure 5.3 gives an overview of REMMIX pipeline in HAREM II.

The first module, the *Candidates Generator*, is responsible for the identification of candidate expressions, henceforth candidates. Candidates are defined in HAREM's guidelines as all sets of terms with initial capital letter. The *Candidates Generator* uses regular

Figure 5.3: REMMIX pipeline in HAREM II

expressions to identify tokens with initial capital letter, taking in consideration the presence of connecting terms with a maximum length of 5 characters (eg., *e*, *em*, *de*, *da*, *do*, *dos*, *das*, *para*, etc.). An example of an expression identified by the *Candidates Generator* is the expression *Presidente da República* (*President of the Republic*). Expressions containing numbers were not covered in this stage and left to be identified and processed by the TIME and VALUE annotators.

REMMIX *Named Entities Classifier* is responsible for categorizing the identified candidates, according to HAREM's guidelines. This module uses two distinct classification approaches. The first is based on very simple rules and manually constructed gazetteers, while the second approach uses information extracted from Wikipedia pages. These two methods can be held jointly or separately. Using two different approaches allowed to understand the advantages and disadvantages of each method, and also to realize how exactly Wikipedia helps enhancing the results.

The TIME and VALUE annotators, responsible for the classification of the temporal and numerical expressions, together with the LOWERCASE annotator, end the *Named Entities Classifier* processing. These annotators are always used by REMMIX, independently of the method used in the previous processing (gazetteers and rules or Wikipedia.).

5.4.1.1 Classification using rules and gazetteers

The first classification approach is based on manually constructed gazetteers in combination with a set of contextual rules. The gazetteers have been developed by us in the context of a project related with IE from neurological reports (Ferreira *et al.*, 2006) and contain lists of expressions belonging to several semantic classes like person names, cities, diseases and symptoms.

Several contextual rules were also created. These allow to categorize a candidate as a function of itself and, eventually, of a portion of text that precedes it (left context) and/or of a portion of text that follows it (right context). These rules use semantically rich words, *signal words*, as the ones describing titles, places, organisation names and others. Table 5.2 lists some examples of words used to annotate the entities belonging to the classes PERSON, LOCAL, ORGANIZATION and EVENT.

Table 5.2: Examples and total number of signal words used in the definition of contextual rules.

CATEGORY	N	SIGNAL WORD	<i>English gloss</i>
PERSON	110	Ministro Chefe Princesa Reitora ...	<i>Minister</i> <i>Head</i> <i>Princess</i> <i>Dean</i>
LOCAL	58	Praça Avenida Rua Cidade ...	<i>Square</i> <i>Avenue</i> <i>Street</i> <i>City</i>
ORGANIZATION	50	Museu Faculdade Sindicato Prefeitura ...	<i>Museum</i> <i>Faculty</i> <i>Union</i> <i>Prefecture</i>
EVENT	28	Campeonato Concerto Congresso Exposição ...	<i>Championship</i> <i>Concert</i> <i>Congress</i> <i>Exhibition</i>

The annotators used in this task (Figure 5.3) begin by splitting the candidate expression into their various terms. They assign a semantic category (and type, if appropriate) if the terms exist in the gazetteers. When this is not achieved, they apply the contextual rules by looking at the candidates and their context for signal words.

One example of the coverage of this module is presented in Example 14:

- (14) **Input:** Ao que parece, Paulo Pinto Mascarenhas tem a convicção firme (...)
 (*'Apparently, Paulo Pinto Mascarenhas has a strong conviction (...).'*)
- Output:** Ao que parece,
 <EM ID='xxx' CATEG='PESSOA' TIPO='INDIVIDUAL'>
 Paulo Pinto Mascarenhas

 tem a convicção firme (...)
 (*'Apparently,*
 <EM ID='xxx' CATEG='PERSON' TIPO='INDIVIDUAL'>
 Paulo Pinto Mascarenhas

has a strong conviction (...).')
 where the name *Paulo* exists in the gazetteer.

5.4.1.2 Classification using Wikipedia

Building and maintaining high-quality gazetteers is an extremely time consuming task. Many methods have been proposed to solve this problem. The automatic extraction of gazetteers from large amounts of texts is an example of a proposed solution (Toral and Munoz, 2006). In MEDINX, the development of large gazetteers, with a wide coverage of the language and capable of recognizing general domain NEs, was not an option due to time and human resources constraints. Given so, Wikipedia was used as an external knowledge source (Bunescu and Pasca, 2006). Although Wikipedia cannot be directly used as a gazetteer, extracting knowledge from Wikipedia is much simpler than from raw texts or from usual Web texts because of its structure.

The Wikipedia annotator of REMMIX uses the first sentence of a Wikipedia article to extract the category label. For example, consider the sentence:

Concluiu os seus estudos de medicina, em 1870, na Universidade de Harvard, onde iniciou a sua carreira como professor de fisiologia em 1872.
 (*He completed his medical studies in 1870, in Harvard University, where he began his career as a Physiology Professor in 1872.*)

The previous annotator, the *Candidates Generator*, starts by identifying the expression

`Universidade de Harvard` as a candidate NE. To get the Wikipedia description of this expression, the candidate is converted into a Wikipedia entity name through the concatenation of the expression words with an underscore, as in `Universidade_de_Harvard` (see Section 3.1.1). Next, the corresponding Wikipedia article is retrieved from the MEDINX Wikipedia dumps database¹, following redirections until a non-redirection page is found.

Although there is no strict formatting rule in Wikipedia, the convention is to start an article with a short sentence defining the entity the article describes. For example, the article for `Universidade_de_Harvard` starts with the sentence:

- (15) *A Universidade Harvard (em inglês Harvard University) é uma das instituições educacionais mais prestigiadas do mundo, bem como a instituição de ensino superior mais antiga dos Estados Unidos da América.*
(Harvard University is one of most prestigious educational institutions in the world, as well as the oldest higher education institution in the United States of America.)

Most of the times, the first sentence of the article contains a *signal expression* which indicates the semantic category of the entity. In the example the signal word is *instituição* (*institution*). The method used in MEDINX concentrates, therefore, in the extraction of the first sentence from an article and the enclosed signal expressions. Unnecessary markups, such as italics, bold faces, and internal links, were eliminated from the article. The markup for internal links, in particular, was appropriately converted into the form displayed to the user. The article was then divided into lines according to the new line code, some HTML tags as `<br>` and a very simple sentence segmentation rule for period. After obtaining the first sentence, simple rules, similar to those used in the classification based on gazetteers and contextual rules, were used. In particular, the first sentence of the Wikipedia article was inspected for signal words that indicate the semantic class of the entity. Some examples of these words, as well as the number of words used by this annotator, are listed in Table 5.3.

An example output of the *Wikipedia annotator*, retrieved from Second HAREM golden collection is given in example 16.

- (16) **Input:** A popularidade do piloto Ayrton Senna na França era comparável à de seu maior rival, Alain Prost, quatro vezes campeão mundial.
(‘The popularity of the pilot Ayrton Senna in France was comparable to the one of his rival, Alain Prost, the four times world champion.’)

¹There are pages for other than usual articles in the Wikipedia data. They are distinguished by a namespace attribute. To retrieve articles, only the namespace 0 was searched, which is the one used for usual articles.

Table 5.3: Examples and total number of signal words used in the extraction of a category from an Wikipedia’s article first sentence.

CATEGORY	N	SIGNAL WORD	<i>English gloss</i>
PERSON	15	imperador engenheiro professor piloto ...	<i>emperor</i> <i>engineer</i> <i>professor</i> <i>pilot</i>
LOCAL	15	planeta cidade ilha continente ...	<i>planet</i> <i>city</i> <i>island</i> <i>continent</i>
ORGANIZATION	12	partido movimento universidade ...	<i>party</i> <i>movement</i> <i>university</i>
EVENT	2	acordo competição	<i>agreement</i> <i>competition</i>

Output: A popularidade do piloto
 <EM ID='xxx' CATEG='PESSOA' TIPO='INDIVIDUAL'>
 Ayrton Senna
 na França era comparável à de seu maior rival,
 <EM ID='yyy' CATEG='PESSOA' TIPO='INDIVIDUAL'>
 Alain Prost
 , quatro vezes campeão mundial.
 (*The popularity of the pilot*
 <EM ID='xxx' CATEG='PERSON' TIPO='INDIVIDUAL'>
 Ayrton Senna
 in France was comparable to the one of his rival,
 <EM ID='yyy' CATEG='PERSON' TIPO='INDIVIDUAL'>
 Alain Prost
 , the four times world champion.)

The *Wikipedia annotator* does not use disambiguation pages, since accurate disambiguation is difficult and can introduces noise. As explained in Section 3.1.1, there are two popular ways for presenting ambiguous entities in Wikipedia. The first is to redirect users to a disambiguation page, and the second is to redirect users to one of the articles. REMMIX’s *Wikipedia annotator* focus on the second case. This method is simple but works well because the article presented in the second case represents, in most cases, the principal meaning of

the ambiguous entities and, therefore, the meaning that most frequently appears in the texts.

Wikipedia articles also allow the association of categories with the entity being described. However, it was observed that an article can have more than one category, and often these categories are not clear hypernyms of an entity. Thus, information regarding categories of an article was also not used in REMMIX.

5.4.1.3 TIME, VALUE and LOWERCASE annotators

The TIME, VALUE and LOWERCASE annotators are the last group of annotators of the pipeline. In particular, the TIME and VALUE annotators start by identifying expressions that contain at least one digit or a word listed in the related gazetteers. For instance, the TIME annotator identifies holiday names and year seasons (*Easter, Carnival, Spring, Summer, ...*), week days, months, frequency adverbs (*daily, every year, ...*), etc. The VALUE annotator uses the gazetteers for unit names, like *metro, Kg, Gb, etc.* and money *euros, dólares, contos, etc.* These annotators also use regular expressions to identify temporal and numerical entities as, for example, *em 25 de Abril [de 1974]*.

The LOWERCASE annotator expands the label given to an entity in order to include the lowercase preceding term, in case it exists and belongs to the list defined in HAREM guidelines. The full list of allowed lowercase words is given in (Mota and Santos, 2008a, page 285). However, the LOWERCASE annotator was also developed in order to analyze the unclassified candidates, i.e., if the identified candidate has not been previously tagged, the precedent lowercase term is analyzed and the appropriate tag is attributed to the expression '**lowercase + candidate**'.

Examples 17 and 18 show, respectively, the output of the LOWERCASE and TIME annotators. In the Example 18, the expression *ex-presidente* is included in the list available in the HAREM guidelines.

(17) **Input:** As fortes chuvas que atingiram ontem (...)

Output: As fortes chuvas que atingiram

<EM ID='xxx' CATEG='TEMPO' TIPO= 'TEMPO_CALEND'>

ontem (...)

(18) **Input:**(...) quando o ex-presidente José Sarney disse que sua maior missão era conduzir o país até às eleições.

Output:(...) quando o

<EM ID='xxx' CATEG='PESSOA' TIPO='CARGO'>

ex-presidente José Sarney disse que sua maior missão era conduzir o país até às eleições.

The results and knowledge obtained with the participation of REMMIX in the Second HAREM contest were used to develop the current version of REMMIX, capable of detecting and categorizing relevant clinical entities and their relations in the PDLs.

5.4.2 REMMIX in Clinical Text

With the participation of REMMIX in the Second HAREM workshop the use of gazetteers in the task of NER was overcome with the use of Wikipedia. In the task of NER from clinical narratives, the creation of useful gazetteers would be even more demanding, as it requires detailed medical expertise. Also, medical terminology changes and evolves constantly. Maintaining such lists would possibly compromise the goals determined for MEDINX, in particular, its scalability, flexibility or coverage. Instead, and following the former approach of using external resources as main knowledge sources, the UMLS Metathesaurus and the ontologies created in this project were used as knowledge resources in this version of REMMIX.

The semantic types of the UMLS Semantic Network were used in MEDINX to related the concepts mentioned in the PDLs with the ones described in UMLS Metathesaurus. Table 5.4 lists the semantic groups identified in UMLS which are relevant to the texts analysed in MEDINX, as well as, their semantic types.

Table 5.4: UMLS semantic types used in MEDINX.

UMLS SEMANTIC GROUPS	UMLS SEMANTIC TYPES
Anatomy	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, Tissue
Disorders	Acquired Abnormality, Anatomical Abnormality, 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

The UMLS database was queried in order to find all the Portuguese terms associated with the concepts categorized by the semantic groups listed in Table 5.4. The information extracted was indexed using Lucene library indexing capabilities. Three different indexes were created, one for each UMLS Semantic group, containing a total of 54,682 different concepts (1,911 anatomical concepts, 42,759 disorders and 10,759 procedures). All the information

included in the ICD-9-CM and Drugs ontologies was also indexed using Lucene. This indexed information was then used as knowledge source in the following MEDINX components. The use of an index, instead of direct access to the resources, ensures a rapid and efficient system performance.

To accomplish the task of extracting information from the PDLs three main components were created: the *Concept Dependent annotator*, *Concept Finding* and the *Relation Extractor*. Each one of these components will be described in the next paragraphs. The REMMIX pipeline used to extract information from the PDLs is illustrated in Figure 5.4.

Figure 5.4: REMMIX pipeline in the clinical setting

5.4.2.1 Context Dependent Annotator

Several problems have been previously analysed regarding the use of simple heuristics to identify the basic units in texts, i.e., the tokens and the sentences. One of the identified problems is related with the use of multiple formats containing ambiguous separators while describing numbers (e.g., 123,456.78 or 123 456,78), dates (e.g., 30/09/2009 or 30-09-09 or 30 Set. 2009), procedure results (e.g., FC - 85bpm, AP-N), and so forth.

The *Context dependent annotator* of MEDINX creates annotations from one or more tokens, using the surrounding tokens as clues. An example of an annotation created from multiple tokens is a range that includes two numbers, a unit of measurement and a dash, as

in 12-12h.

5.4.2.2 Concept Finding

Concept Finding is one of the most critical components of the system. It associates textual mentions to the concepts of the external knowledge sources used in the processing, in order to attribute appropriate categories to the text excerpts.

Several annotators, using different approaches to IE, compose the *Concept Finding*, as illustrated in Figure 5.4.

1. Ontology based annotator: The task of the MEDINX *ontology annotator* is particularly challenging. It classifies the entities in the text with regards to the classes pre-specified by a concept hierarchy.

The MEDINX *ontology annotator* uses the structured conceptualized knowledge described in the MEDINX ontology. In particular, this annotator uses the knowledge provided by the ontology to recognize matching expressions in text, and create the corresponding annotations in the CAS. In case of ambiguity, i.e., in the case the term is associated with two or more concepts (polysemy), the expression being analysed is annotated as belonging to all associated categories. Disambiguating such terms is the task of a subsequent annotator, the *Disambiguation annotator*.

The main idea of this kind of ontology based approach is that the more information is added to the ontology (in a automatic way with the structured output of the system) more will be the information that the annotator will recognize. The use of an ontology driven approach to IE is meant to enhance the semantics of the system, since, even when using a simple method as pattern or string matching from the content of the ontology, the extracted pieces are interpreted with respect to a predefined knowledge model.

2. Lookup annotators: The lookup annotators are a group of UIMA annotators developed to recognize clinical entities through lookup in the external knowledge sources, as the UMLS indexes or the drugs and ICD-9-CM ontologies. Several annotators were developed (see Figure 5.4) to recognize different information in texts, as the *Condition*, *Procedure* and *Anatomical Site* annotators, among others. They use the information added to the CAS by the general NLP component to define a window of tokens to which a matching algorithm should be tested.

The lookup annotators start their processing by looking the first token in text and, in

case this is not listed as a stop word², expands its search for the next token in context, until an interval of four tokens is reached. This value was chosen after analysing the PDLs and realizing that no concept in the reports is described using more than four tokens. In case a match is triggered before the analysis of the entire window of tokens (the four tokens) the annotator saves the matching expression. However, it does not create the final annotation until all the tokens of the interval have been analysed.

For instance, consider the sentence given in Example 19:

- (19) **Acidente Vascular Cerebral antigo, ...**
(Previous Cerebrovascular Accident, ...)

The *lookup annotators* start by analysing the first token *Acidente*, which is a polysemic word. In fact, a simple search in UMLS for this word retrieves 33 results, containing references to *Conditions, Therapeutic or Preventive Procedure* (e.g., prophylaxis against stroke) and even a reference to an *Health Care Related Organization*, the *Instituto Nacional de Distúrbios Neurológicos e do AVC dos Estados Unidos da América* (the *United States National Institute of Neurological Disorders and Stroke*). Thus, and despite matching several concepts in UMLS, the annotators continue their analysis, including the next token in text in the search expression: *Acidente Vascular*. This expression is also matched with several UMLS and ICD-9-CM terms, including the desired expression *Acidente Vascular Cerebral*.

When using the three tokens interval, the expression *Acidente Vascular Cerebral* is analysed and a full match is found in both resources, the UMLS and ICD-9-CM ontology. The match is then saved and the next token *antigo* is added to the search. However, the full expression *Acidente Vascular Cerebral antigo* has no matching results and the annotator marks the spam of text corresponding to the largest previous match. In this case, *Acidente Vascular Cerebral* is classified as belonging to the category *Condition*.

In general, several parameters can be defined while using these annotators:

1. Case matching;
2. Use of lemma;
3. The definition of words that should be skipped (stop words and/or of words with several semantic categories, i.e., polysemic words).
4. Order of the used resources;

²List containing some of the most common, short function words in the PDLs, such as a/o *the*, depois *after*, durante *during*, nas *in/on*, etc

5. The context within a matching string is executed (e.g., interval of tokens, sentence, etc.);
6. The starting point of the matching algorithm (e.g., at every token, after the last match, etc.)

These parameters bring more flexibility to the system components and ensure the capability of using different resources in their processing. For instance, the definition of a list of polysemic words reduces the ambiguity associated with these terms, since the annotators only attribute a category to an entity in case of a full expression match.

An example of this process is given by the word *bloqueio* (*block*), commonly used when referring to a patient *Condition* (e.g., *intraventricular block*, *induction of neuromuscular block*, *heart block*, etc.). However, this word is also used in several other contexts as, for instance, to describe *Therapeutic or Preventive Procedures*, as in *ring block* or *digital block*. An expression containing such word is only semantically categorized in MEDINX in case there is a match of the complete expression with the definition used in the resources.

The lookup annotators also save several information retrieved from the knowledge source used to recognize the expression, as the name of the resource, the expanded version of the text (in case it contains an abbreviation or an acronym) and the CUI (or ICD-9-CM) code of the entity when it is recognized in the UMLS Metathesaurus (or ICD-9-CM ontology).

This set of annotators also use regular expressions in their analysis, which, in conjunction with the information extracted from the external knowledge sources or from simple lists containing, for instance, measurement units or temporal adverbs (similar to the ones used in the HAREM II), attribute categories to textual mentions describing values, sizes, dates and other temporal and numerical expressions.

Negation annotator: One particularly relevant annotator, included in the lookup annotators group, is the *Negation annotator*. Actually, when extracting information from clinical text it is important, not only to recognize clinical entities and relations but also to determine whether these words are negated or not. This is specially relevant in the analysis of patient records. When describing the status of a patient, the physician often reasons by excluding various possible diagnoses and symptoms. Most of the work on detecting negations has been carried out in English (Pyysalo *et al.*, 2006), and short or even no work has been done for other languages, as it is the case of Portuguese.

In MEDINX, negation is interpreted as a relationship between the entity NEGATION and the other entities in text. The *Negation annotator* of MEDINX uses a generalized algorithm, based on the popular NegEx (Chapman *et al.*, 2001). Negation words are specified in a list and used to discover negated phrases within a specific window. In MEDINX trigger words,

such as *sem* (*without*), *nunca* (*never*), *não* (*not*), *excepto* (*except*), *suspende* (*stops*), are first identified in the *Negation annotator*, while the task of relating them with the negated concepts was performed in the *Relations Extraction annotator*.

3. Disambiguation annotator: The previous characterizations of the content of the PDLs allowed to understand the frequent use of polysemic words in the documents, which are treated in the previous annotators as if they would belong to all the semantic categories they are associated with. The *Disambiguation annotator* is responsible for attributing one, and only one, semantic category to such entities. This annotator uses the information concerning the *Segment* of the text in which these entities are enclosed in order to disambiguate their sense.

For instance, the token “IV” is frequently used in the PDLs with two different senses:

- as a VALUE indicating the strength of the patient’s limb, as in *Força muscular III/IV*, or
- indicating the *Administration Route* of a given *Chemical Procedure*, i.e., as an acronym of *Intravenoso* (*Intravenous*).

However, the *Segment* of the PDLs in which the token was referred can be help distinguishing both senses. For example, if IV is mentioned in the context of a *Physical Examination* it will, most probably, refer the result of a *Procedure* and will be annotated as a belonging to the category *Value*. However, when IV is used in the context of the *Therapeutics* performed, it will be annotated as an *Administration Route* of a given drug, since it is certainly related with the *Chemical Procedures* described in the PDL.

5.4.2.3 Relations Extraction annotator

The next step in the pipeline is to discover the relations between the several entities identified and categorize by the previous annotators. The relations between the clinical entities have already been identified and modeled within MEDINX ontology (see Table 4.2). Consequently, this information can be used as a preexisting knowledge base defining that, for instance, an entity of the category *Condition* is **negated** by an entity of the category *Negation*, and so forth.

Two kinds of relations can be roughly distinguished in MEDINX: the hierarchical ones, namely, the **is-a** relation between an entity of a given type and its main category, and the ones between the entities of different categories. The first type of relations are already handled in UIMA through the definition of the hierarchical type system, i.e., the type system

configuration used defines in advance the hierarchical relations, such as *Dosage is a Value* or *Disease is a Condition* (as represented in Figure B.1).

The remaining relations were introduced as attributes in the appropriate UIMA types corresponding to the entity category. The type system defined for each category allows to define the relations of the type with the other types in analyses (see Figure B.1).

Figure 5.5: Part of the MEDINX type system inheritance diagram.

The MEDINX *Relation Extraction annotator* was designed to search the CAS for annotations belonging to the related categories within a predefined window of text. In MEDINX the window was defined according to the punctuation of the text. Specifically, the annotator looks for related entities within a comma separated excerpt of a sentence. In case two related entities exist in that excerpt, the CAS annotation of the entity belonging to the first argument of the relation (see Table 4.2) is updated with the information regarding the extracted relation.

For example, in the sentence:

- (20) Vai medicada com Trivicum 100 id.
 (*She goes medicated with Trivicum 100 id.*)

the previous annotators are responsible for classifying the word *trivicum* as a MEDICATION, *100* as a DOSAGE and *id*³ as a FREQUENCY.

The *Relation Extraction annotator* will then set the appropriate relations between the entities within the defined window (the entire sentence in this case since there is no comma punctuation) taking into consideration the relations defined for each category in the UIMA type system. In the example given, the extracted relations would be:

1. *Trivicum* hasDosage *100*;
2. *Trivicum* hasFrequency *id*.

The attributes `hasDosage` and `hasFrequency` of the annotation correspondent to the MEDICATION entity *trivicum* would then be updated to include the fact that the entity is related with the DOSAGE entity *100* and with the FREQUENCY entity *id*.

5.5 Consumers

A consumer in UIMA is the component responsible for the extraction of the previous analyses results and, also, to select the information that will be part of the output. This information can be used, for instance, to populate the MEDINX ontology or to fill predefined templates.

In MEDINX three main consumers were developed: the *XML consumer*, the *ontology population consumer* and the *template filling consumer*.

The *XML consumer* extracts the identified annotations from the CAS, being those the clinical entities or the relations between these, and writes them in to an XML formatted document.

Two versions of this consumer were developed: the first follows the HAREM XML output format and was used to produce the results of REMMIX in the Second HAREM. The second version is the one used to create the annotated versions of the PDLs to be used in the evaluation of the system.

The HAREM version of the XML consumer was designed to be able to handle alternative annotations, `<ALT>`. This consumer determines the existence of two or more enclosed annotations and tags the corresponding entities within two or more labels separated by `|`, as illustrated in Example 21.

³*id*: daily, from the Latin “in die”

```
(21) <ALT>
      <EM ID="xxx" CATEG="PESSOA" TIPO="CARGO">líder do Sinn Fein</ EM>
      |
      líder do <EM ID="yyy" CATEG="ORGANIZACAO" TIPO="INSTITUICAO">Sinn
      Fein</EM>
      </ALT>
```

The *XML consumer* used to produce the annotated PDLs has, on the other and, the capability of handling the relations between entities as shown in Example 22.

```
(22) Doente com <TIME ID="T1">antecedentes</TIME> de
      <CONDITION TYPE="DISEASE" ID="D1" ABB="hipertensao" HASTIME="T1"
      HASMODIFIER="M1">
      HTA
      </CONDITION>
      <TIME ID="T2">à longa data</TIME>,
      <MODIFIER ID="M1">mal controlada</MODIFIER>.4
```

The *ontology population consumer* transforms the information concerning the clinical entities and relations into a valid OWL file. This consumer uses the model of the ontology developed using the Jena ontology API (see Section 5.2) to create new instances of the ontology and fill the associated properties. This reuse of the previously defined model ensures that the consistency of the ontology is maintained. The created OWL file corresponds to the MEDINX ontology populated with the new information extracted from the PDLs by the system.

The last consumer developed, the *template filling consumer* produces an HTML document with a summary of the health and health related status of the patient, according to the information extracted from the PDL, and encoded according to the ICF classification system. The application of this consumer will be further explored in Chapter 7.

5.6 Summary

This chapter introduced the architecture of the MLP system described in this thesis, MEDINX. Despite being originally intended to be exclusively used as a MLP system, MEDINX has a flexible and adaptable architecture capable of extracting knowledge from other domains, as it is evident with its participation in the Second HAREM contest.

⁴The orthographic error “*à longa data*” is intentionally kept, since it was present in the PDL from where the example sentence was selected.

MEDINX is based on an open source framework, UIMA, and its components use NLP principles to discover significant (clinical) entities and relations in free text documents. The system provides also a mechanism to ingest, process and automatically populate the MEDINX ontology with the extracted information and to use the information contained in several other ontologies in the IE task.

Several methods are used by the MEDINX components to extract information. Semantically rich words, rules, regular expressions are some of the most basic methods used. However, it is through the use of the MEDINX ontology of PDLs that the system is able to assign domain-specific meanings to the terms used in the documents and use these meanings to build the structured representation of the PDLs free-text content.

Evaluation

The evaluation of MEDINX was performed at two distinct phases of its development: the first evaluation took place at the time of its participation in the Second HAREM contest, in 2008, while the evaluation of the system on the task of extracting clinical entities from PDLs was performed between October 2010 and February 2011.

The current chapter presents the results obtained on each evaluation period and, also the combination of the system annotators used in the evaluations, particularly of REMMIX, the main component of the system.

6.1 Evaluation Metrics

The results presented in this chapter were computed according to the standard metrics: precision, recall and F-measure, frequently used in the evaluation of IE systems (Makhoul *et al.*, 1999). In particular, precision is calculated as the ratio of the relevant findings (the entities and relations extracted) in all findings of the system (see eq. 6.1), whereas, the recall is calculated as the ratio of relevant findings within the total numbers of all expected findings (eq. 6.2). The F-measure is the weighted average of the precision and recall (eq. 6.3). In the evaluation process of MEDINX the traditional F-measure (also called F_1 score), representing the harmonic mean of precision and recall, was used (eq. 6.4).

$$precision = \frac{\#ofCorrectValuesExtracted}{\#ofValuesExtracted} \quad (6.1)$$

$$recall = \frac{\#ofCorrectValuesExtracted}{\#ofAllCorrectValues} \quad (6.2)$$

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

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

More specific metrics have been used in some phases of the evaluation and will be further described in the corresponding Sections.

6.2 MEDINX results in Second HAREM contest

The participation of MEDINX in the Second HAREM was mainly intended to understand the benefits of using semi-structured knowledge resources, such as Wikipedia, in the task of NER, instead of creating and maintaining large gazetteers, but also, to evaluate the performance of MEDINX when compared to the state of the art systems in Portuguese NER. The results obtained by MEDINX allowed to answer some of the existing questions, as shown in the following paragraphs.

In HAREM each participant was allowed to submit until three different *runs*¹ of their system. The following three combinations of REMMIX were used:

1. **Gazetteers run:** created using the annotators based on gazetteers and contextual rules;
2. **Wikipedia run:** using only the Wikipedia annotator;
3. **Both run:** which uses, as the name indicates, all developed annotators; first, the annotators based on gazetteers and contextual rules are called and for the remaining unclassified candidates, the Wikipedia annotator is used.

All REMMIX runs extract information related with temporal expressions and values using the TIME and VALUE annotators, and follow HAREM's recommendation to identify and classify the precedent lowercase expressions, by using the LOWERCASE annotator. The configuration of the system in each of the runs is illustrated in Figure 6.1.

In HAREM the participant systems were given the possibility of choosing a subset of categories in which they wanted to be evaluated. MEDINX participated in the contest in a **selective** scenario by classifying all the HAREM categories, but not considering their type.

¹different combination of the system components.

Figure 6.1: REMMIX runs in HAREM II

The results presented in the following paragraphs were evaluated according to the HAREM evaluation modes:

- **strict:** taking into account all the possible alternatives, <ALT>, on a text segment, having each <ALT> a weight equal to the inverse of the number of alternatives in that segment;
- **relaxed:** taking into account the elements of <ALT> which maximize the system classification.

6.2.1 Results

The use of Wikipedia has potential for better performance?

Table 6.1 presents the results of the three different runs in the task of semantic classification with the *strict* and *relaxed* <ALT> evaluation.

Table 6.1: MEDINX results in the task of semantic classification.						
Run	Strict ALT evaluation			Relaxed ALT evaluation		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Gazetteers	61.32 %	29.52 %	0.3985	63.40 %	30.84 %	0.4150
Wiki	58.08 %	23.16 %	0.3312	59.50 %	24.09 %	0.3429
Both	60.50 %	36.15 %	0.4526	62.26 %	37.50 %	0.4681

The analysis of table 6.1, particularly the F-measure values, allows to identify the best

results of the system as the ones obtained while using all the developed annotators, i.e., gazetteers in combination with the contextual rules and Wikipedia, in the *Both* run. This observation is valid in both assessments of <ALT>, the strict and relaxed evaluations. This higher F-measure value is obtained due to the higher recall of this configuration ($\sim 13\%$ higher recall than in the *Wikipedia* run and $\sim 7\%$ more than the *Gazetteers* run, in both evaluations), despite a slight loss of precision when compared to the *Gazetteers* run (less than 1%). The worst results are obtained with the *Wikipedia* run. This is valid for all metrics and in both assessments of ALT. In fact, when compared with the *Both* run the F-measure value decreases, approximately, 12%. The results obtained with the use of gazetteers are higher than the ones obtained when using only Wikipedia, being, however, still lower than the results of the system when using all annotators. When compared to the best run, the *Both* run, it is observed a decrease of, approximately, 5% in the F-measure in both evaluations.

Are all categories equal for Wikipedia?

Table 6.2 presents the results obtained on the semantic classification task for each of the categories considered in Second HAREM.

Table 6.2: MEDINX results for each category in the task of semantic classification.

Category	Best run	Classification		
		Precision	Recall	F-measure
LOCAL	Both	57.00 %	50.89 %	0.5377
PERSON	Both	66.66 %	36.77 %	0.4740
VALUE	Both = Gazetteers	35.89 %	52.02 %	0.4247
ORGANIZATION	Both	58.29 %	23.97 %	0.3397
TIME	Both = Gazetteers	47.44 %	25.38 %	0.3307
EVENT	Both	40.44 %	14.73 %	0.3159
WORK	Both	51.46 %	12.12 %	0.1962
ABSTRACTION	Both = Gazetteers	22.31 %	3.92 %	0.0667
THING	Both = Gazetteers	22.27 %	3.18 %	0.0557

The table indicates a better performance when using Wikipedia in the classification of the categories LOCAL, PERSON, ORGANIZATION, EVENT and WORK, and the use of this knowledge source does not seem to affect the classification of the categories ABSTRACTION and THING. The results for the categories VALUE and TIME are independent of the use of Wikipedia combined with the gazetteers or the use the gazetteers separately. It is observed a decrease of $\sim 1\%$ in the F-measure when classifying these categories only with Wikipedia.

Figure 6.2 presents the distribution of MEDINX results in terms of precision and recall for each of the system runs and categories.

The figure indicates the higher precision of the system when compared to its recall values. It can also be observed that the worst results were obtained for the categories ABSTRACTION

Figure 6.2: Precision and Recall in the task of semantic classification for each run and category.

and THING. The figure also shows that the categories with best recall are LOCAL and VALUE, while the entities of the category PERSON were recognized with best precision.

Is this approach competitive?

Figure 6.3 presents the results obtained by MEDINX for each category, when normalized according to the maximum and minimum performance values of HAREM, i.e., the results were normalized according to the results of the best and worst system for each category. From this analysis we can place the system results in a scale of 0 to 100%, where 100% indicates that our system performed best.

Such analysis shows, for instance, that MEDINX obtained the highest precision when

Figure 6.3: Normalized comparison according to the minimum and maximum values.

classifying the category ABSTRACTION (100%), while the recall and F-measure for this category were the lowest obtained in all participant systems (0%). In general, the image indicates the high precision of MEDINX, having reached a normalized mean value of 71.9%.

Comparison with REMBRANDT

The comparison of the results between the participant systems is not, in general, a straightforward task since each system participated in a different selective scenario. However, in the final meeting of HAREM, in the Second HAREM workshop, we realized the existence of a system with a similar approach to the one followed in MEDINX, the REMBRANDT (Cardoso, 2008). This system uses Wikipedia categories to classify each of

HAREM's categories and participated in the total HAREM scenario.

The comparison with the results published by Cardoso (2008) shows that the F-measure of MEDINX is, in general, lower than the one obtained by REMBRANDT (approximately 11% lower in the best run of both systems), being, however, the precision of the systems very similar (the precision of MEDINX is approximately 3% lower).

When analysing individually each of the categories, MEDINX obtained an higher F-measure ($\sim 2\%$ higher) when classifying the category EVENT and similar values in the category THING. In the remaining categories the F-measure of MEDINX is lower than the one obtained by REMBRANDT. However, when analyzing the results of both systems in terms of precision (the most important metric due to the system main area of application, the clinical area), the differences decrease, with MEDINX obtaining higher precision in the semantic classification of the categories LOCAL, ORGANIZATION, ABSTRACTION and THING.

6.2.2 Discussion

The results obtained by MEDINX suggest that using Wikipedia to extract semantic categories is useful in the task of NER, even if performed by a simple method as the one presented. However, the Wikipedia seems to be relevant when used together with other gazetteers and contextual rules, even if simple, as the ones used in MEDINX. A deeper use of the internal structures of Wikipedia, like its categories, might result in several improvements in the results obtained with this annotator.

When analysing the results of each category individually one can observe that the categories ABSTRACTION and THING do not benefit from the use of Wikipedia. This fact can be related with the higher ambiguity associated with this categories and consequently an higher difficulty while categorizing them. Another factor that can influence these results is the insufficient number of words used to extract the classification from the first sentence of the Wikipedia article.

Generally, it can be concluded that MEDINX is precise, presenting competitive values when compared to the other participant systems. The comparison with REMBRANDT in particular, suggests comparable precisions, while the recall of MEDINX is, in general, lower. The lowest recall values indicate the need to develop more and more complex rules to identify the candidate NEs, as well as the methods to solve conflicts and ambiguities. It is, however, noteworthy that MEDINX is a system intended to process medical reports, a domain where precision has an additional relevance. A system capable of correctly extracting information from PDLs is extremely relevant, more than the extraction of several information with possible noise.

In general, the participation and evaluation of MEDINX in HAREM, even if directed to general domain newspaper texts, allowed us to place the system in the state of the art of Portuguese NER systems and to concretize some experiments regarding the use of semi-structured external knowledge resources as Wikipedia. It was also possible to identify specific areas for improvement, as explained in this discussion. The results obtained encouraged the development of similar, but improved, solutions to the clinical area.

6.3 MEDINX results in Patient Discharge Letters

The evaluation of MEDINX in the task of IE from the authentic PDLs created in HIP, was performed between the 28th of October 2010 and the 28th of February 2011, and is the focus of the current section.

6.3.1 System setting

Similarly to the methodology used in the first evaluation of MEDINX, two distinct system settings were used to extract information from the PDLs, i.e., two *runs*. The first, is mainly based on rules and uses the knowledge contained in MEDINX drugs and ICD-9-CM ontologies, together with the Portuguese vocabularies of UMLS Metathesaurus, to extract significant entities and their relations from the texts. The second system setting uses the structure and the content of the MEDINX ontology in an ontology-driven IE approach. These runs and the components of MEDINX used in each, are represented in Figure 6.4 and can be summarized as follows:

1. **Resources run:** created using the knowledge contained in the available Portuguese vocabularies of UMLS, together with MEDINX drugs and ICD ontologies; and
2. **Ontology+Resources run:** created using the above sources and the semi-automatically populated MEDINX ontology;

In order to use the content of the MEDINX ontology in the analysis of the PDLs, the developed knowledge representation model had to be populated. This process was performed following a semi-automatic approach. Namely, 30 PDLs from the MEDINX development set were automatically analysed using MEDINX *Resources* run setting and its results were written to an output XML file. The file was then manually inspected to find incorrect categories but, mainly, to identify and manually classify the missing entities. This validated file was the input of the automatically ontology population process, performed with Jena API and the model of the ontology previously created.

Figure 6.4: REMMIX runs in PDLs processing.

The *Ontology+Resources* run is thus designed to use all the annotators of the *Resources* run, together with the *Ontology annotator*, capable of processing the information of the MEDINX ontology. This run is intended to evaluate the impact on the results, in terms of precision and recall, of the use of the MEDINX ontology in the system results.

6.3.2 Evaluation Setup

In order to evaluate an IE system, its facts need to be compared to some ground truth. Chapter 3, presented the efforts carried out to create a manually annotated corpus of PDLs, to which MEDINX results could be automatically compared, the MEDINX gold standard. That chapter also presented the difficulties experienced in this task, making it still, at the date of writing, an ongoing task. Given so, and since there was no computer-processable ground truth of proper dimensions to which MEDINX results could be compared to, the evaluation of MEDINX was based in the manual inspection of the automatic results in order to determine their correctness.

For this purpose, an evaluation interface was developed in cooperation with an expert on Web 2.0 technologies and Human-Computer Interaction² to allow a more user friendly visualization and evaluation of the annotations created on each report.

The evaluation process was divided into two phases, as illustrated in Figure 6.5. In the first phase, the *precision evaluation*, the judges were asked to assess whether the facts (clinical

²Our thanks to Klaus Schaefer for his time and availability.

Figure 6.5: Evaluation interface.

entities and their relations) extracted by MEDINX were correct. Each of the extracted entities was shown to the judge, and a visual representation of the relations of each entity was given to help in the task. This process is illustrated in Figure 6.6, where one can observe the entities and relations extracted by the system in the sentence *Ischemic stroke - recent ischemic stroke of the posterior cerebral artery*.

To evaluate the classification given by the system to each entity, the judges could select between the options “Correct”, “Incorrect” or “Don’t know”. In the case of a relationship a “Correct” checkbox was made available. This process is illustrated in Figure 6.7.

In the second phase, intended to assess the recall of the system, the judges were asked to mark up in text the entities not found by the system, and chose their appropriate category. This process is represented in Figure 6.8. A detailed evaluation manual was developed to guide the judges, describing all the necessary steps required to a successful evaluation. This manual is presented in Appendix C.

The 86 documents of the MEDINX test set, were automatically annotated using both settings of the system, the *Resources* and *Ontology+Resources* runs. The judges were asked to evaluate the outcome of this process, i.e., the 172 documents originated by the double annotation of each one of the 86 PDLs of the test set. These annotated PDLs were presented to the judges without any specific order. It was, however, explained the need to evaluate them in the order they appear in the interface, to ensure that the documents go through, at least, a double evaluation process, and, in the best case, each PDL is evaluated by several judges.

Figure 6.6: Detail of the precision assessment: the entities and respective relations extracted from the sentence *Ischemic stroke - recent ischemic stroke of the posterior cerebral artery*.

Figure 6.7: Evaluating entities and relations.

The judges were also asked to ensure that for each precision assessment, the corresponding recall assessment was performed, in order to have complete metrics on each report.

Figure 6.8: Recall assessment.

6.3.3 Judges Expertise

Eight persons from different expertise areas were asked to participate in the evaluation of MEDINX. These judges were selected by convenience and mainly because of their accessibility and proximity. However, at the end of the evaluation period, limited to four months due to the time constraints associated with this type of work, only seven judges had evaluated part of the test set. More specifically, the judges group was constituted by two Computer Scientists (Reviewer1 and Reviewer4), one Linguist (Reviewer2), one Radiologist (Reviewer5), two Psychologists (Reviewers 6 and 7) and one Physician (Reviewer3).

6.3.4 Evaluation

The 172 annotated PDLs were made available online for evaluation during a period of 4 months. In the end of this period, 50 different PDLs have been reviewed, from which 30 belonged to the *Ontology+Resources* run and 20 to the *Resources* run.

The number of judges that evaluated each PDL varies, as shown in Table 6.3. In fact, only two judges, Reviewer4 and Reviewer5, evaluated the 50 documents and only one PDL has been evaluated by all the seven different judges. The total number of entities evaluated in each phase of the assessment by each judge is also represented in Table 6.3 presenting Reviewer5 as the judge which evaluated more entities.

Some descriptive statistics about the number of entities extracted by each of MEDINX runs are presented in Tables 6.4 and 6.5. Table 6.4 presents the values for the *Ontology+Resources* run while Table 6.5 presents the number of entities extracted with the *Resources* run. These tables present the total number of entities extracted in each category (N), but also the average (Mean), the standard deviation (Std. Dev.), the minimum (Min) and maximum (Max) number

Table 6.3: Number of PDLs evaluated by each judge.

Judge	Ontology+Resources	Resources	Total		
	Number of PDLs	Number of PDLs	Number of PDLs	Number of Entities	
				Precision assessment	Recall assessment
Reviewer1	0	1	1	75	2
Reviewer2	4	6	10	893	106
Reviewer3	2	7	9	840	35
Reviewer4	20	30	50	4,159	424
Reviewer5	20	30	50	4,159	487
Reviewer6	4	6	10	893	44
Reviewer7	3	9	12	1 062	98

of entities identified in the PDLs. More detailed statistics, including the the first and third quartile(Q_1 and Q_3), are presented in Appendix D.

Table 6.4: Descriptive statistics for the amount of entities extracted from the 30 reports of the *Ontology+Resources* run.

Category	N	Mean	Std. Dev.	Min	Max
ADMINISTRATION ROUTE	86	3	4	0	13
ANATOMICAL SITE	99	3	2	0	10
CHEMICAL	456	15	7	0	33
CONDITION	764	25	11	2	55
EVOLUTION	22	1	1	0	2
LATERALIZATION	113	4	3	0	12
MODIFIER	227	8	5	1	19
NEGATION	128	4	3	0	12
PROCEDURE	328	11	8	0	40
TIME	301	10	5	0	21
VALUE	331	11	8	0	36

Both tables indicate that the entities with more representativity in the PDLs are the ones belonging to the categories CONDITION and CHEMICAL. In fact, in all analysed PDL and in both runs, was always identified at least one CONDITION. On the other hand, the entities less present in the PDLs belong to the category EVOLUTION, being in average referred in each report only once.

The same analysis can be performed regarding the relations identified by the system in each run. Tables 6.6 and 6.7 present the number of relations extracted from the PDLs and show the relations **hasModifier**, **hasTime** and **hasNegation** as the most frequently identified relations in the PDLs with the *Ontology+Resources* configuration (Table 6.6). In particular, the most frequent relation identified with this run is the **hasModifier** relation, which has

Table 6.5: Descriptive statistics for the amount of entities extracted from the 20 reports of the *Resources* run.

Category	N	Mean	Std. Dev.	Min	Max
ADMINISTRATION ROUTE	46	2	3	0	10
ANATOMICAL SITE	15	1	1	0	3
CHEMICAL	245	12	5	0	18
CONDITION	294	15	8	1	33
EVOLUTION	15	1	1	0	2
LATERALIZATION	54	3	2	0	7
MODIFIER	109	5	4	0	14
NEGATION	73	4	3	0	10
PROCEDURE	126	6	3	0	14
TIME	185	9	5	0	17
VALUE	142	7	5	0	17

been identified in all the evaluated PDLs at least once, and is present in each documents in average six times.

Table 6.6: Descriptive statistics for the amount of relations extracted from the 30 reports of the *Ontology+Resources* run.

Relation	N	Mean	Std. Dev.	Min	Max
HASADMINISTRATIONROUTE	69	2	3	0	13
HASDOSAGE	57	2	2	0	7
HASDURATION	1	0	0	0	1
HASEVOLUTION	6	0	0	0	1
HASFINDING	59	2	2	0	13
HASFREQUENCY	101	3	3	0	9
HASINDICATION	23	1	1	0	3
HASLATERALIZATION	82	3	2	0	10
HASMODIFIER	169	6	3	1	16
HASNEGATION	116	4	3	0	11
HASQUANTITY	78	3	3	0	11
HASSUBSTANCE	1	0	0	0	1
HASTARGET	71	2	2	0	8
HASTIME	156	5	4	0	15
HASVALUE	91	3	3	0	16
INCLUDES	3	0	0	0	1

In the *Resources* run, the three most common relations are the `hasTime`, `hasModifier` and `hasFrequency`. The `hasTime` relation has been identified 80 times in the set of 20 PDLs evaluated with the *Resources* run.

Figure 6.9 illustrates the total number of entities evaluated in the precision assessment of each category, in both runs and by all the judges. It indicates, in general, that more entities

Table 6.7: Descriptive statistics for the amount of relations extracted from the 20 reports of the *Resources* run.

Relation	N	Mean	Std. Dev.	Min	Max
HASADMINISTRATIONROUTE	35	2	3	0	8
HASDOSAGE	35	2	2	0	6
HASDURATION	1	0	0	0	1
HASEVOLUTION	2	0	0	0	1
HASFINDING	15	1	1	0	3
HASFREQUENCY	58	3	3	0	9
HASINDICATION	12	1	1	0	4
HASLATERALIZATION	34	2	1	0	5
HASMODIFIER	64	3	2	0	8
HASNEGATION	46	2	2	0	7
HASQUANTITY	27	1	2	0	6
HASSUBSTANCE	1	0	0	0	1
HASTARGET	12	1	1	0	2
HASTIME	80	4	3	0	12
HASVALUE	33	2	1	0	3
INCLUDES	4	0	0	0	1

from the *Ontology+Resources* run have been evaluated. This is valid for all the categories analyzed. The figure also allows to understand that a similar amount of entities from the categories VALUE and PROCEDURE were identified in the documents and evaluated.

To evaluate the task of automatic identification and category assignment to each entity, the judges were asked to choose between the options *Correct*, *Incorrect* and *Don't know* (Figure 6.7). The distribution of the evaluation, for each of the available options and for each judge is presented in Figure 6.10. The figure indicates a high precision of the system, independently of the considered run. This result will further be analysed in the next paragraphs with the analysis of the precision, recall and F-measure of the system in the tasks of semantic classification and relation extraction.

6.3.5 Interjudge agreement

The agreement between judges is usually quantified by the k statistic (Fleiss, 2000), which is the chance-corrected interjudge agreement, or, more specifically, the agreement among the judges above the expected by chance. However, the use of k has been controversial (Hripcsak and Heitjan, 2002) due to the difficulty interpreting its level. In many NLP related tasks and in NER, in particular, this statistic is particularly not appropriate (Hripcsak and Rothschild, 2005) given the impossibility of calculating the true negative case count. In fact, when manually evaluating clinical entities from text, the number of true negatives in the text

Figure 6.9: Frequency of each category for *Ontology+Resources* and *Resources* runs

Figure 6.10: Evaluation of each reviewer of the *Resources* and *Ontology+Resources* runs

(i.e., words that were correctly not annotated by the system) is poorly defined, because the entities may overlap or vary in length, and the number is expected to be very large,

bringing the expected chance agreement close to zero. Given so, the interjudge agreement was computed as a percentage agreement given the total number of evaluated annotations where the judges agree over the total number of annotations made by the system, and by the Cramer's V (Cramér, 1999) association measure.

The agreement was calculated as a two-way agreement between the judges that annotated more PDL, Reviewer4 and Reviewer5 and also as seven-way agreement between all the judges on the annotation of the single PDL evaluated by all judges. Table 6.8 presents the results obtained for each of the categories considered.

Table 6.8: Inter-judge agreement.

Category	Double judgment		Septuple judgment	
	Cramer's V*	% Agreement	Cramer's V*	% Agreement
ADMINISTRATION ROUTE	0.97	100.0	0.86	100.0
ANATOMICAL SITE	0.55	98.9	0.71	100.0
CHEMICAL	0.62	99.7	0.32	97.8
CONDITION	0.58	95.1	0.62	96.9
EVOLUTION	0.97	100.0	0.97	100.0
LATERALIZATION	0.65	98.2	0.97	100.0
MODIFIER	0.56	96.9	0.51	96.4
NEGATION	0.67	99.2	0.55	96.4
PROCEDURE	0.45	93.9	0.71	100.0
TIME	0.69	98.3	0.52	97.1
VALUE	0.68	97.9	0.97	100.0
Total	0.64	97.7	0.70	98.5

* $p < 0.05$

In general, the agreement results indicate strong agreement between the judges. The percentage of agreement is, in all cases, greater than 93%, while the overall Cramer's V is, in both cases, greater than 0.6 (with $p < 0.05$), showing substantial and significant agreement. The categories to which was reached less agreement are CONDITION in both calculations of the statistic, and PROCEDURE when computing the agreement in the evaluation of the 50 double evaluated PDLs. The entities of the categories ADMINISTRATION ROUTE and EVOLUTION have reached 100% agreement between all the judges and in all the evaluated PDLs.

6.3.6 Evaluation Results

6.3.6.1 Overall Precision, Recall and F-measure

The overall results of MEDINX in terms of precision, recall and F-measure are presented in Table 6.9.

These results indicate high precision of the system in both runs, being its values in both

Table 6.9: Overall results of MEDINX in the task of semantic classification.

Run	Results (%)		
	Precision	Recall	F-measure
<i>Resources</i>	94.48	82.48	86.70
<i>Ontology+Resources</i>	95.92	93.44	94.53

cases close to 95%. In the *Ontology+Resources* run the system obtained similar values in the precision and recall assessment, while in the *Resources* run the recall of the system is, approximately, 11% lower than the precision. The difference in the values of recall from the *Resources* run to the *Ontology+Resources* run is also $\sim 11\%$.

To evaluate the possibility that the difference obtained in the two runs is due to chance, statistical hypothesis testing has been used. The probability p has been calculated according to the χ^2 distribution in order to test the null hypothesis, H_0 , that the two runs do not differ. A lower p indicates the lower probability that the null hypothesis holds. In our case, the lower the significance level, the more likely it is that the results of the two runs are significantly different. In this case $p = 0.35$ ($\chi^2 = 0.88$; $df = 2$) which is greater than the significance level $\alpha = 0.05$. Consequently, H_0 can not be rejected.

Figure 6.11 presents, separately, the boxplots of the precision and recall of the system in each of the runs. The figure shows the distribution of the results when computed according to the evaluation of each PDL (representing the group *Report* in the figure) and to the evaluation performed by each judge (group *Reviewer*). The filled squares represent the values obtained for each element of the group, i.e., each PDL, when considering the *Report* group, or each judge, when considering the *Reviewer* group. The left side of the Figure shows the precision distribution, while the recall values are in the right side.

The boxplots indicate once again the similar precision of the system in both runs and an improvement in terms of recall from the *Resources* run to the *Ontology+Resources* run. When analysing the results grouped by *Report*, the Doc6006286 stands out as one of the analyzed PDLs with highest precision in both runs; while when analysing the same results of the *Reviewers* group, the Reviewer7 is the judge which attributed to the system the highest precision. The lowest precision values are obtained in the Doc700242 and by Reviewer3 in the *Resources* run, while in the *Ontology+Resources* run this rank corresponds to Doc8017363 and the Reviewer1.

The distribution of the results, represented in the figure with the filled black boxes, allows to understand that the evaluation performed by each judge as some degree of variability. This is particularly visible in the evaluation of the precision of the *Resources* run, where the results

Figure 6.11: Distribution of the Precision and Recall of MEDINX in the task of semantic classification, grouped by *Reports* and *Reviewers*.

attributed by each judge are very disperse in the boxplot.

The distribution of the recall results allows to identify the PDLs and the judges which evaluated the system with highest and lowest recall values. In particular, Doc6009911 was evaluated as having 100% recall in the *Resources* run, while Doc6005095 and Doc8011932 were assessed equally in the *Ontology+Resources* run. Reviewer6 has been in both runs the judge with highest recall assessment. On the other hand, Doc5014478 and Reviewer2, and Doc8015753 and Reviewer5 have obtained the lowest recall values, in both *Resources* and *Ontology+Resources* runs, respectively.

A similar analysis was performed for the F-measure of the system and is presented in Figure 6.12. The image illustrates the higher values of the system when evaluated with the *Ontology+Resources* run. The image indicates that the automatic semantic categorization of the entities of Doc6009911 have been evaluated with 100% F-measure in the *Resources* run. The PDL with lowest F-measure in this run was Doc6013389. In both runs, the judge that evaluated the system F-measure with highest value was Reviewer6, while the lowest values were obtained with the evaluations of Reviewer2 in the *Resources* run and Reviewer5 in the *Ontology+Resources* run.

Figure 6.12: F-measure of MEDINX

6.3.6.2 Precision, Recall and F-measure for each Category

Table 6.10 presents the results of the best run of the system in the task of semantic categorization, for each of the categories considered.

The highest classified category is NEGATION with 99.5% precision, 100% recall and 99.8% F-measure. The best run for this category was the *Resources* run, although the values obtained are very similar to the ones obtained in the *Ontology+Resources* run. The run *Ontology+Resources* was, in all other cases, the one with better results.

The obtained p-value according to the χ^2 distribution was, in this case, less than the significance level $\alpha = 0.05$ (in the *Ontology+Resources* run: $\chi^2 = 299.7$; $df = 20$; $p < 0.05$; in the *Resources* run: $\chi^2 = 145.9$; $df = 20$; $p < 0.05$), which allows to reject the null hypothesis (H_0 : there is no difference between categories) with low probability of error and indicates that the results obtained in the different categories are statistically significant in both runs.

Table 6.10: MEDINX results for each category.

Category	Best run	Results (%)		
		Precision	Recall	F-measure
ADMINISTRATIONROUTE	Ontology+Resources	99.7	98.3	99.0
ANATOMICALSITE	Ontology+Resources	86.5	70.2	77.5
CHEMICAL	Ontology+Resources	98.4	96.5	97.4
CONDITION	Ontology+Resources	91.7	95.5	93.6
EVOLUTION	Ontology+Resources	100.0	93.9	96.9
LATERALIZATION	Ontology+Resources	98.8	93.6	96.1
MODIFIER	Ontology+Resources	97.0	85.7	91.0
NEGATION	Resources	99.5	100.0	99.8
PROCEDURE	Ontology+Resources	89.5	96.5	92.9
TIME	Ontology+Resources	98.8	98.8	98.8
VALUE	Ontology+Resources	95.4	99.3	97.3

The next figures present the precision (Figure 6.13), recall (Figure 6.14) and F-measure (Figure 6.15) of the system, in both runs, and for each one of the considered categories.

Figure 6.13: Precision for each category in the *Resources* and *Ontology+Resources* runs

While Figure 6.13 indicates a similar precision of the system in both runs, a clear difference is visualized in Figure 6.14, representing the recall values. All categories except VALUE and NEGATION have improved their recall values in the *Ontology+Resources* run. In the referred categories, the absolute differences are under 1%. This improvement is particularly clear in the categories ANATOMICAL SITE and CONDITION with absolute differences of 46.4% and 29.3% from the *Resources* to the *Ontology+Resources* run. Also the categories PROCEDURE and MODIFIER present high recall improvement between both runs with a difference of,

Figure 6.14: Recall for each category in the *Resources* and *Ontology+Resources* runs

respectively, 17.5% and 14.0% from the *Resources* to the *Ontology+Resources* run.

The F-measure values for each category are represented in Figure 6.15. The categories ANATOMICAL SITE, MODIFIER and PROCEDURE are the categories with lowest F-measure, while NEGATION, ADMINISTRATION ROUTE and TIME present the highest F-measure values.

Figure 6.15: F-measure for each category in the *Resources* and *Ontology+Resources* runs

Figures 6.16 and 6.17 present the distribution of the results in terms of precision and recall for both runs. The dispersion of the values is also illustrated in the image with the

help of the boxplots presented on the top and right side of each image. Besides the clear improvements in the results of the categories ANATOMICAL SITE, CONDITION and MODIFIER from the RESOURCES run to the *Ontology+Resources* run, the boxplots allow to characterize the results obtained in the *Ontology+Resources* run as less disperse, when comparing to the ones obtained in the *Resources* run.

Figure 6.16: Precision x Recall in the *Resources* run

6.3.6.3 Precision of the Relations Extraction task

To evaluate the relations extracted by the system, the judges were asked to determine if the visualized relation was correct or not (Figure 6.7). Given so, as the missing relations were not identified, only the precision results of the relation extraction task are analysed in the following paragraphs.

The results were corrected for entity recognition in order to separate both tasks and ensure that the results do not reflect double penalizations: only the relations between two correctly

Figure 6.17: Precision x Recall in the *Ontology+Resources* run

categorized entities were considered in precision calculation.

Table 6.11 presents the precision of MEDINX in the relation extraction task, for both runs. In general, the precision obtained in both runs is very similar and, approximately, 95%.

The precision of the best run of each of the relations is presented in Table 6.12. The Table identifies several relations assessed at 100%, having those reached this value in both runs of the system. In general, better results are obtained with the *Ontology+Resources* setting. One can observe that the relations with lowest precision are the **hasIndication**, with 69.6%, and **hasTarget**, with 77.5% precision. This is also visible in Figure 6.18, which also indicates that the relation **hasIndication** was the one with highest improvement with the use of the *Ontology+Resources* setting, with a difference of 19.6%.

The boxplots of the precision values obtained in the relation extraction task, are illustrated in Figure 6.19, grouped in terms of *Reports* and *Reviewers*. Very similar results can be

Table 6.11: MEDINX results in the relation extraction task.

Run	Precision (%)
<i>Resources</i>	94.76
<i>Ontology+Resources</i>	94.94

Table 6.12: MEDINX results for each relationship.

Category	Best run	Precision (%)
HASADMINISTRATIONROUTE	Ontology+Resources = Resources	100.0
HASDOSAGE	Ontology+Resources = Resources	100.0
HASDURATION	Ontology+Resources = Resources	100.0
HASEVOLUTION	Ontology+Resources = Resources	100.0
HASFINDING	Resources	93.3
HASFREQUENCY	Ontology+Resources = Resources	100.0
HASINDICATION	Ontology+Resources	69.6
HASLATERALIZATION	Ontology+Resources	96.3
HASMODIFIER	Ontology+Resources	93.5
HASNEGATION	Ontology+Resources	100.0
HASQUANTITY	Ontology+Resources = Resources	100.0
HASSUBSTANCE	Ontology+Resources = Resources	100.0
HASTARGET	Ontology+Resources	77.5
HASTIME	Ontology+Resources	85.9
HASVALUE	Ontology+Resources	97.8
INCLUDES	Ontology+Resources = Resources	100.0

Figure 6.18: Precision for each relationship in the *Resources* and *Ontology+Resources* run

analysed in this image, when considering the different runs. The lowest precision was obtained in the evaluation performed by Reviewer4. In the *Resources* run several PDLs and judges obtained 100% precision; Doc5011036 and Reviewer2 are an example of such result.

In the *Ontology+Resources* run, the highest values were obtained in a particular PDL, the Doc6006286. Reviewer5 was the judge that evaluated the precision of the system as highest in this run. The opposite result was obtained by Reviewer4.

Figure 6.19: Precision on the task of Relation Extraction.

6.3.7 Intrajudge Agreement

The reliability and the consistency of a given judge is given by the intrajudge agreement. Given the time constraints associated with the development of a doctoral thesis, the intrajudge agreement assessment was based on a simple test-retest method. In particular, two judges, namely, Reviewer2 and Reviewer5, were asked to re-evaluate two of the automatically annotated PDLs one month after the end of the evaluation period. This re-evaluation was intended to quantify the coherence between the results obtained in both periods of evaluation.

The selected documents, Doc7002484 and Doc7003767, have been randomly retrieved from the group of the PDLs evaluated by both reviewers and belong, respectively, to the *Ontology+Resources* and *Resources* runs. The computed values for the precision, recall and F-measure in both assessment periods are presented in Table 6.13.

The table indicates that the reviewers have been very coherent in their assessment, with

Table 6.13: Intrajudge agreement.

PDL	Reviewer	1st Evaluation (%)			2nd Evaluation (%)		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Doc7002484	<i>Reviewer2</i>	96.0	96.2	96.1	97.3	96.2	96.8
	<i>Reviewer4</i>	97.3	98.7	98.0	98.0	98.7	98.5
Doc7003767	<i>Reviewer2</i>	90.7	79.6	84.8	90.7	81.2	85.6
	<i>Reviewer4</i>	96.0	86.5	91.0	96.0	80.4	87.5

similar evaluation in both periods. The F-measures obtained in both evaluation periods are similar, with a difference of less than 1% in the first and second evaluation of Doc7002484. Reviewer2 reached a higher agreement in the evaluation of Doc7003767, with an F-measure difference in both assessments of less than 1%, while Reviewer4 had a difference of 3.5% in both assessments. These values are, nevertheless, low and indicate good agreement.

The percentage agreement given the total number of evaluated annotations where the judges agree in both evaluation periods, over the total number of annotations of the documents, is given in Table 6.14, together with the values obtained for the Cramer's V association measure.

Table 6.14: Intrajudge agreement for each of MEDINX categories.

Category	Reviewer2		Reviewer4	
	Cramer's V*	% Agreement	Cramer's V*	% Agreement
ADMINISTRATION ROUTE	0.86	100.0	0.86	100.0
ANATOMICAL SITE	0.71	100.0	0.71	100.0
CHEMICAL	0.49	99.4	0.37	97.8
CONDITION	0.61	95.1	0.45	93.9
EVOLUTION	0.97	100.0	0.97	100.0
LATERALIZATION	0.97	100.0	0.97	100.0
MODIFIER	0.71	100.0	0.71	100.0
NEGATION	0.71	100.0	0.71	100.0
PROCEDURE	0.71	100.0	0.71	100.0
TIME	0.48	97.1	0.71	100.0
VALUE	0.97	100.0	0.97	100.0
Total	0.73	99.5	0.74	99.3

* $p < 0,05$

The table shows that there was substantial and significant agreement of the judges in the two evaluation periods and for each of the categories considered. The percentage of agreement obtained was always above 93% and the values obtained for Cramer's V measure are, in general, above 0.6, indicating high association between the results obtained in the two evaluation periods.

6.3.8 Discussion

MEDINX evaluation on the set of authentic PDLs was performed through human manual inspection of the annotated PDLs of the MEDINX test set. Seven judges from different expertise areas performed the task of categorizing each of the automatic annotations assigned by MEDINX. However, manual inspection of such kind of documents is an extremely demanding task, even if developed in a user friendly interface like the one used in MEDINX. This difficulty was reflected in the number of documents evaluated, only 50 of the originally presented 172, and in the lack of recall assessment in the task of relation extraction.

Two different configurations of REMMIX were evaluated: one designed to use the information available the clinical knowledge representations used by the system, the *Resources* run, and the other using the MEDINX ontology. This separation was intended to assess the advantages of using a clinical ontology, as MEDINX ontology, to process clinical text.

The general results, using the usual metrics of Precision, Recall and F-measure, indicate a very good performance of the system. Although the comparison with similar systems developed for other languages is not straightforward, MEDINX is clearly at the level of the international state of the art systems. The overall recall and precision of the system of, approximately, 96% and 93% in MEDINX best run, allows to position MEDINX as one of the best clinical information extraction systems. MedLEE, one of the similar systems with best results worldwide, presents 81% recall and 91% precision in the identification of six medical conditions from radiology reports.

Precision has been, however, a major concern while developing the architecture and components that constitute the MLP system. Being MEDINX intended to process clinical text, it is extremely relevant that it is capable of correctly extracting the entities and relations described in text. The overall precision of 96% reflects the approach followed in the system design to ensure the extraction of reliable information. Also important is the capability of MEDINX of automatically populating the PDLs knowledge representation model used in its several applications. Only a precise system is capable of producing a correct, consistent and concise ontology. However, the completeness of such ontology is only ensured by a system with high recall. The values obtained in MEDINX evaluation allow to ensure that the ontology follows the standard principles considered during its creation.

When analysing the results obtained for each of the considered categories, the ANATOMICAL SITE category stands as the category with the lowest results in both precision and recall assessment of the system. Several problems have been found in the entities belonging to this category. The UMLS Metathesaurus has been used in the extraction of such entities and the information present in this knowledge source concerning the semantic category *Anatomy*, is reduced and extremely diverse. In fact, UMLS does not contain any translation

to Portuguese of the vocabularies related with this semantic type and, consequently, the information used is noisy and compromised the system performance. This result suggests the need to use more knowledge resources developed specifically to Portuguese and to process clinical reports.

In the task of relation extraction the system obtained a precision of 100% in most of the relations extracted by the system. Despite the use of a very simple method to extract information from the clinical texts, an overall precision of 95% was reached in the best run. However, some of the worst results obtained demonstrate that the method is not capable of leading with the lack of punctuation, common in some of the PDLs segments. For instance, the *Physical Examination* segment contains information regarding the physical procedures performed and the results of the patient. This is also the segment where the relations with worst precision, the **hasIndication** and **hasTarget** relations, are most commonly found. However, physicians often use in this segment ungrammatical language (based on pseudo-lists and tables) to describe the patient's information (see Section 3.1.3.3). Without the presence of punctuation, the MEDINX *relation extractor* can not specify a correct window for its processing and, consequently, extracts incorrect relations among the modeled data.

6.4 Summary

This chapter presented the methodology used to evaluate MEDINX. Two different evaluations were performed, the first developed at the time of the participation of MEDINX in the Second HAREM contest, and the second on the MEDINX test set. The system evaluation on the set of clinical discharge summaries was performed by human manual inspection of the system output. Seven judges, from different expertise areas, as Computer Science, Medicine, Radiology, Linguistic and Psychology, participated in this task.

The results obtained indicate, in general, that MEDINX is a very precise system having reached a precision of 63% in the task of IE from general domain newspaper texts and 95% in the task of IE from clinical discharge documents. The results analysis indicate that the creation of a formal representation of the information contained in the PDLs was determinant to enhance the recall of the system. The results obtained for each of the analysed semantic categories indicate the need for several improvements, as the the creation of more knowledge sources in Portuguese.

Example Applications

The previous chapters have introduced the medical information extraction system MEDINX and the ontologies developed in the context of the PDLs processing. They have also presented the methods to automatically populate these ontologies and the evaluation results of the system. This chapter will give brief overviews of three example projects developed to use MEDINX capabilities in real applications. These projects were designed to answer common clinical problems related with the automatic coding of the diagnoses and other health-related conditions described in the PDLs, and also to automatically review the content and completeness of the documents against specific criteria. The presented example applications were designed to support automatic code assignment and automatic clinical reports auditing.

7.1 Automatic Code Assignment support

Currently, and particularly in Portuguese hospitals, a great deal of human time and effort is expended in assigning codes to diseases, and other health-related conditions, described in medical records. Often the standards used to code the documents are extremely complex and only expert annotators with extensive training can perform it. In order to assign codes to a medical record, the coder must select from thousands of codes the smallest set appropriate to a given situation. Since this coding determines mainly reimbursement, it is important to accomplish this task as accurately as possible.

Providing assistance to manual coding is an important research area in Medical Informatics since many decades. Chapter 2 introduced some of the systems developed to automatically code clinical text, with focus on the systems developed in the context of the Medical NLP challenge (Pestian *et al.*, 2007a). These systems were designed to code radiology reports using the ICD-9-CM classification system.

In MEDINX we used the knowledge contained in the ontologies of the international classification systems ICD-9-CM and ICF to support automatic code assignment to the PDLs. Consequently, the MEDINX tool for automatic code assignment is based on knowledge-intensive methods, i.e., methods that represent both the coding system and the clinical text written in natural language.

7.1.1 Automatic ICD-9-CM code assignment

The NER component of MEDINX, REMMIX (presented in section 5.4.2), uses the ICD-9-CM ontology in the identification and categorization of the entities of the type *Disease*. In case of a full match of the expression described in the PDL with the description of a given code in the ICD-9-CM ontology, REMMIX assigns the ICD-9-CM code to the annotations of the entity being analyzed. The code reference is saved in the new annotation attributes and can be used by a subsequent annotator in the remaining processing, as, for instance, by the consumers of the system.

This feature of MEDINX allows to automatically assign ICD-9-CM codes to the diseases described in the PDLs. In fact, the graphical interface developed to evaluate MEDINX has already been designed to deal with this information, as presented in Figure 7.1. The image presents the entity *Osteoartrose* which has been categorized by MEDINX as a *Disease* and associated with the ICD-9-CM code 715.9.

The screenshot shows a web-based interface for evaluating MEDINX. The main area displays a clinical text with several segments highlighted in grey boxes. The text is organized into three main sections: 'Evolução', 'Destino', and 'TA'. The 'Evolução' section contains a paragraph about a patient's condition and treatment. The 'Destino' section contains a paragraph about a transfer to a hospital. The 'TA' section contains a paragraph about a therapeutic adjustment. A yellow pop-up window titled 'Entity' is overlaid on the right side of the text. It contains the following information: 'Category : CONDITION', 'Type : DISEASE', 'ICD9 : 715.9', and 'Evaluation : Correct (selected), Incorrect, Don't know'. There are also buttons for 'Cancel', 'Save', and 'Enter a comment...'.

Figure 7.1: Automatic ICD-9-CM code assignment with MEDINX.

7.1.2 Automatic ICF code assignment

WHO describes ICF as the classification of health and health-related domains (World Health Organization, 2010b). The ICF domains are classified from both, body individual,

and societal perspectives. This classification can be used to describe how functional problems can result in difficulties carrying out tasks, and how these problems are manifested in a person's environment (Section 4.3). In fact, ICD-9-CM codes give little detail about the effect of a patient's health status on the individual's daily activities. They do not reflect, for instance, that two individuals may have the same medical diagnosis, but very different levels of functioning.

The main objective of MEDINX ICF code assignment application is the automatic extraction of information from the PDLs related with ICF components, namely the body functions and structures, the activities and participation, and the environmental factors. These last, in particular, are generally less explored in code assignment tasks (manual or automatic) and present relevant information concerning the physical, social and environmental conditions in which people live. The application aggregates all the relevant information in an automatically ICF coded PDL summary. To this purpose a new annotator, the *ICF annotator*, has been developed and integrated in REMMIX pipeline. Figure 7.2 reflects the integration of the ICF annotator in REMMIX.

Figure 7.2: REMMIX pipeline with the ICF annotator.

This annotator has been placed at the end of REMMIX pipeline in order to be able to access all the annotations added to the CAS by the previous annotators. In particular, the ICF annotator reads the content of the MEDINX ICF ontology regarding the codes definition, and identifies the annotations that reflect this description. In case of a match, the entity annotation is updated with the correspondent ICF code.

To create a summary of the extracted information, the MEDINX *Template filling* consumer

was used. In particular, this consumer wraps up the information concerning the ICF code, the code description, the expression used in the PDLs to describe the entity, and the segment where the entity has been extracted from in the MEDINX ICF summary. This summary can also integrate the information concerning the qualifiers of the extracted entity, as, the related *modifiers*, *negations* and/or *lateralization* entities, and can also include the temporal information of the entity, i.e., the related *Time* entities.

Impairments of Body Structures				
List of Body Structures	PDL text	Qualifiers	Time	Segment
s320 structure of mouth	desvio da comissura labial [direita]	ligeiro	entrada	Exame Físico
s1 structures of the nervous system	edemas [periféricos]	sem	entrada	Exame Físico
s320 structure of mouth	desvio da comissura labial [direita]	ligeiro	data de alta	Evolução
s1 structures of the nervous system	edemas [periféricos]	sem	data de alta	Evolução

Impairments of Activities and Participation				
List of Activities and Participation	PDL text	Qualifiers	Time	Segment
d5 self-care	autonomo	---	---	História Clínica
d510 washing oneself	independente na sua higiene	---	---	História Clínica

Figure 7.3: Automatically created ICF summary of the PDL presented in Figure 7.4.

An example of the output of this application, on a PDL retrieved from the MEDINX development set, is given in Figure 7.3. The figure shows part of the automatically produced summary of the information described in the report of Figure 7.4. Namely, the summary lists the entities that belong to the ICF categories *body structures* and *activities and participation*. The full summary can be analysed in Appendix E, where all the entities automatically coded by the system are presented. These entities correspond to the underlined text in Figure 7.4.

7.2 Audit system

The previous chapters have already demonstrated that the clinical data included in the PDLs constitute a rich source of information about the patient medical conditions, procedures and treatments results. However, such documents are only helpful if they are complete and represent all relevant information concerning the inpatient period.

Currently, the analysis of the content and completeness of the PDLs is a process that needs to be performed manually, by expert physicians in the health institutions. The MEDINX clinical audit system main objective is to help, not only the physicians, but also the hospital administrators and managers in this task, and automatically analyse the content of the PDLs.

The clinical audit system uses the automatically populated MEDINX ontology, containing the structured information extracted from the PDLs and performs an automatic analysis of the content and completeness of the documents. In particular, the MEDINX audit system

<p>Motivo de Internamento A.V.C. isquêmico em fase isodensa.</p> <p>História Clínica Doente de 67 anos trazido ao <u>S.U.</u> por, subitamente, ter ficado com a <u>voz arrastada</u>. Sem outras queixas acompanhantes. <u>Autônomo e independente na sua higiene</u>.</p> <p>Exame Físico A <u>entrada</u> apresentava-se <u>consciente, orientado e pouco colaborante</u>. <u>Pupilas isocóricas e reativas</u>. <u>Disartrico</u>. <u>Ligeiro desvio da comissura labial direita</u>. Sem lateralização motora. <u>Corado e hidratado</u>. <u>Apiretico</u>. <u>Eupneico</u>. <u>AC - arritmica</u>. <u>AP - sem alterações</u>. <u>Sem edemas periféricos</u>.</p> <p>Terapêutica Efectuada <u>AAS</u>, <u>Enoxaparina</u>, <u>Ranitidina</u>, <u>Amiodarona</u>, <u>Furosemido</u>, <u>Nitratos transdermicos</u>, <u>Carvedilol</u>, <u>DNI</u>, <u>Digoxina</u>, <u>Enalapril</u>, <u>Amoxicilina+CA</u>.</p> <p>DIAGNOSTICOS: 1 - Acidente Vascular Cerebral Isquêmico. 2 - Fibrilhação Auricular Crônica. 3 - Hipertensão Arterial. 4 - Infecção Urinária por germen não identificado.</p> <p>Destino Alta para o <u>domicílio</u> mantendo medicação <u>excepto Ticlopidina</u> que é substituída pelo <u>Clopidogrel</u> 75 mg 1id. Vai medicado com <u>Amoxicilina /AC</u> 12/12 horas por mais 3 dias. <u>Pedida Consulta Externa de Terapia da fala</u>.</p> <p>Evolução Foi observado pelo <u>Terapeuta da Fala</u> que sugeriu re-avaliação em <u>Consulta Externa</u>. A <u>data de alta</u> apresenta-se <u>consciente, orientado e colaborante</u>. <u>Disartrico</u>. Mantem <u>ligeiro desvio da comissura labial</u>. Sem déficits de novo. <u>Corado e hidratado</u>. <u>Apiretico</u>. <u>Eupneico</u>. <u>Normotenso</u> (120/ 70 mmHg). ACP - sem alterações. Abdomen - livre. <u>Sem edemas periféricos</u>.</p>	<p>Reason for Admission Stroke at isodense phase.</p> <p>Medical History Patient of 67 years brought to E.R. after having sudden drool. No other complaints. Autonomous and independent in his hygiene.</p> <p>Physical examination At time of admission the patient was conscious, oriented and not so cooperative. Isochoric and reactive pupils. Dysarthria. Slight deviation of the right lip. Without motor lateralization. Flushed and hydrated. Afebrile. Eupneic. CA - arrhythmia. PA - unchanged. No peripheral edema.</p> <p>Therapy Performed AAS, Enoxaparin, ranitidine, amiodarone, furosemide, transdermal nitrates, Carvedilol, DNI, Digoxin, Enalapril, Amoxicillin + CA.</p> <p>DIAGNOSES: 1 - ischemic stroke. 2 - Chronic Atrial Fibrillation. 3 - Hypertension. 4 - Urinary Tract Infection by unidentified germ.</p> <p>Destination Discharge to home keeping medication except Ticlopidine substituted by Clopidogrel 75 mg 1id. Goes medicated with Amoxicillin/AC 12/12 hours for 3 days more. Ordered to Outpatient Speech Therapy.</p> <p>Evolution Observed by the Speech Therapist who suggested re-evaluation in Outpatient. The date of discharge the patient was conscious, oriented and cooperative. Dysarthria. Maintains slight deviation of the lip. No new deficits. Flushed and hydrated. Afebrile. Eupneic. Normotensive (120/70 mmHg). CPA - without amendment. Abdomen - free. No peripheral edema.</p>
---	---

Figure 7.4: Style-preserving illustration of the Patient Discharge Letter used to create the ICF Summary presented in Figure 7.3. The underlined text corresponds to the ICF entities automatically coded by the system.

uses the RDF query language SPARQL (introduced in section 2.3.1) to query the MEDINX ontology. Several queries were design to illustrate the different kinds of information that can be retrieved with such application.

The first group of queries was developed in order to obtain the five most commonly mentioned medical conditions, procedures and drugs (active ingredient or medication). The results of this query are listed in Table 7.1.

The table indicates, for instance, that the most common disease described in the documents is, as expected, *Hypertension*. However, the information retrieved allows also to understand that most of the patients admitted with hypertension related diagnoses, also have other important chronic conditions, as heart diseases or diabetes. Such information, has not only clinical relevance, as it allows to obtain specific information about the medical condition of the inpatients, but it is also important in administrative and statistical management, as it allows fast access to information concerning, for instance, the most frequent procedures

Table 7.1: Five most common Diseases, Procedures and Drugs mentioned in the PDLs.

Rank	Diseases	Procedures	Drugs
1	HTA (<i>Hypertension</i>)	TA (<i>Blood Pressure determination</i>)	AAS (<i>Acetylsalicylic Acid</i>)
2	AVC isquémico (<i>Ischemic Stroke</i>)	AP (<i>Pulmonary Auscultation</i>)	enalapril
3	AVC (<i>Stroke</i>)	AC (<i>Cardiac Auscultation</i>)	captopril
4	DM (<i>Diabetes mellitus</i>)	FC (<i>Heart Rate</i>)	enoxaparina (<i>enoxaparin</i>)
5	sopros (<i>Murmur (heart)</i>)	TAC-CE (<i>X-ray Computed Tomography</i>)	sinvastatina (<i>simvastatin</i>)

performed in the hospital, or the most commonly administered drugs.

Another group of queries has been designed to use the information obtained with the evaluation of the system. More specifically, the results of the evaluation of MEDINX in the task of IE from the PDLs, allowed to obtain several statistics related with the number of entities from each category identified in the texts. These values (presented in Section 6.3.4 and in more detail in Appendix D) indicate, for instance, that in all analysed PDLs, the MEDINX test set, were identified at least two medical conditions in each document, and that 75% of the analysed documents contained reference to, at least, 17 medical conditions.

Another example is given by the analyses of automatically identified chemical procedures (medication and active substances). The results allow to understand that in average a PDL contains 15 references to this entities, and only 25% of the PDLs refer more than 17 medications and active substances.

Using this information, a new rule can be idealized to obtain a list of reports describing less than 17 *Conditions*. The information retrieved is presented in Figure 7.5 and has several use cases, as, for instance, to suggest to the physician a review of the document produced, since it could be incomplete or not contain all the information desired in such documents.

Figure 7.5 also presents the interface developed to create the queries, and the SPARQL code used to implement the previously described rule.

Another idealized rule is presented in Figure 7.6 and describes a more complex scenario. This rule finds the episodes containing references to less than 17 *Conditions*, but also containing more than 17 *Chemical Procedures*, particularly, *Medications* and *Active Substances*.

The outcome of this rule is presented in Figure 7.7, where the *Episodes* matching the developed query are listed. The number of identified *Conditions* and *Chemical Procedures* in

Rule

Name: Episodes with less than 17 conditions

Severity: Info ☐ Warning ☐ Error ☒

Query:

```
SELECT ?episode (count(?condition) as ?count)
WHERE {
  ?episode rdf:type m:Episode.
  ?episode m:Episode_describes_Condition ?condition
}
GROUP BY ?episode ?count
HAVING ($count < 17)
```

Edit Run

Message: The episode {\$episode} has only {\$count} conditions.

Figure 7.5: Audit System Rule Editor.

Rule

Name: Episodes with less than 17 conditions and more than 17 medications and active substances

Severity: Info ☐ Warning ☐ Error ☒

Query:

```
SELECT ?episode (count(?condition) as ?conditionCount) (count(?activeSubstance ) as
?asCount) (count(?medication ) as ?medCount)
WHERE {
  {
    ?episode rdf:type m:Episode.
    ?episode m:Episode_describes_Condition ?condition.
  }
  UNION
  {
    ?episode rdf:type m:Episode.
    ?episode m:Episode_describes_Procedure ?activeSubstance.
    ?activeSubstance rdf:type m:ActiveSubstance
  }
  UNION
  {
    ?episode rdf:type m:Episode.
    ?episode m:Episode_describes_Procedure ?medication.
    ?medication rdf:type m:Medication
  }
}
GROUP BY ?episode ?conditionCount ?medCount
HAVING ($conditionCount < 17 && ( ?medCount + $asCount) > 17)
```

Edit Run

Message: Episode {\$episode} has {\$conditionCount} Conditions and {\$asCount} active substances and {\$medCount} medications

Figure 7.6: SPARQL query *Episodes with less than 17 conditions and more than 17 medications*.

this PDLs present an absolute deviation from the trigger values used in the query and were, therefore, identified. These values could indicate an inconsistent report and a possible need for revision.

Rule
 Name : Episodes with less than 17 conditions and more than 17 medications and active substances
 Severity : Info ☐ Warning ☐ Error ☒

Severity	Message
Error	Episode Episode_5007662 has 14 Conditions and 11 active substances and 8 medications
Error	Episode Episode_8000356 has 14 Conditions and 18 active substances and 14 medications
Error	Episode Episode_5011622 has 15 Conditions and 10 active substances and 8 medications

Query :

[Edit](#) [Run](#)

Message : Episode {Sepisode} has {SconditionCount} Conditions and {SasCount} active substances and {SmedCount} medications

Figure 7.7: Audit System Results interface.

7.3 Summary

This chapter presented some example applications that make use of MEDINX capabilities. These applications use the system mechanisms to read, process and utilize external resources to solve specific clinical problems. In particular, the chapter presented some domains of application of MEDINX ontologies and examples of possible reuse of these conceptualizations. The MEDINX support to automatic code assignment and the audit system presented in the previous paragraphs, use the knowledge contained in the ontologies created within this work to help clinicians in difficult and time consuming tasks that can only be performed by experts.

By presenting these simple applications of the system we present examples of the possible real applications and how NLP technologies can be connected to develop more comprehensive solutions that establish themselves in the hospital setting.

Conclusions

Health information systems are becoming more frequent worldwide and are expected to lower costs and improve health care quality through improved access to health information. However, free unstructured text is still a common source of information in this type of systems, preventing the information included in clinical narrative text to be used in several relevant medical applications, such as clinical decision support systems. There is, thus, a growing need for systems that automatically extract information from narrative data as manual extraction of this information is a costly, time consuming and error prone process.

MEDINX, the IE system presented in this document, is the first IE system for Portuguese hospital discharge letters. MEDINX uses an ontology-driven approach to the task of IE from the clinical narratives. By using ontologies, the system converts a complete report into a relational information model and increases the expressive power of the extraction rules. The precision and recall of MedInX were determined in the evaluation of the system against a set of 50 discharge documents validated by 7 different reviewers. The systems obtained an F-measure of 95% in the task of identifying and classifying relevant clinical entities. These results are particularly relevant in a system intended to be used in the clinical area, a domain where the impact of errors and missing data has particular relevance, as it can influence medical decisions and, consequently, the patients' health.

MEDINX has been designed to be modular system and allow easy introduction of new components. In addition, the hypertension related components can be replaced by other domain related modules to allow the adaptation of MedInX to an ever changing and evolving number of knowledge representation models within and outside the medical space.

8.1 Future work

There are numerous opportunities for extending and improving this work. For example, new sources of knowledge could be created; a complete gold standard of the clinical records could be developed (following the started approach) opening possibilities for the development and application of new models for information extraction; new applications, directed to real clinical problems and using new reasoning methods can be found and implemented; new ways of combining existing knowledge could be explored and new applications for the ontologies can appear. Another promising potential of language technology is to interactively improve, interpret and code during text entry so that the resulting structured, coded, free text can be validated immediately. These challenges can also be addressed in co-operation with medical experts to increase and ensure the clinical applicability of the system.

Free text is handy for information entry but a challenge for information extraction. Language technology can alleviate some of these problems in *retrospective* analysis by offering a more semantically informed interpretation and abstraction. The critical bottleneck today is namely information handover and reuse, and extensive text is simply not used nor is useful. Interactively validated, semantically processed text could be more usable and support abstraction, visualization. The development of query tools that could be used during consultation and live practice would greatly benefit clinicians and patients.

In a more general view, HLT can improve intelligibility of health information and in this way, support clinical decision making and empower citizens cognitively. Currently, the records are produced and used by health care professionals. But the individuals, as the patients or the citizens in general, could also benefit with the use of such technologies to access their health data, and in the future, even enter data to their records.

HLT tools have also the potential to enhance the quality and efficiency of care by improved access of health information. This tools give the health professionals more time for direct care. This has positive impacts on patient safety as well as efficiency and profitability of health care services. The technology can allow massive analyses of textual patient records, and use of this knowledge-rich resource to enrich clinical evidence base. This enables better and more versatile usage of knowledge gained with previous patients when caring for new patients, supporting, this way, individualized care. Text mining capabilities can also be used by creating practice-based evidence.

In general, the ability to base care decisions on accurate, timely and customized views on relevant information is likely to cause a paradigm shift in decision making and administration in health care.

Bibliography

- J. ABERDEEN, J. BURGER, D. DAY, L. HIRSCHMAN, P. ROBINSON, AND M. VILAIN (1995). MITRE: description of the Alembic system used for MUC-6. In *Proceedings of the 6th conference on Message Understanding*, MUC6 '95, pages 141–155. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 1-55860-402-2.
- R. AGRAWAL AND R. SRIKANT (1994). Fast algorithms for Mining Association Rules. In *20th International Conference Very Large Data Bases*, volume 1215, pages 487–499.
- H. ALLVIN, E. CARLSSON, H. DALIANIS, R. DANIELSSON-OJALA, V. DAUDARAVICIUS, M. HASSEL, D. KOKKINAKIS, H. LUNDGREN-LAINE, G. NILSSON, O. NYTRØ, S. SALANTERA, M. SKEPPSTEDT, H. SUOMINEN, AND S. VELUPILLAI (2010). *Characteristics and Analysis of Finnish and Swedish Clinical Intensive Care Nursing Narratives*, pages 53–60. Second Louhi Workshop on Text and Data Mining of Health Documents (Louhi-10).
- J. ALMEIDA AND U. PINTO (1995). Jspell – um módulo para análise léxica genérica de linguagem natural. In *Actas do X Encontro da Associação Portuguesa de Linguística*, pages 1–15. Évora 1994.
- C. AMARAL, H. FIGUEIRA, A. MENDES, P. MENDES, C. PINTO, AND T. VEIGA (2008). Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM. In C. MOTA AND D. SANTOS (editors), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 171–179. Linguatca, Aveiro, Portugal.
- S. ANANIADOU AND J. MCNAUGHT (2005). *Text Mining for Biology And Biomedicine*. Artech House, Inc., Norwood, MA, USA. ISBN 158053984X.
- C. AONE, L. HALVERSON, T. HAMPTON, AND M. RAMOS-SANTACRUZ (1998). SRA:

- description of the IE2 system used for MUC-7. In *Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufmann Publishers, San Francisco, California.
- APACHE SOFTWARE FOUNDATION (2011). Apache software license. Last accessed 01/03/2011. URL <http://www.apache.org/licenses/>
- D. E. APPELT AND D. J. ISRAEL (1999). Introduction to Information Extraction Technology. In *A Tutorial prepared for the 16th International Joint Conference on Artificial Intelligence (IJCAI)*. Stockholm, Sweden.
- A. ARONSON, O. BODENREIDER, D. DEMMER-FUSHMAN, K. FUNG, V. LEE, AND J. MORK (2007). From indexing the biomedical literature to coding clinical text. In *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, pages 105–112. Prague, Czech Republic.
- S. AUER, C. BIZER, G. KOBILAROV, J. LEHMANN, R. CYGANIAK, AND Z. IVES (2008). DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, pages 722–735.
- F. BAADER, D. CALVANESE, D. L. MCGUINNESS, D. NARDI, AND P. F. PATEL-SCHNEIDER (editors) (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press. ISBN 0-521-78176-0.
- F. BARRETO, A. BRANCO, E. FERREIRA, A. MENDES, M. FERN, A. B. D. NASCIMENTO, F. NUNES, AND J. R. SILVA (2006). Open resources and tools for the shallow processing of portuguese: The tagshare project. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.
- V. BASHYAM, G. DIVITA, D. B. BENNETT, A. C. BROWNE, AND R. K. TAIRA (2007). A normalized lexical lookup approach to identifying UMLS concepts in free text. *Stud Health Technol Inform*, 129(Pt 1):545–549.
- J. J. BERMAN (2002). Confidentiality issues for medical data miners. *Artif Intell Med*, 26(1-2):25–36.
- A. BERNARAS, I. LARESGOITI, AND J. CORERA (1996). Building and reusing ontologies for electrical network applications. In W. WAHLSTER (editor), *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI 96): Budapest, Hungary: August 11-16*, pages 298–302. Wiley. ISBN 0471968099.
- T. BERNERS-LEE, R. CAILLIAU, A. LUOTONEN, H. F. NIELSEN, AND A. SECRET (1994). The world-wide web. *Communications of the ACM*, 37:76–82.

- T. BERNERS-LEE AND M. FISCHETTI (1999). *Weaving the Web; The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper Audio. ISBN 0694521256.
- T. BERNERS-LEE, J. HENDLER, AND O. LASSILA (2001). The semantic web. *Scientific American*, pages 28–37.
- O. BODENREIDER AND A. T. MCCRAY (2003). Exploring semantic groups through visual approaches. *J Biomed Inform*, 36(6):414–432.
- L. BORREGO AND P. QUARESMA (2010). Criação de uma Ontologia e Respectiva Povoação a Partir do Processamento de Relatórios Médicos. In *Jornadas de Informática da Universidade de Évora*, pages 1–8. Informatics Department, University of Evora.
- W. N. BORST (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Enschede, The Netherlands, Doctoral Thesis of the University of Twente.
- M. M. BOUAMRANE, A. RECTOR, AND M. HURRELL (2008). Gathering precise patient medical history with an ontology-driven adaptive questionnaire. In *Proc. 21st IEEE International Symposium on Computer-Based Medical Systems CBMS '08*, pages 539–541.
- D. BRICKLEY AND R. GUHA (2004). RDF Vocabulary Description Language 1.0: RDF Schema. Last accessed 01/12/2010.
URL <http://www.w3.org/TR/rdf-schema/>
- E. BRILL (1992). A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155. Trento, Itália.
- P. BUITELAAR AND M. SIEGEL (2006). Ontology-based Information Extraction with SOBA. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 2321–2324. Genoa, Italy.
- R. BUNESCU AND M. PASCA (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*.
- M. E. CALIFF AND R. J. MOONEY (1998). Relational Learning of Pattern-Match Rules for Information Extraction. In *Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 6–11. AAAI Press, Menlo Park, CA.
- D. A. CAMPBELL AND S. B. JOHNSON (2001). Comparing syntactic complexity in medical and non-medical corpora. *Proc AMIA Symp*, pages 90–94.
- H. CAO, G. HRIPCSAK, AND M. MARKATOU (2007). A statistical methodology for analyzing co-occurrence data from a large sample. *J Biomed Inform*, 40(3):343–352.

- H. CAO, M. MARKATOU, G. MELTON, M. CHIANG, AND G. HRIPCSAK (2005). Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. *AMIA Annu Symp Proc*, pages 106–110.
- N. CARDOSO (2008). REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In C. MOTA AND D. SANTOS (editors), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 171–179. Linguateca, Aveiro, Portugal.
- CARNIGIE MELLON UNIVERSITY (2010). UIMA Component Repository. Last accessed 01/05/2010.
URL <http://uima.lti.cs.cmu.edu>
- P. CARVALHO, H. G. OLIVEIRA, D. SANTOS, C. FREITAS, AND C. MOTA (2008). Segundo harem: Modelo geral, novidades e avaliação. In C. MOTA AND D. SANTOS (editors), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 1, pages 11–31. Linguateca, Aveiro, Portugal.
- A. C. CASTILLA, S. S. FURUIE, AND E. A. MENDONÇA (2007). Multilingual information retrieval in thoracic radiology: feasibility study. *Stud Health Technol Inform*, 129(Pt 1):387–391.
- W. W. CHAPMAN, W. BRIDEWELL, P. HANBURY, G. F. COOPER, AND B. G. BUCHANAN (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34(5):301–310.
- W. W. CHAPMAN AND J. N. DOWLING (2006). Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. *J Biomed Inform*, 39(2):196–208.
- W. W. CHAPMAN, J. N. DOWLING, AND G. HRIPCSAK (2008). Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform*, 77(2):107–113.
- E. S. CHEN, G. HRIPCSAK, H. XU, M. MARKATOU, AND C. FRIEDMAN (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc*, 15(1):87–98.
- L. M. CHRISTENSEN, P. J. HAUG, AND M. FISZMAN (2002). MPLUS: a probabilistic medical language understanding system. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 29–36. Association for Computational Linguistics, Morristown, NJ, USA.

- J. CHUNG AND S. MURPHY (2005). Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports. *AMIA Annu Symp Proc*, pages 131–135.
- K. J. CIOS AND G. W. MOORE (2002). Uniqueness of medical data mining. *Artif Intell Med*, 26(1-2):1–24.
- F. CIRAVEGNA (2001). Adaptive information extraction from text by rule induction and generalization. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence(IJCAI)*.
- F. CIRAVEGNA, A. DINGLI, J. IRIA, AND Y. WILKS (2003). Multi-strategy definition of annotation services in melita. In H. CUNNINGHAM, Y. DING, AND A. KIRYAKOV (editors), *Proceedings of Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference*. Sanibel Island, Florida.
- A. CODEN, G. SAVOVA, I. SOMINSKY, M. TANENBLATT, J. MASANZ, K. SCHULER, J. COOPER, W. GUAN, AND P. C. DE GROEN (2008). Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *Journal of Biomedical Informatics*, 42(5):937–949.
- A. R. CODEN, S. V. PAKHOMOV, R. K. ANDO, P. H. DUFFY, AND C. G. CHUTE (2005). Domain-specific language models and lexicons for tagging. *J Biomed Inform*, 38(6):422–430.
- A. M. COHEN AND W. R. HERSH (2005). A survey of current work in biomedical text mining. *Brief Bioinformatics*, 6(1):57–71.
- COMPUTATIONAL MEDICINE CENTER (2007). International challenge: Classifying clinical free text using natural language processing. Last accessed 01/12/2010.
URL <http://www.computationalmedicine.org/challenge/index.php>
- J. COWIE AND W. LEHNERT (1996). Information extraction. *Commun. ACM*, 39(1):80–91.
- K. CRAMMER, M. DREDZE, K. GANCHEV, P. P. TALUKDAR, AND S. CARROLL (2007). Automatic code assignment to medical text. In *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, pages 129–136. Association for Computational Linguistics, Morristown, NJ, USA.
- H. CRAMÉR (1999). *Mathematical methods of statistics*. Number 978-0691005478. Princeton University Press.

- J. P. S. CUNHA, I. CRUZ, I. OLIVEIRA, A. S. PEREIRA, C. T. COSTA, A. M. OLIVEIRA, AND A. PEREIRA (2006). The RTS project: Promoting secure and effective clinical telematic communication within the Aveiro region. In *Em eHealth 2006 High Level Conference*, pages 1–10.
- H. CUNNINGHAM (2000). *Software Architecture for Language Engineering*. PhD Thesis, University of Sheffield.
- H. CUNNINGHAM, D. MAYNARD, K. BONTCHEVA, AND V. TABLAN (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.
- H. CUNNINGHAM, D. MAYNARD, K. BONTCHEVA, V. TABLAN, AND C. URSU (2000). The GATE User Guide.
- D. G. DA SAÚDE (2006). Programa Nacional de Vacinação. Orientações Técnicas.
- G. DASILVA AND D. DWIGGEN (1980). Towards a prolog text grammar. *ACM SIGART Newsletter*, 73:20–25.
- R. DAVIS, H. E. SHROBE, AND P. SZOLOVITS (1993). What is a knowledge representation? *AI Magazine*, 14(1):17–33.
- D. DAY, J. ABERDEEN, L. HIRSCHMAN, R. KOZIEROK, P. ROBINSON, AND M. VILAIN (1997). Mixed-initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 348–355. Association for Computational Linguistics, Washington, DC, USA.
- L. R. S. DE LIMA, A. H. F. LAENDER, AND B. A. RIBEIRO-NETO (1998). A hierarchical approach to the automatic categorization of medical documents. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 132–139. ACM, New York, NY, USA. ISBN 1-58113-061-9.
- G. F. DEJONG (1982a). *An overview of the FRUMP system*, chapter 5, pages 149–176. Lawrence Erlbaum Associates, Hillsdale, NJ.
- G. F. DEJONG (1982b). *Strategies for Natural Language Processing*, chapter An Overview of the FRUMP System, pages 149–176. Lawrence Erlbaum and Associates, Hillsdale.
- J. DOMINGUE (1998). Tadzebao And Webonto: Discussing, Browsing, Editing Ontologies On The Web. In *11th Knowledge Acquisition for Knowledge-Based Systems Workshop*, pages 1–20. Banff, Canada.

- P. L. ELKIN, S. H. BROWN, B. A. BAUER, C. S. HUSSER, W. CARRUTH, L. R. BERGSTROM, AND D. L. WAHNER-ROEDLER (2005). A controlled trial of automated classification of negation from clinical notes. *BMC Med Inform Decis Mak*, 5:13.
- J. S. ELKINS, C. FRIEDMAN, B. BODEN-ALBALA, R. L. SACCO, AND G. HRIPCSAK (2000). Coding neuroradiology reports for the northern manhattan stroke study: a comparison of natural language processing and manual review. *Comput Biomed Res*, 33(1):1–10.
- D. W. EMBLEY, D. M. CAMPBELL, R. D. SMITH, AND S. W. LIDDLE (1998). Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the seventh international conference on Information and knowledge management*, CIKM '98, pages 52–59. ACM, New York, NY, USA. ISBN 1-58113-061-9.
- EUROPEAN COMMISSION (2008). Benchmarking ict use among general practitioners in europe. Technical report, Information Society and Media Directorate General, Bonn.
- EUROPEAN COMMISSION (2009). Semantic interoperability for better health and safer healthcare.
URL http://ec.europa.eu/information_society/
- EUROPEAN LANGUAGE RESOURCES ASSOCIATION (2010). Last accessed 01/05/2010.
URL <http://www.elra.info>
- G. EYSENBACH (2001). What is e-health? *Journal of Medical Internet Research*, 3(2):e20.
URL <http://www.jmir.org/2001/2/e20/>
- R. FARKAS AND G. SZARVAS (2008). Automatic construction of rule-based icd-9-cm coding systems. *BMC Bioinformatics*, 9 Suppl 3:S10.
- A. FARQUHAR, R. FIKES, AND J. RICE (1997). The ontolingua server: a tool for collaborative ontology construction. *International Journal of Human-Computer Studies*, 46(6):707–727.
- R. FELDMAN AND I. DAGAN (1995). Knowledge discovery in textual databases (kdt). In *1st International Conference on Knowledge Discovery*, pages 112–117. Montreal.
- L. FERREIRA, N. JAKOB, AND I. GUREVYCH (2008a). A comparative study of feature extraction algorithms in customer reviews. In *Proc. IEEE International Conference on Semantic Computing*, pages 144–151.
- L. FERREIRA, C. OLIVEIRA, A. TEIXEIRA, AND J. CUNHA (2009). Extraccao de informacao de relatórios médicos. *Linguamática*, 1(1):89–101.
URL <http://linguamatica.com/index.php/linguamatica/article/view/12/12>

- L. FERREIRA, A. TEIXEIRA, AND J. AO PAULO DA SILVA CUNHA (2010). Information extraction from portuguese hospital discharge letters. In C. G. MATEO, F. C. DÍAZ, AND F. M. PAZÓ (editors), *FALA 2010: VI Jornadas en Tecnologia del Habla, Speech and Language Technologies for Iberian Languages*, pages 39–42. Vigo, Spain. ISBN 978-84-8158-510-0.
- L. FERREIRA, A. TEIXEIRA, AND J. P. S. CUNHA (2008b). REMMA - Reconhecimento de Entidades Mencionadas do MedAlert. In C. MOTA AND D. SANTOS (editors), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguatca, Aveiro, Portugal.
- L. FERREIRA, A. TEIXEIRA, AND J. P. DA SILVA CUNHA (2006). Information extraction from medical reports. In *3rd International Workshop on Natural Language Understanding and Cognitive Science*.
- L. FERREIRA, A. TEIXEIRA, AND J. P. DA SILVA CUNHA (2008c). Ontology-driven vaccination information extraction. In *5th International Workshop on Natural Language Processing and Cognitive Science*.
- L. FERREIRA, A. TEIXEIRA, AND J. P. DA SILVA CUNHA (2008d). *Perspectivas sobre a Linguatca. Actas do encontro Linguatca: 10 anos*, chapter Linguatca e Processamento de Linguagem Natural na Área da Saúde: Alguns Comentários e Sugestões, pages 43–48. Linguatca. ISBN 978-989-20-1445-6.
URL <http://www.linguatca.pt/LivroL10/>
- D. FERRUCCI AND A. LALLY (2004). UIMA an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.
- J. FLEISS (2000). *Statistical Methods for Rates & Proportions*. John Wiley & Sons Inc. ISBN 9780471137276.
- K. FRANZÉN, G. ERIKSSON, F. OLSSON, L. ASKER, P. LIDÉN, AND J. CÖSTER (2002). Protein names and how to find them. *Int J Med Inform*, 67(1-3):49–61.
- D. FREITAG (1998). Toward general-purpose learning for information extraction. In *Proceedings of the 17th international conference on Computational linguistics*, pages 404–408. Association for Computational Linguistics.
- D. FREITAG AND A. MCCALLUM (2000). Information extraction with HMM structures learned by stochastic optimization. In *AAAI/IAAI*, pages 584–589.

- C. FREITAS, P. CARVALHO, H. G. OLIVEIRA, C. MOTA, AND D. SANTOS (2010). Second HAREM: Advancing the state of the art of Named Entity Recognition in Portuguese. In *The seventh international conference on Language Resources and Evaluation (LREC 2010)*. Malta.
- F. FREITAS, S. SCHULZ, AND E. MORAES (2009). Survey of current terminologies and ontologies in biology and medicine. *RECIIS. Electronic Journal of Communication, Information and Innovation in Health*, 3:1–13.
- F. FREITAS, H. STUCKENSCHMIDT, AND N. NOY (2006). Ontology issues and applications guest editors’s introduction. *Journal of the Brazilian Computer Society*, 11:5–16.
- J. FRIEDLIN AND C. J. McDONALD (2006). A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA Annu Symp Proc*, pages 269–273.
- C. FRIEDMAN (2000). A broad-coverage natural language processing system. *Proc AMIA Symp*, pages 270–274.
- C. FRIEDMAN, P. O. ALDERSON, J. H. AUSTIN, J. J. CIMINO, AND S. B. JOHNSON (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1(2):161–174.
- C. FRIEDMAN AND G. HRIPCSAK (1999). Natural language processing and its future in medicine. *Acad Med*, 74(8):890–895.
- C. FRIEDMAN, S. B. JOHNSON, B. FORMAN, AND J. STARREN (1995). Architectural requirements for a multipurpose natural language processor in the clinical environment. *Proc Annu Symp Comput Appl Med Care*, pages 347–351.
- C. FRIEDMAN, C. KNIRSCH, L. SHAGINA, AND G. HRIPCSAK (1999). Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *Proc AMIA Symp*, pages 256–260.
- P. GAMALLO (2011). Treetagger para o português. Last accessed 01/03/2011.
URL <http://gramatica.usc.es/gamallo/tagger.htm>
- J. H. GENNARI, M. A. MUSEN, R. W. FERGERSON, W. E. GROSSO, M. CRUBÓZY, H. ERIKSSON, N. F. NOY, AND S. W. TU (2002). The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. *International Journal of Human-Computer Studies*, 58:89–123.

- I. GOLDSTEIN, A. ARZRUMTSYAN, AND O. UZUNER (2007). Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. *AMIA Annu Symp Proc*, pages 279–283.
- A. GÓMEZ-PÉREZ, M. FERNÁNDEZ-LÓPEZ, AND O. CORCHO (2004). *Ontological engineering: with examples from the areas of knowledge management, e-commerce and the Semantic Web*. Advanced information and knowledge processing. Springer. ISBN 9781852335519.
- GOOGLE (2011a). Last accessed 01/03/2011.
URL <http://www.google.pt/>
- GOOGLE (2011b). Google health. Last accessed 01/03/2011.
URL <http://www.google.com/health/>
- S. GRIMM, P. HITZLER, AND A. ABECKER (2007). Knowledge representation and ontologies. *Knowledge Creation Diffusion Utilization*, 2642:535–540.
- T. R. GRUBER (1993a). Toward principles for the design of ontologies used for knowledge sharing. In *International Journal of Human-Computer Studies*, pages 907–928. Kluwer Academic Publishers.
- T. R. GRUBER (1993b). A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5:199–220.
- N. GUARINO (1998). Formal ontology and information systems. In *First International Conference on Formal Ontology in Information Systems*, pages 3–15. IOS Press, Trento, Italy.
- N. GUARINO AND P. GIARETTA (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32.
- I. GUREVYCH AND M.-C. MÜLLER (2008). Information Extraction with the Darmstadt Knowledge Processing Software Repository. In *Proceedings of the Workshop on Linguistic Processing Pipelines*.
- J. HABIB (2010). EhRs, meaningful use, and a model emr. *Drug Benefit Trends.*, 22(4):99–101.
- U. HAHN, E. BUYKO, R. LANDEFELD, M. MÜHLHAUSEN, M. POPRAT, K. TOMANEK, AND J. WERMTER (2008). An overview of jcore, the julie lab uima component repository. In U. HAHN, T. GÖTZ, E. W. BROWN, H. CUNNINGHAM, AND E. NYBERG (editors), *Proceedings of the LREC-2008 Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, 6th International Conference on Language Resources and Evaluation*, pages 1–7. ELRA, Marrakech, Morocco.

- U. HAHN, M. ROMACKER, AND S. SCHULZ (2002). Creating knowledge repositories from biomedical reports: The medsyndikate text mining system. In *Pac Symp Biocomput*, pages 338–349.
- U. HAHN AND J. WERMTER (2004). High-performance tagging on medical texts. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04. Association for Computational Linguistics, Morristown, NJ, USA.
- H. HARKEMA, A. SETZER, R. GAIZAUSKAS, AND M. HEPPLER (2005). Mining and modelling temporal clinical data. In *Processings of the UK e-Science All Hands Meeting*, pages 507–514.
- E. HATCHER AND O. GOSPODNETIC (2004). *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA. ISBN 1932394281.
- P. J. HAUG, S. KOEHLER, L. M. LAU, P. WANG, R. ROCHA, AND S. M. HUFF (1995). Experience with a mixed semantic/syntactic parser. *Proc Annu Symp Comput Appl Med Care*, pages 284–288.
- P. J. HAUG, D. L. RANUM, AND P. R. FREDERICK (1990). Computerized extraction of coded findings from free-text radiologic reports. work in progress. *Radiology*, 174(2):543–548.
- HEALTH INFORMATION MANAGEMENT SYSTEMS SOCIETY (2010). Cited 01/12/2010.
URL <http://www.himss.org/ASP/index.asp>
- M. A. HEARST (1999). Untangling text data mining. In *37th Annual Meeting of the Association for Computational Linguistics*. University of Maryland.
- T. HEINZE, M. LIGHT, AND F. SCHILDER (2008). Experiences with uima for online information extraction at thomson corporation. In U. HAHN, T. GÖTZ, E. W. BROWN, H. CUNNINGHAM, AND E. NYBERG (editors), *Proceedings of the LREC-2008 Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, 6th International Conference on Language Resources and Evaluation*, pages 8 – 11. ELRA, Marrakech, Morocco.
- M. HEPPLER (2000). Independence and commitment: Assumptions for rapid training and execution of rule-based pos taggers. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*. Hong Kong.
- W. R. HERSH, H. MÜLLER, J. R. JENSEN, J. YANG, P. N. GORMAN, AND P. RUCH (2006). Advancing biomedical image retrieval: development and analysis of a test collection. *J Am Med Inform Assoc*, 13(5):488–496.

- L. HIRSCHMAN (1998). The evolution of evaluation: lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12:281–305.
- J. R. HOBBS (1993). The generic information extraction system. In *MUC5 '93: Proceedings of the 5th conference on Message understanding*, pages 87–91. Association for Computational Linguistics, Morristown, NJ, USA. ISBN 1-55860-336-0.
- J. R. HOBBS (2002). Information extraction from biomedical text. *J Biomed Inform*, 35(4):260–264.
- G. HRIPCSAK AND D. F. HEITJAN (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35:99–110.
- G. HRIPCSAK, G. J. KUPERMAN, AND C. FRIEDMAN (1998). Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inf Med*, 37(1):1–7.
- G. HRIPCSAK AND A. S. ROTHSCILD (2005). Agreement, the f-measure, and reliability in information retrieval. *J. of Am. Medical Informatics Association*, 12(3):296–298.
- G. HRIPCSAK, L. ZHOU, S. PARSONS, A. K. DAS, AND S. B. JOHNSON (2005). Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. *J Am Med Inform Assoc*, 12(1):55–63.
- Y. HUANG AND H. J. LOWE (2007). A novel hybrid approach to automated negation detection in clinical radiology reports. *J Am Med Inform Assoc*, 14(3):304–311.
- S. B. HUFFMAN (1995). Learning information extraction patterns from examples. In *Learning for Natural Language Processing*, pages 246–260.
- i2B2 (2006). i2b2 (informatics for integrating biology and the bedside) website. Cited 01/12/2010.
URL <https://www.i2b2.org/>
- IBM (2011). Watson in the world. Last accessed 01/03/2011.
URL <http://www-03.ibm.com/innovation/us/watson/index.html>
- INFARMED (2011). Last accessed 01/03/2011.
URL www.infarmed.pt
- INTERNATIONAL HEALTH TERMINOLOGY STANDARDS DEVELOPMENT ORGANIZATION (2010). Last accessed 01/12/2010.

- P. JACKSON AND I. MOULINIER (2002). *Natural Language Processing for online applications; text retrieval, extraction and categorization*. John Benjamins Publ., Amsterdam, Philadelphia.
- N. L. JAIN AND C. FRIEDMAN (1997). Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc AMIA Annu Fall Symp*, pages 829–833.
- JENA (2010). Cited 01/12/2010.
URL <http://jena.sourceforge.net/index.html>
- A. JIMENO, E. JIMENEZ-RUIZ, V. LEE, S. GAUDAN, R. BERLANGA, AND D. REBHOLZ-SCHUHMANN (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC bioinformatics*, 9 Suppl 3(Suppl 3).
- A. K. JOSHI (1991). Natural language processing. *Science*, 253(5025):1242–1249.
- JULIE (2010). Julie lab. Cited 01/05/2010.
URL <http://www.julielab.de>
- D. JURAFSKY AND J. H. MARTIN (2000). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, Upper Saddle River, New Jersey.
- G. KASNECI, M. RAMANATH, F. SUCHANEK, AND G. WEIKUM (2008). The yago-naga approach to knowledge discovery. *SIGMOD Rec.*, 37(4):41–47.
- J. KAZAMA AND K. TORISAWA (2007). Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.
- J.-D. KIM, T. OHTA, Y. TATEISI, AND J. TSUJII (2003). Genia corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–i182.
- J.-D. KIM, T. OHTA, AND J. TSUJII (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.
- G. KLYNE AND J. J. CARROLL (2004). Resource description framework (rdf): Concepts and abstract syntax.
URL <http://www.w3.org/TR/rdf-concepts/>
- KNOWTATOR (2010). Cited 01/05/2010.
URL <http://bionlp.sourceforge.net/Knowtator>

- M. KRAUTHAMMER AND G. HRIPCSAK (2001). A knowledge model for the interpretation and visualization of nlp-parsed discharged summaries. *Proc AMIA Symp*, pages 339–343.
- R. KUKAFKA, M. E. BALES, A. BURKHARDT, AND C. FRIEDMAN (2006). Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health. *J Am Med Inform Assoc*, 13(5):508–515.
- A. KUMAR, P. CICCARESE, S. QUAGLINI, M. STEFANELLI, E. CAFFI, AND L. BOIOCCHI (2003). Relating umls semantic types and task-based ontology to computer-interpretable clinical practice guidelines. *Stud Health Technol Inform*, 95:469–474.
- M. KUNZE AND D. RÖSNER (2008). Uima for nlp based researchers’ workplaces in medical domains. In U. HAHN, T. GÖTZ, E. W. BROWN, H. CUNNINGHAM, AND E. NYBERG (editors), *Proceedings of the LREC-2008 Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, 6th International Conference on Language Resources and Evaluation*, pages 20–23. ELRA, Marrakech, Morocco.
- N. KUSHMERICK, D. S. WELD, AND R. B. DOORENBOS (1997). Wrapper induction for information extraction. In *Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 729–737.
- L. LARKEY AND W. B. CROFT (1995). Automatic assignment of icd9 codes to discharge summaries. Technical report.
- O. LASSILA AND R. SWICK (1998). Resource Description Framework (RDF) Model and Syntax. W3C, World Wide Web Consortium. [Http://www.w3.org/TR/WD-rdf-syntax/](http://www.w3.org/TR/WD-rdf-syntax/).
- C. LINDBERG (1990). The unified medical language system (umls) of the national library of medicine. *J Am Med Rec Assoc*, 61(5):40–42.
- D. LINDBERG, B. HUMPHREYS, AND A. MCCRAY (1993). Unified medical language systems. *Methods of Information in Medicine*, 32(4):281–291.
- LINGUATECA (2008). Last accessed 01/03/2011.
URL <http://www.linguateca.pt/HAREM/>
- LINGUISTIC DATA CONSORTIUM (2010). Cited 01/05/2010.
URL <http://www ldc.upenn.edu>
- H. LIU, Y. A. LUSSIER, AND C. FRIEDMAN (2001). Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method. *Journal of Biomedical Informatics*, 34(4):249–261.
URL <http://dx.doi.org/10.1006/jbin.2001.1023>

- K. LIU, W. CHAPMAN, R. HWA, AND R. S. CROWLEY (2007). Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *J Am Med Inform Assoc*, 14(5):641–650.
- K. LIU, K. J. MITCHELL, W. W. CHAPMAN, AND R. S. CROWLEY (2005). Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. *AMIA Annu Symp Proc*, pages 460–464.
- K. LLOYD, M. CELLA, M. TANENBLATT, AND A. CODEN (2009). Analysis of clinical uncertainties by health professionals and patients: an example from mental health. *BMC Med Inform Decis Mak*, 9:34.
- W. LONG (2005). Extracting diagnoses from discharge summaries. *AMIA Annu Symp Proc*, pages 470–474.
- C. LOVIS, R. H. BAUD, AND P. PLANCHE (2000). Power of expression in the electronic patient record: structured data or narrative text? *Int J Med Inform*, 58-59:101–110.
- A. MAEDCHE, E. MAEDCHE, G. NEUMANN, AND S. STAAB (2002). Bootstrapping an ontology-based information extraction system. In *Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web*. Springer.
- J. MAKHOUL, F. KUBALA, R. SCHWARTZ, AND R. WEISCHEDEL (1999). Performance measures for information extraction.
- M. MANDEL (2006). Integrated annotation of biomedical text: creating the pennbioie corpus. In *Text Mining, Ontologies and Natural Language Processing in Biomedicine*. Manchester.
- M. P. MARCUS, M. A. MARCINKIEWICZ, AND B. SANTORINI (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19:313–330.
- P. J. MASCHE (2004). *Multilingual Information Extraction*. Master Thesis, Dept. of Computer Science, Faculty of Science, University of Helsinki.
- D. MAYNARD, H. CUNNINGHAM, K. BONTCHEVA, R. CATIZONE, G. DEMETRIOU, R. GAIZAUSKAS, O. HAMZA, M. HEPPLER, P. HERRING, B. MITCHELL, M. OAKES, W. PETERS, A. SETZER, M. STEVENSON, V. TABLAN, C. URSU, AND Y. WILKS (2000). A survey of uses of gate. Technical Report CS-00-06, Department of Computer Science, University of Sheffield.
- A. T. MCCRAY, A. R. ARONSON, A. C. BROWNE, T. C. RINDFLESCH, A. RAZI, AND S. SRINIVASAN (1993). Umls knowledge for biomedical language processing. *Bull Med Libr Assoc*, 81(2):184–194.

- A. MCGRAY, J. SPONSLER, B. BRYLAWSKI, AND A. BROWNE (1987). The role of lexical knowledge in biomedical text understanding. In *SCAMC*, pages 103–107. IEEE.
- J. MCNAUGHT AND W. BLACK (2005). *Text Mining for Biology And Biomedicine*, chapter Information Extraction, pages 143–178. Artech House, Inc.
- MEDICAL SUBJECT HEADINGS (2010). Cited 01/12/2010.
URL <http://www.nlm.nih.gov/mesh/>
- S. MEYSTRE AND P. J. HAUG (2006). Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform*, 39(6):589–599.
- S. M. MEYSTRE AND P. J. HAUG (2005). Comparing natural language processing tools to extract medical problems from narrative text. *AMIA Annu Symp Proc*, pages 525–529.
- S. M. MEYSTRE, G. K. SAVOVA, K. C. KIPPER-SCHULER, AND J. F. HURDLE (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, pages 128–144.
- MICROSOFT (2007). Health vault. Last accessed 01/03/2011.
URL <http://www.healthvault.com>
- A. MIKHEEV, C. GROVER, AND M. MOENS (1998). Description of the LTG system used for MUC-7. In *Proceedings of 7th Message Understanding Conference (MUC-7)*.
- A. MIKHEEV, M. MOENS, AND C. GROVER (1999). Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 1–8. Association for Computational Linguistics, Stroudsburg, PA, USA.
- S. MILLER, M. CRYSTAL, H. FOX, L. RAMSHAW, R. SCHWARTZ, R. STONE, R. WEISCHEDEL, AND A. GROUP (1998). Algorithms that learn to extract information; bbn: Description of the sift system as used for muc-7. In *Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufmann Publishers, San Francisco, California.
- MINISTÉRIO DA SAÚDE AND ADMINISTRAÇÃO CENTRAL DO SISTEMA DE SAÚDE (2009). RSE - Registo de Saúde Electrónico: Documento de Estado da Arte. Technical report.
- K. J. MITCHELL, M. J. BECICH, J. J. BERMAN, W. W. CHAPMAN, J. GILBERTSON, D. GUPTA, J. HARRISON, E. LEGOWSKI, AND R. S. CROWLEY (2004). Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. *Stud Health Technol Inform*, 107(Pt 1):663–667.

- R. MITKOV (2003). *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.)*. Oxford University Press. ISBN 0198238827.
- MITRE (2010). Callisto. Last accessed 01/05/2010.
URL <http://callisto.mitre.org>
- MMAX2 (2010). Last accessed 01/05/2010.
URL <http://mmax.eml-research.de>
- C. MOTA (2008). R3m, uma participação minimalista no segundo harem. In C. MOTA AND D. SANTOS (editors), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 181–193. Linguatca, Aveiro, Portugal.
- C. MOTA AND D. SANTOS (editors) (2008a). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguatca.
- C. MOTA AND D. SANTOS (editors) (2008b). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguatca. ISBN: 978-989-20-1656-6.
URL <http://www.linguatca.pt/LivroSegundoHAREM/>
- I. MUSLEA, S. MINTON, AND C. A. KNOBLOCK (2001). Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*, 4(1-2):93–114.
- P. G. MUTALIK, A. DESHPANDE, AND P. M. NADKARNI (2001). Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the umls. *J Am Med Inform Assoc*, 8(6):598–609.
- C. MÜLLER, T. ZESCH, M.-C. MÜLLER, D. BERNHARD, K. IGNATOVA, I. GUREVYCH, AND M. MÜHLHÄUSER (2008). Flexible uima components for information retrieval research. In U. HAHN, T. GÖTZ, E. W. BROWN, H. CUNNINGHAM, AND E. NYBERG (editors), *Proceedings of the LREC-2008 Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, 6th International Conference on Language Resources and Evaluation*, pages 24–27. ELRA, Marrakech, Morocco.
- H. MÜLLER, T. DESELAERS, T. DESERNO, E. KIM, AND W. HERSH (2006). Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In *In: CLEF 2006 Proceedings. Lecture Notes in Computer Science (2007)*, pages 595–608.
- NANCY CHINCHOR (1998). Overview of MUC-7. In *SemaNet '02: Building and Using Semantic Networks*, pages 38–44.

- NATIONAL CANCER INSTITUTE (2010). cancer text information extraction system (caties). Last accessed 01/12/2010.
- NATIONAL INSTITUTE OF STATISTICS (2009). Last accessed 01/01/2011.
URL <http://www.ine.pt/>
- NATURAL LANGUAGE SOFTWARE REGISTRY (2010). Last Accessed 01/05/2010.
URL <http://registry.dfki.de>
- R. NECHES, R. FIKES, T. FININ, T. GRUBER, R. PATIL, T. SENATOR, AND W. R. SWARTOUT (1991). Enabling technology for knowledge sharing. *AI Mag.*, 12:36–56.
- S. J. NELSON, T. POWELL, AND B. L. HUMPHREYS (2006). The Unified Medical Language System (UMLS) Project. Last accessed 01/12/2010.
URL <http://www.nlm.nih.gov/mesh/umlsforelis.html>
- S. J. NELSON AND J.-L. SCHULMAN (2007). A multilingual vocabulary project - managing the maintenance environment. In *European Association for Health Information & Libraries (EAHIL) Workshop*. Krakow, Poland.
- N. F. NOY AND D. L. MCGUINNESS (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Online.
- C. NÉDELLEC (2005). Learning language in logic - genic interaction extraction challenge. In *Proceedings of the Learning Language in Logic 2005 Workshop at the International Conference on Machine Learning*.
- OBJECT MANAGEMENT GROUP (2011). Unified modeling language. Last accessed 01/03/2011.
URL <http://www.uml.org>
- OBO (2010). Obo-edit. an introduction to obo ontologies. Last accessed 01/12/2010.
URL http://oboedit.org/docs/html/An_Introduction_to_OBO%20ontologies.htm
- P. OGREN (2006a). knowtator: A plug-in for creating training and evaluation data sets for biomedical natural language systems. In *Proceedings of the 9th International Protégé Conference*, pages 73–76. Stanford, California.
- P. OGREN, G. SAVOVA, AND C. CHUTE (2008). Constructing evaluation corpora for automated clinical named entity recognition. In B. M. J. M. J. O. S. P. D. T. NICOLETTA CALZOLARI (CONFERENCE CHAIR), KHALID CHOUKRI (editor), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco. ISBN 2-9517408-4-0. <Http://www.lrec-conf.org/proceedings/lrec2008/>.

- P. V. OGREN (2006b). Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275. Association for Computational Linguistics, Morristown, NJ, USA.
- P. V. OGREN, G. SAVOVA, J. D. BUNTROCK, AND C. G. CHUTE (2006). Building and evaluating annotated corpora for medical nlp systems. *AMIA Annu Symp Proc*, page 1050.
- H. OH, C. RIZO, M. ENKIN, AND A. JADAD (2005). What is ehealth (3): A systematic review of published definitions. *Journal of Medical Internet Research*, 7(1):e1.
URL <http://www.jmir.org/2005/1/e1/>
- M. OLEJNIK, P. NOHAMA, P. S. CANCIAN, AND S. SCHULZ (2010). Performance analysis of a pos tagger applied to discharge summaries in portuguese. *Stud Health Technol Inform*, 160(Pt 2):959–63.
- I. C. OLIVEIRA AND J. P. S. CUNHA (2011). Integration services to enable regional shared electronic health records. In I. PRESS (editor), *XXIII International Conference of the European Federation for Medical Informatics (MIE 2011)*. Oslo, Norway. Accepted for publication.
- OPEN BIOMEDICAL ONTOLOGIES (2010). Last accessed 01/12/2010.
URL <http://www.obofoundry.org>
- S. PAKHOMOV, J. BUNTROCK, AND P. DUFFY (2005). High throughput modularized nlp system for clinical text. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, ACL '05, pages 25–28. Association for Computational Linguistics, Morristown, NJ, USA.
- S. V. PAKHOMOV, A. CODEN, AND C. G. CHUTE (2006). Developing a corpus of clinical notes manually annotated for part-of-speech. *Int J Med Inform*, 75(6):418–429.
- P. F. PATEL-SCHNEIDER, P. HAYES, AND I. HORROCKS (2004). Owl web ontology language semantics and abstract syntax. Last accessed 01/12/2010.
URL <http://www.w3.org/TR/owl-semantics/>
- J. PATRICK, Y. ZHANG, AND Y. WANG (2007). Developing feature types for classifying clinical notes. In *Biological, translational, and clinical language processing*, pages 191–192. Association for Computational Linguistics, Prague, Czech Republic.
- L. PESHKIN AND A. PFEFFER (2003). Bayesian information extraction network. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence(IJCAI)*.

- J. PESTIAN, C. BREW, P. MATYKIEWICZ, D. HOVERMALE, N. JOHNSON, K. B. COHEN, AND W. DUCH (2007a). A shared task involving multi-label classification of clinical free text. In ACL (editor), *Proceedings of ACL BioNLP*. Association of Computational Linguistics, Prague.
- J. P. PESTIAN, C. BREW, P. MATYKIEWICZ, D. J. HOVERMALE, N. JOHNSON, K. B. COHEN, AND W. DUCH (2007b). A shared task involving multi-label classification of clinical free text. In *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, pages 97–104. Association for Computational Linguistics, Morristown, NJ, USA.
- D. B. PETERSON (2005). International classification of functioning, disability and health: An introduction for rehabilitation psychologists. *Rehabilitation Psychology*, 50(2):105 – 112.
- E. PRUD'HOMMEAUX AND A. SEABORNE (2008). Sparql query language for rdf. Last accessed 01/03/2011.
URL <http://www.w3.org/TR/rdf-sparql-query/>
- S. PYYSALO, T. SALAKOSKI, S. AUBIN, AND A. NAZARENKO (2006). Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. *BMC Bioinformatics*, 7(Suppl 3).
- A.-M. RASSINOUX, S. E. FOR THE IMIA YEARBOOK SECTION ON KNOWLEDGE REPRESENTATION, AND MANAGEMENT (2009). Knowledge representation and management: towards interoperable medical terminologies. *Yearb Med Inform*, pages 99–102.
- S. RAY AND M. CRAVEN (2001). Representing sentence structure in hidden markov models for information extraction. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-2001)*.
- R. REIS AND J. J. D. DE ALMEIDA (1997). Etiquetador morfo-sintático para o português. In *Actas do XIII Encontro da Associação Portuguesa de Linguística*, volume 2, pages 209–222. Associação Portuguesa de Linguística, Lisboa, Portugal.
- E. RILOFF (1993). Automatically constructing a dictionary for information extraction tasks. In *National Conference on Artificial Intelligence*, pages 811–816.
- T. C. RINDFLESCH, S. V. PAKHOMOV, M. FISZMAN, H. KILICOGU, AND V. R. SANCHEZ (2005). Medical facts to support inferencing in natural language processing. *AMIA Annu Symp Proc*, pages 634–638.
- A. ROBERTS, R. GAIZAUSKAS, M. HEPPEL, N. DAVIS, G. DEMETRIOU, Y. GUO, J. KOLA, I. ROBERTS, A. SETZER, A. TAPURIA, AND B. WHEELDIN (2007). The clef corpus: Semantic annotation of clinical text. In J. M. TEICH, J. SUERMONDT, AND G. HRIPCSAK

- (editors), *American Medical Informatics Association 2007 Proceedings. Biomedical and Health Informatics: From Foundations to Applications to Policy*, pages 625–629. American Medical Informatics Association, Chicago, IL, USA.
- B. ROSARIO (2005). Multi-way relation classification: Application to protein-protein interactions. In *In Empirical Methods in Natural Language Processing (EMNLP)*.
- B. ROSARIO AND M. A. HEARST (2004). Classifying semantic relations in bioscience texts. In *In Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics*, pages 431–438.
- RTS REDE TELEMÁTICA DA SAUDE (2010). Last accessed 01/03/2010.
URL <http://www.rtsaude.pt/>
- D. L. RUBIN, N. H. SHAH, AND N. F. NOY (2008). Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, 9(1):75–90.
- M. RUIZ-CASADO, E. ALFONSECA, AND P. CASTELLS (2006). From wikipedia to semantic relationships: a semi-automated annotation approach. In *1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conference (ESWC 2006)*.
- J. RUMBAUGH, I. JACOBSON, AND G. BOOCH (1999). *The Unified Modeling Language Reference Manual*. Addison-Wesley Professional. ISBN 020130998X.
- J. C. SAGER (1994). Terminology: Custodian of knowledge and means of knowledge transfer. *Terminology*, 1:7–15(9).
- N. SAGER (1981). *Natural Language Information Processing: A Computer Grammar of English and its Applications*. Addison-Wesley Longman Publishing Co., Inc. ISBN 0201067692.
- N. SAGER, C. FRIEDMAN, AND E. CHI (1986). The analysis and processing of clinical narrative. In R. SALAMON, B. BLUM, AND M. JORGENSEN (editors), *Medinfo 86*, pages 1101–1105. Elsevier, Amsterdam.
- D. SANTOS AND N. CARDOSO (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguatca. ISBN: 978-989-20-0731-1.
URL <http://www.linguatca.pt/LivroHAREM/>
- D. SANTOS, N. CARDOSO, P. CARVALHO, I. DORNESCU, S. HARTRUMPF, J. LEVELING, AND Y. SKALBAN (2008a). Getting geographical answers from Wikipedia: the GIKIP pilot at CLEF. In *Working notes for the Cross Language Evaluation Forum, CLEF'2008*.

- D. SANTOS, C. FREITAS, H. G. OLIVEIRA, AND P. CARVALHO (2008b). Second harem: new challenges and old wisdom. In A. TEIXEIRA, V. L. S. DE LIMA, L. C. DE OLIVEIRA, AND P. QUARESMA (editors), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume 5190, pages 212–215. Springer Verlag, Aveiro, Portugal.
- D. SANTOS, N. SECO, N. CARDOSO, AND R. VILELA (2006). Harem: An advanced ner evaluation contest for portuguese. In N. CALZOLARI, K. CHOUKRI, A. GANGEMI, B. MAEGAARD, J. MARIANI, J. ODJIK, AND D. TAPIAS (editors), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 1986–1991. Genoa, Italy.
- G. SAVOVA, K. KIPPER-SCHULER, J. BUNTROCK, AND C. CHUTE (2008). Uima-based clinical information extraction system. In U. HAHN, T. GÖTZ, E. W. BROWN, H. CUNNINGHAM, AND E. NYBERG (editors), *Proceedings of the LREC-2008 Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, 6th International Conference on Language Resources and Evaluation*, pages 39–42. ELRA, Marrakech, Morocco.
- G. SCHADOW AND C. J. McDONALD (2003). Extracting structured information from free text pathology reports. *AMIA Annu Symp Proc*, pages 584–588.
- H. SCHMID (1995). TreeTagger, a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universidade de Estugarda*.
- S. SCHULZ, M. BOEKER, H. STENZHORN, AND J. NIGGEMANN (2009). Granularity issues in the alignment of upper ontologies. *Methods Inf Med*, 48(2):184–189.
- N. SECO (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, chapter MUC vs HAREM: a contrastive perspective, pages 35–41. Linguateca.
- T. SIBANDA, T. HE, P. SZOLOVITS, AND O. UZUNER (2006). Syntactically-informed semantic category recognition in discharge summaries. *AMIA Annu Symp Proc*, pages 714–718.
- B. SMITH, M. ASHBURNER, C. ROSSE, J. BARD, W. BUG, W. CEUSTERS, L. J. GOLDBERG, K. EILBECK, A. IRELAND, C. J. MUNGALL, N. LEONTIS, P. ROCCA-SERRA, A. RUTTENBERG, S.-A. SANSONE, R. H. SCHEUERMANN, N. SHAH, P. L. WHETZEL, AND S. LEWIS (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.
- R. SNYDER-HALPERN, S. CORCORAN-PERRY, AND S. NARAYAN (2001). Developing clinical practice environments supporting the knowledge work of nurses. *Comput Nurs*, 19(1):17–23; quiz 24–6.

- S. SODERLAND (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272.
- S. SODERLAND, D. FISHER, J. ASELTINE, AND W. LEHNERT (1995). CRYSTAL: Inducing a conceptual dictionary. In C. MELLISH (editor), *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1314–1319. Morgan Kaufmann, San Francisco.
- J. F. SOWA (2000). *Knowledge representation: logical, philosophical and computational foundations*. Brooks/Cole Publishing Co., Pacific Grove, CA, USA. ISBN 0-534-94965-7.
- K. SPACKMAN (2004). Snomed ct milestones: endorsements are added to already-impressive standards credentials. *Healthcare Informatics*, 21(9):54–56.
- V. N. STROETMANN, J.-P. THIERRY, K. A. STROETMANN, AND A. DOBREV (2007). ehealth for safety - impact of ict on patient safety and risk management. Technical report, European Commission Directorate-General Information Society and Media, ICT for Health.
- R. STUDER, V. R. BENJAMINS, AND D. FENSEL (1998). Knowledge engineering: principles and methods. *Data Knowl. Eng.*, 25:161–197.
- F. M. SUCHANEK, G. IFRIM, AND G. WEIKUM (2006). Combining linguistic and statistical analysis to extract relations from web documents. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 712–717. ACM, New York, NY, USA. ISBN 1-59593-339-5.
- F. M. SUCHANEK, G. KASNECI, AND G. WEIKUM (2007). Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, New York, NY, USA. ISBN 978-1-59593-654-7.
- F. M. SUCHANEK, M. SOZIO, AND G. WEIKUM (2009). Sofie: a self-organizing framework for information extraction. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 631–640. ACM, New York, NY, USA. ISBN 978-1-60558-487-4.
- H. SUOMINEN (2009). *Machine Learning and Clinical Text: Supporting Health Information Flow*. PhD Thesis, Turku Centre for Computer Science, Turku, Finland.
- D. R. SWANSON (1999). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*, 38(4):228–233.
- B. SWARTOUT, P. RAMESH, K. KNIGHT, AND T. RUSS (1997). Toward Distributed Use of Large-Scale Ontologies. In *AAAI Symposium on Ontological Engineering*, pages 138–148. Stanford University, California.

- V. TABLAN (2009). *Toward Portable Information Extraction*. PhD Thesis, The University of Sheffield.
- L. TANABE, N. XIE, L. H. THOM, W. MATTEN, AND W. J. WILBUR (2005). Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S3.
- A. THORODDSEN, K. SARANTO, A. EHRENBURG, AND W. SERMEUS (2009). Models, standards and structures of nursing documentation in european countries. *Stud Health Technol Inform*, 146:327–331.
- A. TORAL AND R. MUNOZ (2006). A proposal to automatically build and maintain gazetteers for named entity recognition using wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*.
- A. TURCHIN, N. S. KOLATKAR, R. W. GRANT, E. C. MAKHNI, M. L. PENDERGRASS, AND J. S. EINBINDER (2006). Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc*, 13(6):691–695.
- UBIQUITUOUS KNOWLEDGE PROCESSING LAB (2011). Last accessed 01/03/2011.
URL <http://www.ukp.tu-darmstadt.de/>
- A. UIMA (2011). Last accessed 01/03/2011.
URL <http://incubator.apache.org/uima>
- US DEPARTMENT OF HEALTH & HUMAN SERVICES (1996). Health insurance portability and accountability act of 1996. HIPAA 1996.
- M. USCHOLD, V. R. BENJAMINS, B. CH, A. GOMEZ-PEREZ, N. GUARINO, AND R. JASPER (1999). A framework for understanding and classifying ontology applications. In *Proceedings of the IJCAI99 Workshop on Ontologies*, pages 16–21.
- USNLM (2008). *UMLS Knowledge Sources*. United States National Library of Medicine.
- USNLM (2011). Pubmed. Last accessed 01/03/2011.
URL <http://www.ncbi.nlm.nih.gov/pubmed>
- O. UZUNER, I. GOLDSTEIN, Y. LUO, AND I. KOHANE (2008). Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, 15(1):14–24.
- O. UZUNER, Y. LUO, AND P. SZOLOVITS (2007). Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14(5):550–563.

- K. VERSPOOR, W. B. JR., C. ROEDER, AND L. HUNTER (2009). Abstracting the types away from a uima type system. In C. CHIARCOS, R. E. DE CASTILHO, AND M. STEDE (editors), *Von der Form zur Bedeutung: Text automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference*, pages 249–256. Tuebingen.
- J. VOSS (2005). Measuring Wikipedia. In *10th International Conference of the International Society for Scientometrics and Informatics*, pages 221–231.
- W3C (2010). Last accessed 01/12/2010.
URL <http://www.w3c.org>
- W3C OWL WORKING GROUP (2009). OWL 2 Web Ontology Language Document Overview. Last accessed 01/12/2010.
URL <http://www.w3.org/TR/owl2-overview/>
- S. H. WALSH (2004). The clinician’s perspective on electronic health records and how they can affect patient care. *BMJ*, 328(7449):1184–1187.
- WIKIPEDIA (2011). Last accessed 01/03/2011.
URL <http://www.wikipedia.org/>
- A. B. WILCOX, S. P. NARUS, AND W. A. BOWES (2002). Using natural language processing to analyze physician modifications to data entry templates. *Proc AMIA Symp*, pages 899–903.
- D. C. WIMALASURIYA AND D. DOU (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323.
- WORDFREAK (2010). Last accessed 01/05/2010.
URL <http://wordfreak.sourceforge.net>
- WORLD HEALTH ORGANIZATION (2010a). International classification of diseases. Last accessed 01/12/2010.
URL <http://www.who.int/classifications/icd/en/>
- WORLD HEALTH ORGANIZATION (2010b). International classification of functioning, disability and health (icf). Last accessed 01/12/2010.
URL <http://www.who.int/classifications/icf/en/>
- F. WU AND D. S. WELD (2007). Autonomously semantifying wikipedia. In *CIKM ’07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM, New York, NY, USA.

- B. YILDIZ (2007). *Ontology-Driven Information Extraction*. PhD Thesis, Vienna University of Technology.
- T. ZESCH, C. MÜLLER, AND I. GUREVYCH (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.
- L. ZHOU, C. FRIEDMAN, S. PARSONS, AND G. HRIPCSAK (2005). System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. *AMIA Annu Symp Proc*, pages 869–873.
- L. ZHOU, G. B. MELTON, S. PARSONS, AND G. HRIPCSAK (2006). A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform*, 39(4):424–439.
- L. ZHOU, S. PARSONS, AND G. HRIPCSAK (2008). The evaluation of a temporal reasoning system in processing clinical discharge summaries. *J Am Med Inform Assoc*, 15(1):99–106.
- P. ZWEIGENBAUM, D. DEMNER-FUSHMAN, H. YU, AND K. B. COHEN (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8(5):358–375.

Appendix **A**

Annotation Guidelines

MEDALERT - SISTEMA DE PROCESSAMENTO DE LINGUAGEM MÉDICA

Directivas de Anotação

Liliana Ferreira

Instituto de Engenharia Electrónica e Telemática de Aveiro
Universidade de Aveiro
lsferreira@ua.pt

Abril 2009

Resumo

O projecto MedAlert tem por objectivo o desenvolvimento de um sistema capaz de inferir de uma forma automática irregularidades/dúvidas suscitadas pelas decisões tomadas pelos profissionais de saúde. Uma parte importante da informação requerida por este tipo de sistemas é originada em formato de texto, embora, actualmente e em instituições *EHR-savvy*, em formato electrónico. O MedAlert usa extracção de informação para tornar esta informação não estruturada disponível, através da identificação de entidades semânticas e de relacionamentos entre estas. Para tal, possui actualmente um *corpus* de 48 229 textos médicos relativos a episódios de internamento no Hospital Infante D. Pedro, Aveiro. Define-se neste relatório o processo de selecção de um subconjunto deste *corpus* para anotação manual das entidades clínicas e seus relacionamentos, bem como a metodologia de anotação de modo a garantir a consistência da anotação e melhorar a concordância entre anotadores.

Conteúdo

1	Introdução	4
2	O <i>corpus</i>	7
3	Metodologia de Anotação	10
3.1	Regras gerais de anotação	11
3.1.1	Recursividade das anotações	11
3.1.2	Vagueza na classificação semântica	11
3.1.3	Escolha das entidades	12
3.1.4	Anotação dos relacionamentos entre entidades	12
3.1.5	Relacionamentos múltiplos de uma entidade	12
3.1.6	Equivalência entre relacionamentos	13
3.2	Entidades	13
3.3	Relacionamentos	17
3.3.1	Relacionamento de identidade	18
3.4	Esquema de anotação	19
3.5	Exemplos de anotação	20
3.5.1	Motivo de Internamento	20
3.5.2	História Clínica	21
3.5.3	Exame Físico	23
3.5.4	Evolução	25
3.5.5	Terapêutica Efectuada	27
4	Software de anotação	30
4.1	Instalação	31
4.2	Abrir os projectos de anotação	31
4.3	Anotação	34
4.3.1	Anotação de entidades	34
4.3.2	Anotação de relacionamentos	36
4.4	Criar XML com anotação	38
	Bibliografia	38

Lista de Figuras

1.1	Processo Iterativo de anotação de relatórios.	5
2.1	Estrutura do ficheiro disponibilizado.	7
2.2	Distribuição de tokens, frases e textos no <i>corpus</i>	9
3.1	Exemplo ilustrativo de anotação.	11
3.2	Esquema de anotação	19
4.1	Estrutura do directório knowtator.	30
4.2	Protégé GUI	32
4.3	Projecto de anotação no Knowtator.	33
4.4	Botões de navegação na colecção de texto.	34
4.5	Seleccionar os documentos a anotar	34
4.6	Anotação de entidades.	36
4.7	Anotação de relacionamentos.	37
4.8	Anotação de correferência.	37

Lista de Tabelas

1.1	Equivalência entre concordância e métricas de IE	6
2.1	Relatórios MedAlert	8
2.2	Distribuição de erros ortográficos	9
3.1	Relacionamentos MedAlert.	18

Capítulo 1

Introdução

Este documento descreve o *corpus* de relatórios médicos actual e define os passos necessários à construção de uma colecção dourada para o projecto MedAlert. Esta colecção irá permitir:

1. Focar e clarificar os requisitos;
2. Gerar um *gold standard* contra o qual serão calculados os resultados;
3. Gerar dados para o desenvolvimento do sistema: as regras de extracção podem ser criadas automaticamente ou manualmente; podem ser criados modelos estatísticos do texto através de algoritmos de *machine learning*.

A construção de uma colecção dourada para o projecto MedAlert pressupõe a existência de um *corpus* de documentos médicos anotados semanticamente, quer com múltiplas entidades, quer com as suas relações.

Uma vez que o processo de anotação faz uso de um *corpus* de grandes dimensões, com vários géneros de textos e envolve vários anotadores, surge a necessidade de definir uma metodologia de anotação. Esta deverá seguir os *standards* estabelecidos em processamento de linguagem natural, tais como:

- os anotadores deverão concordar com as directivas;
- os documentos deverão ser anotados por pelo menos dois anotadores;
- os documentos só serão usados se a taxa de concordância entre anotadores ultrapassar um determinado *threshold*
- as diferenças entre anotadores deverão ser resolvidas por um terceiro anotador.

De modo a garantir da qualidade da colecção é importante que todos os documentos sejam anotados pelo mesmo *standard* e que as várias questões

quer surjam ao anotar os relatórios sejam previamente resolvidas. Por exemplo, deverá *AVC isquémico protuberancial* ser anotado como *problema* ou como *problema* e *local anatómico*?

Estas directivas pretendem garantir a consistência, descrevendo em detalhe o que deve e o que não deve ser anotado:

- como decidir se duas entidades estão relacionadas,
- como lidar com correferência,

bem como outros casos especiais referidos ao longo do documento.

Estas deverão também apresentar uma sequência de passos, uma receita, que os anotadores deverão seguir quando trabalham com os documentos, de modo a minimizar os erros de omissão.

O desenvolvimento de directivas de anotação foi realizado através de um processo rigoroso e iterativo, criado de modo a garantir consistência. Este processo é apresentado na Figura 1.1.

Figura 1.1: Processo Iterativo de anotação de relatórios.

Em detalhe o processo teve:

1. Dupla anotação:

Um documento anotado por uma única pessoa pode reflectir vários problemas, como os valores ou erros frequentemente efectuados por um único anotador. A anotação dupla é uma forma comum de minimizar estes problemas, na qual cada documento é anotado independentemente por dois ou mais anotadores, e o conjunto de anotações comparado de modo a determinar a concordância.

2. Métricas de Concordância

A taxa de concordância entre anotadores é medida através do *inter annotator agreement* (IAA):

$$IAA = \frac{\text{concordância}}{\text{concordância} + \text{não concordância}} \quad (1.1)$$

Esta métrica é semelhante a outras usadas em avaliações de IE, tal como mostrado na Tabela 1.1

Tabela 1.1: Equivalência entre concordância e métricas de IE

Métricas de concordância	Métricas de avaliação de IE
Concordância	2 x correcto
Não-concordância	espúria + em falta
IAA	medida F

Os pares de anotações deverão ser rejeitados se a concordância não ultrapassar um determinado *threshold*. Na definição destas directivas foi alcançado um índice de concordância de **100%**.

As directivas de anotação são apresentadas neste documento. O capítulo 2 apresenta alguns dados sobre o *corpus* existente. A metodologia de anotação é apresentada no capítulo 3, onde são também apresentadas as entidades e os relacionamentos a serem anotados. O capítulo 4 apresenta o software de anotação a usar.

Capítulo 2

O corpus

O *corpus* MedAlert é relativo a episódios de internamento ocorridos (entre MM e MM de YYYY) no Hospital Infante D.Pedro em Aveiro e é actualmente constituído por 48 229 textos, relativos a:

- Motivo de internamento;
- História clínica;
- Exame físico;
- Evolução;
- Terapêutica;
- Destino.

O *corpus* foi disponibilizado num ficheiro XML com o esquema da figura 2.1.

Figura 2.1: Estrutura do ficheiro disponibilizado.

Posteriormente, foram criados seis ficheiros XML contendo apenas os textos relativos a cada uma das estruturas do relatório médico. Os valores totais de textos por estrutura podem ser analisados na tabela 2.1.

Na criação destes documentos é de notar a conversão de alguns caracteres para o formato específico XML, tais como os caracteres < e &, bem como a alteração de *encoding* de ISO-8859-1 para UTF-8, *encoding* a usar no *software* de anotação seleccionado.

A tabela 2.1 apresenta a distribuição de informação nos relatórios.

Tabela 2.1: Relatórios MedAlert

Documento	Tokens	Frases	Textos
Motivo Internamento	104 833	11851	8 563
História Clínica	1 179 960	56 202	9 775
Exame Físico	414 558	37 499	7 071
Evolução	474 303	26 663	8 106
Terapêutica Efectuada	332 017	11 569	8 363
Destino	219 189	13 834	6 351
Total	2 724 860	157 618	48 229

Uma primeira análise ao nível de verificação ortográfica foi realizada recorrendo ao verificador ortográfico JaSpell [1]. Os resultados são apresentados na tabela 2.2. Esta tabela apresenta não só o total de palavras identificadas como erros ortográficos, mas também a quantidade de palavras que são conhecidas e desconhecidas pelo verificador JaSpell. Este verificador ortográfico recorre a dicionários gerais de português e embora contenha uma lista de nomes próprios e acrónimos que deverão ser ignorados, não contém informação específica da área médica. É possível verificar que a quantidade de palavras desconhecidas pelo verificador ortográfico em textos com mais informação relativa a doenças e medicação (Motivo de Internamento e Terapêutica Efectuada) são os que apresentam mais erros, devido essencialmente ao uso de linguagem médica mais específica.

A distribuição de tokens, frases e textos por tipo de texto é apresentada na Figura 2.2.

Tabela 2.2: Distribuição de erros ortográficos

Documento	Desconhecida	Conhecida	Total	$\frac{total_i}{numTokens_i}$
Motivo Internamento	7 370	14 387	21 757	21%
História Clínica	60 884	108 697	169 581	14%
Exame Físico	15 194	39 471	54 665	13%
Evolução	21 502	42 772	64 274	14%
Terapêutica Efectuada	71 975	25 487	97 462	29%
Destino	12 212	21 636	33 848	15%
Total			441 487	16%

Figura 2.2: Distribuição de tokens, frases e textos no *corpus*

Capítulo 3

Metodologia de Anotação

Pretende-se construir uma colecção dourada anotada semanticamente, isto é, um *corpus* que inclua anotações relativas a *entidades* e aos *relacionamentos* entre estas entidades. Importa, primeiro, definir estes dois conceitos, no contexto médico. *Entidade* é algo real referido no texto, como por exemplo, a medicação prescrita, os exames realizados, etc. Os *relacionamentos* são então ligações entre entidades como o resultado de um exame, a medicação indicada para uma patologia, etc.

A anotação deverá ser introduzida no texto, e marcar uma área do texto com a categoria correspondente: medicação, local anatómico, etc. A anotação deverá também marcar palavras que modificam marcações, tais como negação e caracterização, e marcar ligações entre marcações.

Duas ou mais marcações podem referir-se à mesma *coisa* real, devendo neste caso ser marcadas como *correferências*. A Figura 3.1 apresenta alguns aspectos relevantes da anotação.

Figura 3.1: Exemplo ilustrativo de anotação.

3.1 Regras gerais de anotação

Os espaços imediatamente a seguir e antes da entidade a anotar não deverão ser anotados. Se a entidade contém espaços, esses devem manter-se inalterados.

(3.1) ...por <PROBLEMA>parestesias</PROBLEMA> do
<LOCALANATOMICO>membro inferior direito</LOCALANATOMICO> ...

As aspas, parênteses, pelicas ou travessões não são para incluir na anotação se englobarem a entidade como um todo. No entanto, são para incluir caso se apliquem a partes da entidade ou façam parte integrante da mesma.

(3.2) <PROBLEMA>Hipertensão arterial</PROBLEMA>
(<CARACTERIZACAO>controlada</CARACTERIZACAO>)

3.1.1 Recursividade das anotações

São permitidas anotações dentro de anotações.

(3.3) Abdómen <RESULTADO><NEGACAO>sem</NEGACAO> alterações
evidentes</RESULTADO>.

3.1.2 Vagueza na classificação semântica

No caso de haver dúvida entre várias categorias, ou existirem duas ou mais categorias para uma dada entidade, esta deve ser anotada com as várias categorias adequadas.

(3.4) <LOCALANATOMICO|EXAME>Abdómen<LOCALANATOMICO|EXAME> sem
alterações evidentes.

3.1.3 Escolha das entidades

Devido à utilização futura desta anotação em sistemas de extracção de informação médica, é extremamente importante a anotação do maior número de entidades com uma interpretação separada possível, em oposto à entidade máxima. Deste modo, deverão ser anotadas todas as entidades que contêm o menor número de palavras, anotando, assim, a maior quantidade de informação disponível no texto. Observem-se os seguintes exemplos.

(3.5) Certo: <PROBLEMA>Bronquite</PROBLEMA>
<CARACTERIZACAO>Aguda</CARACTERIZACAO>
Errado: <PROBLEMA>Bronquite Aguda</PROBLEMA>

No entanto, é necessário notar que esta escolha da entidade mínima não deve entrar em conflito com a definição da entidade, tal como é ilustrado de seguida.

(3.6) Certo: Abdómen <RESULTADO><NEGACAO>sem</NEGACAO> alterações evidentes</RESULTADO>.

Errado: Abdómen <NEGACAO>sem</NEGACAO> <RESULTADO>alterações evidentes</RESULTADO>.

Notar que o RESULTADO é o facto de **não** haver alterações evidentes e não apenas alterações evidentes. O contrário se verifica no exemplo 3.1.3, onde o PROBLEMA descrito é a existência de Bronquite com uma caracterização adicional de ser Aguda.

(3.7) Certo: <PROBLEMA>Insuficiência</LOCALANATOMICO>Cardíaca</LOCALANATOMICO></PROBLEMA>...

Errado: <PROBLEMA>Insuficiência</PROBLEMA><LOCALANATOMICO>Cardíaca</LOCALANATOMICO>...

3.1.4 Anotação dos relacionamentos entre entidades

Só deverão ser marcados os relacionamentos definidos no esquema de anotação definido para as estruturas dos relatórios em análise, bem como apenas deverão ser marcados relacionamentos existentes no mesmo texto, isto é, não há relacionamentos entre entidades de textos distintos.

3.1.5 Relacionamentos múltiplos de uma entidade

Uma entidade pode possuir diferentes relacionamentos com mais do que uma outra entidade, tendo de ser marcados, nesse caso, todos os relacionamentos existentes, desde que definidos no esquema de anotação.

3.1.6 Equivalência entre relacionamentos

A marcação de um relacionamento implica o seu inverso, não sendo necessário marcar ambos os relacionamentos.

3.2 Entidades

Apresenta-se de seguida uma lista de todas as entidades definidas, bem como a sua descrição e exemplário.

1. **Problema** Todas as referências a estados físicos, condições, sinais, patologias e diagnósticos. Esta categoria pode ser dividida nos seguintes tipos:

- Sinal
 - (a) **Prostração** marcada.
 - (b) **Hemorragia** digestiva alta.

- (c) **Edema** generalizado.
- (d) Foi internado por **febre**, dispneia, **expectoração** muco-purulenta e hemoptoica...
- (e) ... **sem imagens de condensação**...
- Sintoma
 - (a) **Poliartralgias** MIs.
 - (b) Foi internado por febre, **dispneia**...
- Diagnóstico
 - (a) **Dpoc** agudizada.
 - (b) Quadro **semi-comatoso**.
 - (c) **Anemia** NN.
 - (d) **Insuficiência renal**.
 - (e) Antecedentes de **alcoolismo** crónico, **dislipidemia**, **HTA** ...
- Patologia
 - (a) **Bronquite** Aguda.
 - (b) ... cuja EDA revelou **ulcera** gástrica...
 - (c) A doente tem antecedentes de **Diabetes Mellitus** tipo 2.

Em caso de dúvida, ou não aplicação dos tipos listados, anotar como PROBLEMA.

2. **Local Anatómico** Todas as referências a sítios anatómicos.
 - (a) Poliartralgias **MI**s, predominantemente na **articulação tibiotársica** esquerda...
 - (b) Hemorragia **digestiva** alta.
3. **Tempo** As entidades de tempo devem incluir todas as palavras que indiquem o momento temporal, isto é, que respondam à pergunta <PREP> *quando?* mesmo que a resposta não indique uma data ou ocasião temporal exacta. Não existem restrições na categoria morfológica da palavra. A categoria pode ser dividida nos seguintes tipos:
 - Tempo Calendário
 - (a) Recorreu no dia **14/07/2006** ao Serviço de Urgência...
 - (b) Deu entrada no SU em **08/01/05**...
 - Duração
 - (a) **Durante o internamento**, foi-se desenvolvendo ...
 - (b) Mantiveram-se manobras de Suporte Avançado de Vida **durante 15 minutos**, sem sucesso.

- Frequência
 - (a) ... a repetir **cada meia hora** em SOS.

Caso nenhum dos tipos descreva a entidade em análise, esta deverá ser marcada apenas com a categoria **TEMPO**:

- (a) **Antecedentes** de alcoolismo crónico.
- (b) HTA, AVC **há 2 anos**.
- (c) **Ao 3º dia de internamento**...
- (d) ...tendo sido realizada biópsia na endoscopia **à data de alta**.

4. **Exame** Engloba todas as referências a exames descritos nos relatórios. Pode ser dividida nos seguintes tipos:

- Físico
 - (a) **Auscultação pulmonar**.
- Analítico
 - (a) tendo sido realizada **biópsia** na endoscopia à data de alta.
- Imagiológico
 - (a) ... cuja **EDA** revelou ulcera gástrica...
 - (b) tendo sido realizada biópsia na **endoscopia** à data de alta.
 - (c) **Radiologicamente** apresentava infiltrado heterogenio bilateral.

Em caso de dúvida, ou não aplicação dos tipos listados, anotar como EXAME.

5. **Resultado** Todas as referências a resultados de exames **que não incluam entidades categorizadas por PROBLEMA**.

- (a) Abdómen **sem alterações evidentes**.

6. **Valor** Expressões numéricas referentes a valores. Podem referir quantidades absolutas ou relativas (mais comuns nestes tipo de documentos) ou classificações. Os itens numéricos a marcar ordem no texto não são considerados entidades. Quando há uma referência a um intervalo de valores, os seus limites devem ser anotados como uma única entidade. As unidades que caracterizam as expressões numéricas devem ser incluídas na anotação. A categoria **VALOR** foi dividida nos seguintes tipos:

- Quantidade
 - (a) ...no contexto de DPOC agudizada com hipoxímia de **53**, sem leucocitose e PCR de **17,8**.

- (b) Furosemide **20mg** ev id, ...
 - Dosagem
 - (a) Diovan **80**, Captopril, Diamicron LM **30**,...
 - Classificação
 - (a)
7. **Terapêutica** Todas as referências a procedimentos com interesse no tratamento curativo ou preventivo de doenças. Encontra-se dividida nos seguintes tipos e subtipos:
- Química
 - Medicação
 - * Componente Activo
 - (a) Encontrava-se medicada com : Diovan 80, Captopril, Diamicron LM 30, **Ticlopidina**, **Sinvastatina** e Lexotan 1,5mg.
 - (b) medicado em ambulatório com Hygroton, Captopril25, Adalat CR 30mg e **AAS150mg**.
 - * Nome Comercial
 - (a) Encontrava-se medicada com : **Diovan 80**, **Captopril**, **Diamicron LM 30**, Ticlopidina, Sinvastatina e **Lexotan 1,5mg**.
 - (b) medicado em ambulatório com **Hygroton**, **Captopril25**, **Adalat CR 30mg** e AAS150mg.
 - Substância
 - (a) Intoxicação voluntária com **Paraquat**.
 - (b) Saturação de **O2** de 83%.
 - Física
 - (a)
 - Cirúrgica
 - (a) ...no seguimento de uma **artroplastia** total do joelho esquerdo.
8. **Via de administração** Via de administração de uma terapêutica química.
- (a) Ac. clavulínico (**ev+ po**), claritromicina (**po**), enoxaparina (**sc**), ranitidina (**ev+ po**), prednisolona (**ev**), furosemido (**ev+ po**), salbutamol (**nebulização**), brometo de ipatropio (**neb.**) e O2 por **sonda nasal**.

Alguns acrónimos usados para designar vias de administração mais comuns são listadas de seguida:

- IV - intravenosa
 - EV - endovenosa
 - IM - intramuscular
 - SC - subcutânea
 - ID - intradérmica
 - IT - intratecal
 - IO - intra-óssea
 - ...
9. **Evolução** Acompanhamento e avaliação os sinais e sintomas de um paciente, com ou sem tratamento.
- (a) A DPOC teve **evolução clínica** favorável.
 - (b) Após resposição do trânsito intestinal, teve boa **evolução** do estado neurológico.
10. **Local** Referências a sítios específicos.
- (a) Com internamento em **Coimbra**...
 - (b) Terapêutica proposta para **domicílio**: ...
11. **Caracterização** Todos as indicações de caracterização.
- (a) Bronquite **Aguda**.
 - (b) Anemia **NN**.
 - (c) Prostação **marcada**.
 - (d) Hemorragia digestiva **alta**.
 - (e) Foi internado por febre, dispneia, expectoração **muco-purulenta** e **hemoptoica**
12. **Negação** Todos as indicações de negação.
- (a) ...**sem** imagens de condensação.
 - (b) ...**sem** leucocitose.

3.3 Relacionamentos

Os relacionamentos definidos no esquema de anotação estão resumidos na Tabela 3.1.

Dos vários tipos de relacionamento destacam-se os relacionamentos de *inclusão*, estabelecidos quando uma entidade faz parte de outra, e os de *localização* ou de *ocorrência em*. Os primeiros têm como única restrição a exigência de que as entidades relacionadas sejam da mesma categoria, tendo sido nomeados de *inclui*. Os segundos ocorrem entre entidades das categorias PROBLEMA ou EXAME e entidades da categoria LOCAL ANATOMICO, indicando a localização de um problema ou de um exame.

Outros relacionamentos, de âmbito mais alargado, foram definidos:

- os relacionamentos **caracterizado por** ou **negado por** entre as entidades PROBLEMA e RESULTADO e as entidades CARACTERIZACAO e NEGACAO;
- relacionamento **datado de**, que ocorre nas entidades entre as quais possa existir uma indicação temporal;
- o relacionamento **quantificado por** ocorre entre entidades com possibilidade de quantificação e as entidades da categoria VALOR;
- o relacionamento **indica** relaciona as entidades das categorias PROBLEMA e RESULTADO e a categoria EXAME, relacionando um exame com o seu resultado.

Tabela 3.1: Relacionamentos MedAlert.

	Relacionamento	Categorias a que se aplica
1	inclui	EXAME, PROBLEMA, LOCAL ANATOMICO, TERAPEUTICA
2	tem_alvo	PROBLEMA ou EXAME ou TERAPEUTICA e LOCAL ANATOMICO
3	ocorre_em	TERAPEUTICA e EXAME e LOCAL
4	caracterizado_por	PROBLEMA ou LOCAL ANATOMICO ou RESULTADO e CARACTERIZACAO
5	negado_por	RESULTADO ou PROBLEMA e NEGACAO
6	datado_de	EXAME ou PROBLEMA ou RESULTADO ou MEDICACAO ou POSOLOGIA e TEMPO
7	quantificado_por	POSOLOGIA ou RESULTADO ou PROBLEMA e VALOR
8	indica	PROBLEMA ou RESULTADO e EXAME
9	indicada	TERAPEUTICA e PROBLEMA
10	evolui	TERAPEUTICA ou PROBLEMA e EVOLUCAO
11	tem_duracao	TERAPEUTICAQUIMICA e DURACAO
12	tem_frequencia	TERAPEUTICAQUIMICA e FREQUENCIA
13	tem_quantidade	TERAPEUTICAQUIMICA e QUANTIDADE
14	tem_dosagem	TERAPEUTICAQUIMICA e DOSAGEM
15	tem_quimicos	PROBLEMA e TERAPEUTICAQUIMICA
16	aplicada_em	TERAPEUTICAQUIMICA e VIAADMINISTRACAO

3.3.1 Relacionamento de identidade

Para além dos relacionamentos anteriores, foi ainda definido o relacionamento de identidade estabelecido entre entidades que tenham o mesmo referente, ou seja, que designem a mesma entidade: *correferência*. Desta forma, este relacionamento só existe entre entidades que pertençam à mesma categoria.

- (3.8) Recorreu no dia 14/07/2006 ao SU por parestesias do **membro inferior direito** e diminuição da força muscular **deste membro**.

3.4 Esquema de anotação

Figura 3.2: Esquema de anotação

A secção seguinte apresenta alguns exemplos de anotação que deverão servir de guia na anotação.

3.5 Exemplos de anotação

3.5.1 Motivo de Internamento

Definição

Descreve a patologia/diagnóstico de presunção na ocasião do internamento, bem como as suas características e localização.

Formato

Frase simples, frequentemente um ou mais diagnósticos ou uma lista de problemas.

Exemplo de anotação

Prostração marcada, quadro semi-comatoso, poliartralgias MIS e edema generalizado. Anemia NN, insuficiência renal.

```
<PROBLEMA ID='p1' REL='c1' TIPOREL='caracterizado'>
Prostração
</PROBLEMA>
<CARACTERIZACAO ID='c1'>
marcada
</CARACTERIZACAO>
, quadro
<PROBLEMA ID='p2'>
semi-comatoso
</PROBLEMA>
,
<PROBLEMA ID='p3' REL='la1' TIPOREL='tem_alvo'>
poliartralgias
</PROBLEMA>
<LOCALANATOMICO ID='la1'>
MIS
</LOCALANATOMICO>
e
<PROBLEMA ID='p4' REL='c2' TIPOREL='caracterizado'>
edema
</PROBLEMA>
<CARACTERIZACAO ID='c2'>
generalizado
```

```

</CARACTERIZACAO>
.
<PROBLEMA ID='p5' REL='c3' TIPOREL='caracterizado'>
Anemia
</PROBLEMA>
<CARACTERIZACAO ID='c3'>
NN
</CARACTERIZACAO>
,
<PROBLEMA ID='p6' REL='la2' TIPOREL='tem.alvo'>
insuficiência
<LOCALANATOMICO ID='la2'>
renal
</LOCALANATOMICO>
</PROBLEMA>
.

```

Fontes de conhecimento a usar no reconhecimento automático:

- Vocabulário disponibilizado por DeSC (UMLS traduzido),
- Lista de análises e exames HIP,
- Wikipedia,
- ICD-9 CM
- ...

3.5.2 História Clínica

Definição

Descreve o historial clínico mais relevante do doente à data de internamento. Inclui informação como idade, doenças relevantes, características do doente no início do tratamento, observações médicas, resultados de exames efectuados no início do internamento e medicação crónica.

Formato

Texto livre.

Exemplo de anotação

Doente de 56 anos deu entrada no SU em 08/01/05 com prostração marcada, poliartralgias dos MIs, predominantemente na articulação tibiotársica esquerda e edemas generalizados. Antecedentes de alcoolismo crónico, dislipidémia, HTA e hiperuricémia.

```
Doente de 56 anos deu entrada no SU
<TEMPOCALENDARIO ID='t1'>
08/01/05
</TEMPOCALENDARIO>
com
<PROBLEMA ID='p1' REL='t1' TIPOREL='datado' REL='c1' TIPOREL='caracterizado'>
prostração
</PROBLEMA>
<CARACTERIZACAO ID='c1'>
marcada
</CARACTERIZACAO>
,
<PROBLEMA ID='p2' REL='la1' TIPOREL='tem.alvo' REL='t1' TIPOREL='datado'>
poliartralgias
</PROBLEMA>
dos
<LOCALANATOMICO ID='la1'> REL='la2' TIPOREL='inclui'>
MIs
</LOCALANATOMICO>
, predominantemente na
<LOCALANATOMICO ID='la2' REL='c2' TIPOREL='caracterizado'>
articulação tibiotársica
</LOCALANATOMICO>
<CARACTERIZACAO ID='c2'>
esquerda
</CARACTERIZACAO>
e
<PROBLEMA ID='p3' REL='c3' TIPOREL='caracterizado' REL='t1' TIPOREL='datado'>
edemas
</PROBLEMA>
<CARACTERIZACAO ID='c3'>
generalizados
</CARACTERIZACAO>
.
<TEMPO ID='t2'>
Antecedentes
</TEMPO>
```



```

de
<PROBLEMA ID='p4' REL='t2' TIPOREL='datado' REL='c4' TIPOREL='caracterizado'>
alcoolismo
</PROBLEMA>
<CARACTERIZACAO ID='c4'>
crónico
</CARACTERIZACAO>
,
<PROBLEMA ID='p5' REL='t2' TIPOREL='datado'>
displidémia
</PROBLEMA>
,
<PROBLEMA ID='p6' REL='t2' TIPOREL='datado'>
HTA
</PROBLEMA>
e
<PROBLEMA ID='p7' REL='t2' TIPOREL='datado'>
hiperuricémia
</PROBLEMA>
.

```

Fontes de conhecimento a usar no reconhecimento automático:

- Vocabulário disponibilizado por DeSC (UMLS traduzido),
- ICD-9 CM,
- Lista de exames efectuados no HIP,
- lista de medicação comercializada em Portugal.
- Wikipedia

3.5.3 Exame Físico

Definição

Descrição de exames relevantes efectuados.

Formato

Frase simples contendo vários exames separados por pontuação.

Exemplo de anotação

Ao exame objectivo de entrada, apresentava-se apirético, T.A.=115/74mmHg , F.R.=80 bpm, dispneico, sem cianose. À auscultação pulmonar, presença de roncos bilaterais, sem outras alterações audíveis.

```

Ao
<EXAME ID='e1' REL='t1' TIPOREL='datado' REL='p1' TIPOREL='indica'
REL='e2' TIPOREL='inclui' REL='e3' TIPOREL='inclui' REL='p2' TIPOREL='indica'
REL='p3' TIPOREL='indica'>
exame objectivo
</EXAME>
de
<TEMPO ID='t1'>
entrada
</TEMPO>
, apresentava-se
<PROBLEMA ID='p1'>
apirético
</PROBLEMA>
,
<EXAME ID='e2' REL='r1' TIPOREL='indica'>
T.A.
</EXAME>
=
<RESULTADO|VALOR ID='r1|v1'>
115/74mmHg
</RESULTADO|VALOR>
,
<EXAME ID='e3' REL='r2' TIPOREL='indica'>
F.R.
</EXAME>
=
<RESULTADO|VALOR ID='r2'|'v2'>
80 bpm
</RESULTADO|VALOR>
,
<PROBLEMA ID='p2'>
dispneico
</PROBLEMA>
,
<PROBLEMA ID='p3' REL='n1' TIPOREL='negado'>
<NEGACAO ID='n1'>

```

```

sem
</NEGACAO>
cianose
</PROBLEMA>
.   À
<EXAME ID='e4' REL='la1' TIPOREL='tem_alvo' REL='p4' TIPOREL='indica'
REL='r3' TIPOREL='indica'>
auscultação
<LOCALANATOMICO ID='la1'>
pulmonar
</LOCALANATOMICO>
</EXAME>
, presença de
<PROBLEMA ID='p4' REL='la2' TIPOREL='tem_alvo'>
roncos
<LOCALANATOMICO ID='la2'>
bilaterais
</LOCALANATOMICO>
</PROBLEMA>
,
<RESULTADO ID='r3' REL='n2' TIPOREL='negado'>
<NEGACAO ID='n2'>
sem outras
</NEGACAO>
alterações audíveis
</RESULTADO>
.

```

Fonte de conhecimento

- Lista de exames e análises efectuados no HIP.

3.5.4 Evolução

Definição

Descreve a evolução apresentada pelo doente durante o tempo de internamento. Inclui informação sobre os exames e medicação efectuados.

Formato:

Texto livre.

Exemplo de anotação

Ao 3º dia de internamento apresentou um quadro clínico de Edema agudo do Pulmão de natureza hipertensiva que teve boa evolução após otimização terapêutica. A DPOC agudizada teve evolução clínica favorável.

```

<TEMPO ID='t1'>
Ao 3º dia de internamento
</TEMPO>
apresentou um quadro clínico de
<PROBLEMA ID='p1' REL='c1' TIPOREL='caracterizado' REL='c2' TIPOREL='caracterizado'
REL='la1' TIPOREL='tem_alvo' REL='e1' TIPOREL='evolui'>
Edema
</PROBLEMA>
<CARACTERIZACAO ID='c1'>
agudo
</CARACTERIZACAO>
do
<LOCALANATOMICO ID='la1'>
Pulmão
</LOCALANATOMICO>
de
<CARACTERIZACAO ID='c2'>
natureza hipertensiva
</CARACTERIZACAO>
que teve
<CARACTERIZACAO ID='c3'>
boa
</CARACTERIZACAO>
<EVOLUCAO ID='e1' REL='c3' TIPOREL='caracterizado' REL='t2' TIPOREL='datado'>
evolução
</EVOLUCAO>
<TEMPO ID='t2'>
após
</TEMPO>
<TERAPEUTICA ID='tp1'>
otimização terapêutica
</TERAPEUTICA>
. A
<PROBLEMA ID='p2' REL='c4' TIPOREL='caracterizado' REL='e2' TIPOREL='evolui'>
DPOC
</PROBLEMA>
<CARACTERIZACAO ID='c4'>

```

```

agudizada
</CARACTERIZACAO>
teve
<EVOLUCAO ID='e2' REL='c5' TIPOREL='caracterizado'>
evolução clínica
</EVOLUCAO>
<CARACTERIZACAO ID='c5'>
favoravel
</CARACTERIZACAO>
.

```

Fonte de conhecimento

- Vocabulário disponibilizado por DeSC (UMLS traduzido),
- ICD-9 CM,
- Lista de exames efectuados no HIP,
- lista de medicação comercializada em Portugal (a obter).
- Wikipedia

3.5.5 Terapêutica Efectuada

Definição

Descreve a medicação

Formato

Frase simples, medicação separada por pontuação.

Exemplo de anotação

Furosemide 20mg ev id, tiamina 100mg ev id, glibenclamida 5mg po de manhã, 2U GV. Terapêutica proposta para domicílio: Thirax 0.1 id, indapamida, enalapril 5.

```

<MEDICACAO ID='m1' REL='q1' TIPOREL='tem_quantidade' REL='va1' TIPOREL='aplicada_em'
REL='f1' TIPOREL='tem_frequencia'>
Furosemide
</MEDICACAO>
<QUANTIDADE ID='q1'>
20mg
</QUANTIDADE>

```

```

<VIAADMINISTRACAO ID='va1'>
ev
</VIAADMINISTRACAO>
<FREQUENCIA ID='f1'>
id
</FREQUENCIA>
,
<MEDICACAO ID='m2' REL='q2' TIPOREL='tem_quantidade' REL='va2' TIPOREL='aplicada_em'
REL='f1' TIPOREL='tem_frequencia'>
tiamina
</MEDICACAO>
<QUANTIDADE ID='q2'>
100mg
</QUANTIDADE>
<VIAADMINISTRACAO ID='va2'>
ev
</VIAADMINISTRACAO>
<FREQUENCIA ID='f2'>
id
</FREQUENCIA>
,
<MEDICACAO ID='m3' REL='q3' TIPOREL='tem_quantidade' REL='va3' TIPOREL='aplicada_em'
REL='t1' TIPOREL='datado'>
glibenclamida
</MEDICACAO>
<QUANTIDADE ID='q3'>
5mg
</QUANTIDADE>
<VIAADMINISTRACAO ID='va3'>
po
</VIAADMINISTRACAO>
<TEMPO ID='t1'>
de manhã
</TEMPO>
,
<QUANTIDADE ID='q4'>
2U
</QUANTIDADE>
<SUBSTANCIA ID='s1' REL='q4' TIPOREL='tem_quantidade'>
GV
</SUBSTANCIA>
.
<TERAPEUTICA ID='tp1' REL='l1' TIPOREL='ocorre_em' REL='m4' TIPOREL='inclui'
REL='m5' TIPOREL='inclui' REL='m6' TIPOREL='inclui'>

```

```

Terapêutica proposta
</TERAPEUTICA>
para
<LOCAL ID='l1'>
domicílio
</LOCAL>
:
<MEDICACAO ID='m4' REL='d1' TIPOREL='tem_dosagem' REL='f3' TIPOREL='tem_frequencia'>
Thirax
</MEDICACAO>
<DOSAGEM ID='d1'>
0.1
</DOSAGEM>
<FREQUENCIA ID='f3'>
id
</FREQUENCIA>
,
<MEDICACAO ID='m5'>
indapamida
</MEDICACAO>
,
<MEDICACAO ID='m6' REL='d2' TIPOREL='tem_dosagem'>
enalapril
</MEDICACAO>
<DOSAGEM ID='d2'>
5
</DOSAGEM>
.

```

Fonte de Conhecimento

- Lista de medicação comercializada em Portugal.

Capítulo 4

Software de anotação

De modo a realizar a anotação de uma forma consistente os esquemas de anotação foram modelados como ontologias Protégé-Frames. A anotação deverá ser realizada usando o *plugin* Knowtator para Protégé. Este foi escolhido, após uma avaliação de outras ferramentas disponíveis (...), pelo facto de lidar com relacionamentos.

De modo a iniciar a anotação será necessário obter o directório `knowtator` disponibilizado em www.ieeta.pt/~lsferreira/knowtator.zip. A pasta descompactada deverá ter a estrutura apresentada na Figura 4.

Figura 4.1: Estrutura do directório knowtator.

4.1 Instalação

1. Instalar o Protégé¹ versão 3.3.
2. Fazer o download do Knowtator².
3. Descompactar o `knowtator-VERSION.zip` para o directório de *plugins* do Protégé (por exemplo, `C:/Program Files/Protege_3.3.1/plugins`)
4. Verificar que o directório `<protege-home>/plugins/edu.uchsc.ccp.knowtator` existe.
5. Fechar o Protégé caso ainda esteja a correr.

4.2 Abrir os projectos de anotação

Para abrir o projectos de anotação seguir os seguintes passos:

1. Abrir o Protégé e fechar a caixa de boas vindas.
2. Seleccionar Menu -> File -> Open.

3. Abrir o directório **Projectos** e seleccionar o ficheiro `esquema.pprj`. Deverá obter uma imagem semelhante à da Figura 4.2.

¹<http://protege.stanford.edu/>

²<http://knowtator.sourceforge.net/>

Figura 4.2: Protégé GUI

4. O separador **knowtator** deverá aparecer automaticamente na janela. Caso este não apareça automaticamente realizar os seguintes passos:

- Seleccionar Menu -> Project -> Configure... e escolher o separador Tab Widgets na caixa de diálogo.

- Seleccionar a *checkbox* com o nome **knowtator** e clicar OK.

5. O projecto de anotação está pronto a ser usado. Verificar se o **knowtator** foi correctamente carregado clicando no separador **knowtator**. Deverá ter uma imagem semelhante à da Figura 4.3.

Figura 4.3: Projecto de anotação no Knowtator.

4.3 Anotação

Para iniciar a anotação é necessário carregar os ficheiros de texto correspondentes à estrutura a anotar. Para tal, usar os botões de navegação (Figura 4.4) para seleccionar o directório adequado.

Figura 4.4: Botões de navegação na colecção de texto.

Figura 4.5: Seleccionar os documentos a anotar

Deverá poder visualizar o primeiro relatório no separador **knowtator**. Os botões de navegação laterais também deverão estar activados.

4.3.1 Anotação de entidades

Para anotar uma porção de texto correspondente à categoria *Local Anatómico*:

1. Seleccionar a porção de texto a anotar.

2. Clicar na classe *Local Anatômico* do esquema de anotação. Uma caixa de diálogo irá aparecer.

3. Seleccionar *create annotation...* na caixa de diálogo.
4. A primeira anotação foi criada e deverá estar seleccionada. Todos os detalhes desta anotação aparecem no lado direito da janela.

Figura 4.6: Anotação de entidades.

4.3.2 Anotação de relacionamentos

De modo a inserir os relacionamentos existentes entre as várias entidades:

1. Clicar na entidade a relacionar.
2. Seleccionar no lado direito do separador o relacionamento a anotar.
3. Clicar no botão .
4. Seleccionar na caixa de diálogo a anotação correspondente à entidade a ser relacionada.

Note-se que só são apresentadas as anotações correspondentes à classe da anotação a ser relacionada. Caso exista apenas uma anotação, esta é automaticamente adicionada ao relacionamento.

Figura 4.7: Anotação de relacionamentos.

Correferência

Sempre que a mesma entidade for mencionada duas vezes no mesmo texto:

1. Clicar na classe *Correferência* no esquema de anotação.
2. Seleccionar **Create Correferência annotation** no menu. Uma anotação vazia é criada e seleccionada.

Figura 4.8: Anotação de correferência.

3. Clicar no botão para o relacionamento *Correferência[CLASSE]*.
4. Seleccionar na caixa de diálogo duas (ou mais) anotações que correspondam à mesma entidade.

4.4 Criar XML com anotação

Após a conclusão do processo de anotação a melhor forma de retirar as anotações do knowtator é através da exportação para XML. Para tal, seleccionar a opção do menu **Knowtator -> Export annotations to XML** e seguir as instruções. Este processo criará um ficheiro XML por texto. Os ficheiros XML deverão ser enviados para lsferreira@ua.pt.

Bibliografia

- [Santos & Cardoso 2007] Diana Santos & Nuno Cardoso (eds.). Reconhecimento de entidades mencionadas em português: Documentacao e actas do HAREM, a primeira avaliacao conjunta na área, *Linguateca. 2007*,
http://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf
- [Mota & Santos 2008] Cristina Mota & Diana Santos (eds.). Desafios na avaliacao conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, *Linguateca. 2008* ,
www.linguateca.pt
- [Roberts et al.] , A. Roberts and R. Gaizauskas and M. Hepple and G. Demetriou and Y. Guo and A. Setzer. 2008.Semantic Annotation of Clinical Text: The CLEF Corpus. In Proceedings of the LREC 2008, *Workshop on Building and Evaluating Resources for Biomedical Text Mining*,
<http://nlp.shef.ac.uk/clef/>
- [Ogren] , Philip V. Ogren. Knowtator: A plug-in for creating training and evaluation data sets for Biomedical Natural Language systems . In Proceedings of the 9th International Protege Conference. 73–76,
<http://knowtator.sourceforge.net/>

Appendix B

MEDINX type system

Figure B.1: Entity type system inheritance diagram.

Figure B.2: MediNXConcept type system inheritance diagram.

Appendix **C**

Evaluation Manual

MedInX

Manual do Avaliador

Liliana Ferreira
Instituto de Engenharia Electrónica e Telemática de Aveiro
Universidade de Aveiro

`lsferreira@ua.pt`

Outubro de 2010

Resumo

O projecto MedInX (Medical Information eXtraction) tem por objectivo o desenvolvimento de um sistema capaz de processar linguagem médica. Uma parte importante da informação requerida por este tipo de sistemas é originada em formato de texto, embora, actualmente e cada vez mais, em formato electrónico.

O MedInX usa tecnologias de extracção automática de informação para tornar esta informação não estruturada disponível, através da identificação de entidades semânticas e de relações entre estas. Para tal, possui actualmente um *corpus* de 917 textos médicos relativos a episódios de internamento no Hospital Infante D. Pedro, Aveiro.

Este relatório pretende funcionar como um manual do processo de avaliação de um subconjunto deste *corpus*, o qual foi anotado automaticamente pelo sistema com as entidades clínicas e suas relações. Pretende-se com este manual garantir a consistência da avaliação e melhorar a concordância entre avaliadores.

Índice

1. Iniciar a avaliação	4
2. Avaliação da Precisão (Precision)	5
3. Avaliação da Abrangência (Recall)	11
4. Verificação do estado actual da avaliação	13
5. Terminar a avaliação	13

1. Iniciar a avaliação

Para iniciar a avaliação deverá introduzir o endereço <http://www.ieeta.pt/~lsferreira/jsp/> no *browser* Mozilla Firefox¹ e inserir o nome de utilizador (*User*) e respectiva *Password* no formulário apresentado (Figura 1).

Figura 1 – Login MedInX

Ao clicar em `Login` será reencaminhado para a página inicial da avaliação (Figura 2).

¹ <http://www.mozilla-europe.org/pt/firefox/>

Figura 2 – Início da Avaliação

A avaliação encontra-se dividida em duas fases: uma correspondente à avaliação da precisão do sistema e outra relativa à abrangência. O procedimento para cada uma destas fases é descrito de seguida.

2. Avaliação da Precisão (Precision)

Esta fase da avaliação destina-se a averiguar qual a quantidade de entidades identificadas correctamente pelo sistema. Assim, pretende-se que o avaliador determine a validade das entidades, bem como das respectivas categorias. O processo a seguir é descrito de seguida.

Para iniciar a avaliação deverá seleccionar na caixa Review correspondente ao relatório a avaliar a opção Precision (Figura 3).

Figura 3 – Iniciar a avaliação da Precisão

Será reencaminhado para a página da avaliação da Precisão do relatório escolhido, onde poderá visualizar o texto do relatório (Figura 4).

As palavras marcadas a cinzento correspondem a entidades identificadas pelo sistema. Ao passar o *rato* na área acinzentada será possível visualizar quais os relacionamentos que a entidade em questão tem com as restantes entidades do texto (Figura 4).

Figura 4 – Precisão

Para uma melhor compreensão do processo de anotação do MedInX, apresenta-se nas Tabelas 1 e 2 todas as categorias e tipos das entidades e relacionamentos utilizados na extracção de entidades bem como uma pequena descrição de cada uma.

Categoria	Tipo	Descrição
ANATOMICAL SITE		Todas as referências a sítios anatómicos.
CONDITION	DISEASE	Todas as referências a estados físicos, condições, sinais, patologias e diagnósticos.
	SIGNSYMPTOM	
	FINDING	
	PATHOLOGY	
EVOLUTION		Todas as referências ao acompanhamento e evolução das condições de um doente.
PROCEDURE	THERAPEUTIC	Engloba todas as referências aos procedimentos descritos nos relatórios, sejam estes terapêuticos, de diagnóstico ou laboratoriais.
	LABORATORY	
	DIAGNOSTIC	
CHEMICAL	MEDICATION	Procedimentos efectuados com recurso a substâncias químicas, sejam estas medicação, substâncias activas ou outras substâncias.
	ACTIVESUBSTANCE	
	SUBSTANCE	
MODIFIER		Todas as expressões caracterizadoras de uma entidade.
NEGATION		Todas as expressões que neguem uma entidade. Estas expressões devem ser marcadas separadamente da entidade a que se referem.
LATERALIZATION		Todas as referências à lateralização de uma condição, procedimento ou local anatómico(esquerda, direita, superior, etc...)
ADMINISTRATIONROUTE		Via de administração de um procedimento químico.
TIME	CALENDARTIME	As entidades de tempo devem incluir todas as palavras que indiquem o momento temporal, isto é, que respondam à pergunta <preposição> quando? mesmo que a resposta não indique uma data ou ocasião temporal exacta. Não existem restrições à categoria morfológica da palavra.
	DURATION	
	FREQUENCY	
VALUE	CLASSIFICATION	Expressões numéricas referentes a valores. Podem referir quantidades absolutas ou relativas ou classificações. Quando há uma referência a um intervalo de valores, os seus limites devem ser anotados como uma única entidade. As unidades que caracterizam as expressões numéricas devem ser incluídas na anotação.
	DOSAGE	
	QUANTITY	

Tabela 1 – Entidades MedInX

Relações	Categorias a que se aplica
INCLUDES	Relação entre duas entidades das categorias CONDITION ou PROCEDURE
HASMODIFIER	Relação entre as entidades das categorias CONDITION, PROCEDURE, EVOLUTION, ANATOMICALSITE e as entidades de MODIFIER
HASLATERALIZATION	Relação entre as entidades das categorias ANATOMICALSITE, CONDITION e as entidades de LATERALIZATION
HASNEGATION	Relação entre as entidades das categorias EVOLUTION, MODIFIER, PROCEDURE, CONDITION e as entidades de NEGATION
HASEVOLUTION	Relação entre as entidades das categorias CONDITION e EVOLUTION
HASTARGET	Relação entre as entidades das categorias PROCEDURE e CONDITION e as entidades de ANATOMICALSITE
HASTIME	Relação entre as entidades das categorias EVOLUTION, PROCEDURE e CONDITION e as entidades de TIME
HASVALUE	Relação entre as entidades das categorias CONDITION e PROCEDURE e as entidades de VALUE
HASSUBSTANCE	Relação entre as entidades da categoria CONDITION e as entidades do tipo SUBSTANCE
HASDURATION	Relação entre as entidades da categoria CHEMICAL e as entidades do tipo DURATION
HASFREQUENCY	Relação entre as entidades da categoria CHEMICAL e as entidades do tipo FREQUENCY
HASQUANTITY	Relação entre as entidades da categoria CHEMICAL e as entidades do tipo QUANTITY
HASDOSAGE	Relação entre as entidades da categoria CHEMICAL e as entidades do tipo DOSAGE
HASADMINISTRATIONROUTE	Relação entre as entidades da categoria CHEMICAL e as entidades da categoria ADMINISTRATIONROUTE
HASINDICATION	Relação entre as entidades da categoria PROCEDURE e as entidades da categoria CONDITION
HASFINDING	Relação entre as entidades da categoria PROCEDURE e as entidades do tipo FINDING

Tabela 2 – Relações MedInX

Ao clicar na entidade seleccionada aparecerá um caixa de diálogo com todas as características da entidade em análise (Figura 5). A avaliação é efectuada ao seleccionar uma das opções da área Evaluation, nomeadamente Correct; Incorrect; Don't know. Esta última opção deverá ser usada no caso de não ter a certeza sobre a correcta atribuição da categoria à entidade.

Caso não concorde com algum dos parâmetros descritos na caixa de diálogo, ou caso considere que a atribuição está incompleta poderá adicionar essa informação sob a forma de comentário clicando na opção Enter a comment...

Também poderá usar este procedimento sempre que existam dúvidas ou caso identifique algum factor digno de registo.

Figura 5 – Avaliação das entidades

Caso a entidade a ser avaliada esteja relacionada com outras entidades do texto a caixa de diálogo terá uma secção específica para a avaliação dessas relações: a secção *Relations* (Figura 6).

Nesta secção é possível visualizar o nome das relações, o texto da entidade com a qual esta está relacionada e a *checkbox* para avaliação, na qual deverá indicar se a entidade está correcta ou não.

Figura 6 - Avaliação das relações entre entidades

Para finalizar a avaliação da entidade seleccione a opção *Save*. A entidade avaliada assumirá uma cor diferente: verde no caso de estar correcta (Figura 7), vermelha no caso de incorrecção e violeta para as avaliadas como *Don't know*.

Note que todas as avaliações podem ser alteradas até à submissão final do documento bastará para isso seleccionar a entidade que pretende alterar e efectuar as devidas alterações.

MedInX

Reports Results

Review - Precision (Out7003767Resources.xml)

You can see and edit the result below. If the result is ok, you can press the save button to "Submit" button the annotations.

7003767

Motivo Internamento

AVC Isquémico

História Clínica

Doente de 82 anos de idade trazida ao SU a 18/03/07 por afasia e hemiparésia à direita de início súbito, sem outras queixas acompanhantes. Trata-se de uma doente com antecedentes de patologia osteoarticular degenerativa e HTA, medicada em ambulatório com Triticum 100 (R) id, Co-Diovan (R) id, Zolpidem (R) id, Apton 20 (R) id, Arcoxia (R) id, AAS 100 (R) id ..

Figura 7 – Entidade avaliada como correcta.

Após a avaliação de todas as entidades poderá gravar a sua avaliação clicando na opção **Submit** no final do relatório (Figura 8).

Também poderá utilizar a opção **Submit** para submeter a avaliação efectuada até ao momento, mesmo que incompleta. O sistema guardará a informação adicionada no momento da submissão e poderá retomar em qualquer altura a avaliação seleccionando o relatório em questão na página inicial do sistema.

A qualquer momento poderá parar a anotação e evitar a submissão do relatório através da opção **Cancel**.

Visual Source

Cancel Submit

Contact | About | Help | Logout

Figura 8 – Submeter uma avaliação

3. Avaliação da Abrangência (Recall)

Com a avaliação da abrangência pretende-se perceber qual a quantidade de entidades relevantes para o sistema identificadas.

Assim, nesta fase da avaliação pretende-se que o avaliador identifique no texto entidades que o sistema não reconheceu e lhes atribua a categoria adequada. Para tal deverá seguir os procedimentos descritos de seguida.

Na página inicial da avaliação deverá seleccionar na caixa Review correspondente ao relatório a avaliar a opção Recall (Figura 3).

Será, assim, reencaminhado para a página da avaliação da Abrangência do relatório escolhido, onde poderá visualizar o texto do relatório (Figura 9).

Figura 9 – Abrangência

As entidades identificadas pelo sistema são mais uma vez marcadas a cinzento, embora não esteja disponível desta vez qualquer caixa de diálogo para avaliação.

Caso identifique no texto uma expressão correspondente a uma entidade (Tabela 1) deverá seleccionar a área do texto correspondente à entidade a ser adicionada e imediatamente será apresentada uma caixa de diálogo onde pode seleccionar a informação correspondente à nova entidade (Figura 10).

Figura 10 – Nova entidade

Na caixa de diálogo deverá seleccionar a Categoria e o Tipo adequados à expressão em questão (Figura 11).

Figura 11 – Atribuição de categoria à entidade.

Ao clicar na opção Save a nova entidade será criada e marcada a vermelho no texto (Figura 12).

por afasia e hemiparésia à direita de início súbito, sem outras
 tes de **patologia osteoarticular degenerativa** e HTA, medicada em aml
 pidem (R) id, Apton 20 (R) id, Arcoxia (R) id, AAS 100

Figura 12 – Nova entidade

Quando terminar a verificação de todo o texto deverá submeter a avaliação clicando em **Submit** (ou cancelar caso não esteja satisfeito com a avaliação efectuada). Será reencaminhado para a página inicial da avaliação.

4. Verificação do estado actual da avaliação

Na página inicial da avaliação poderá acompanhar o estado actual da sua avaliação (Figura 13). De cada vez que um relatório é submetido o nome do avaliador é adicionado na coluna correspondente ao tipo de avaliação efectuada (*Precision* ou *Recall*). Cada avaliador visualiza apenas o estado da sua própria avaliação.

The screenshot shows a web browser window with the MedInX Reports Results page. The page has a blue header with 'Reports' and 'Results' tabs. Below the header, there is a section titled 'Reports' with the text 'You can see all uploaded report in here.' Below this is a table with four columns: 'Name', 'Review (Precision)', 'Review (Recall)', and 'Review'. The table lists several reports, including 'Out7002484MedInX.xml', 'Out7008548MedInX.xml', 'Out7003767Resources.xml', 'Out5005733MedInX.xml', 'Out8002805MedInX.xml', 'Out8010941Resources.xml', 'Out5013467MedInX.xml', and 'Out7000242Resources.xml'. The 'Review (Precision)' and 'Review (Recall)' columns show the reviewer's name, and the 'Review' column shows a dropdown menu for selecting a review.

Name	Review (Precision)	Review (Recall)	Review
Out7002484MedInX.xml	-	-	...
Out7008548MedInX.xml	-	-	...
Out7003767Resources.xml	Ililana	Ililana	...
Out5005733MedInX.xml	-	-	...
Out8002805MedInX.xml	-	-	...
Out8010941Resources.xml	-	-	...
Out5013467MedInX.xml	-	-	...
Out7000242Resources.xml	-	-	...

Figura 13 – Relatório avaliado.

5. Terminar a avaliação

A qualquer momento poderá sair da página de avaliação bastando para tal seleccionar a opção **Logout** apresentada no final da página (Figura 14).

The screenshot shows the footer of the MedInX Reports Results page. It contains a table with three rows of reports: 'Out5015708MedInX.xml', 'Out8013982Resources.xml', and 'Out9000049Resources.xml'. Below the table is a footer bar with links for 'Contact', 'About', 'Help', and 'Logout'.

Out5015708MedInX.xml	-	-	...
Out8013982Resources.xml	-	-	...
Out9000049Resources.xml	-	-	...

Contact | About | Help | Logout

Appendix D

MEDINX Results

Table D.1: Descriptive statistics for the amount of entities extracted from the 30 reports of the *Ontology+Resources* run.

Entity	N	Mean	Standard Deviation	Minimum	Q ₁	Q ₂	Maximum
Administration Route	86	3	4	0	0	4	13
Anatomical Site	99	3	2	0	1	4	10
Chemical	456	15	7	0	11	17	33
Condition	764	25	11	2	17	31	55
Evolution	22	1	1	0	0	1	2
Lateralization	113	4	3	0	2	5	12
Modifier	227	8	5	1	5	10	19
Negation	128	4	3	0	1	6	12
Procedure	328	11	8	0	6	13	40
Time	301	10	5	0	6	16	21
Value	331	11	8	0	6	15	36

Table D.2: Descriptive statistics for the amount of entities extracted from the 20 reports of the *Resources* run.

Entity	N	Mean	Standard Deviation	Minimum	Q ₁	Q ₂	Maximum
Administration Route	46	2	3	0	0	4	10
Anatomical Site	15	1	1	0	0	1	3
Chemical	245	12	5	0	11	17	18
Condition	294	15	8	1	10	18	33
Evolution	15	1	1	0	0	1	2
Lateralization	54	3	2	0	1	4	7
Modifier	109	5	4	0	3	9	14
Negation	73	4	3	0	2	4	10
Procedure	126	6	3	0	4	9	14
Time	185	9	5	0	6	13	17
Value	142	7	5	0	3	10	17

Table D.3: Descriptive statistics for the amount of relations extracted from the 30 reports of the *Ontology+Resources* run.

Relation	N	Mean	Standard Deviation	Minimum	Q ₁	Q ₂	Maximum
hasAdministrationRoute	69	2	3	0	0	0	13
hasDosage	57	2	2	0	1	3	7
hasDuration	1	0	0	0	0	5	1
hasEvolution	6	0	0	0	3	8	1
hasFinding	59	2	2	0	0	0	13
hasFrequency	101	3	3	0	0	3	9
hasIndication	23	1	1	0	0	4	3
hasLateralization	82	3	2	0	1	2	10
hasModifier	169	6	3	1	1	6	16
hasNegation	116	4	3	0	1	8	11
hasQuantity	78	3	3	0	0	0	11
hasSubstance	1	0	0	0	0	2	1
hasTarget	71	2	2	0	0	4	8
hasTime	156	5	4	0	1	4	15
hasValue	91	3	3	0	3	7	16
includes	3	0	0	0	0	0	1

Table D.4: Descriptive statistics for the amount of relations extracted from the 20 reports of the *Resources* run.

Relation	N	Mean	Standard Deviation	Minimum	Q ₁	Q ₂	Maximum
hasAdministrationRoute	35	2	3	0	0	0	8
hasDosage	35	2	2	0	0	1	6
hasDuration	1	0	0	0	0	5	1
hasEvolution	2	0	0	0	2	5	1
hasFinding	15	1	1	0	0	0	3
hasFrequency	58	3	3	0	0	3	9
hasIndication	12	1	1	0	0	2	4
hasLateralization	34	2	1	0	0	1	5
hasModifier	64	3	2	0	1	3	8
hasNegation	46	2	2	0	1	2	7
hasQuantity	27	1	2	0	0	0	6
hasSubstance	1	0	0	0	0	1	1
hasTarget	12	1	1	0	0	3	2
hasTime	80	4	3	0	1	2	12
hasValue	33	2	1	0	1	5	3
includes	4	0	0	0	0	0	1

Appendix	E
----------	----------

ICF Summary

MedInX ICF Summary

Episode 5007459

Impairments of Body Functions

List of Body Functions	PDL text	Qualifiers	Time	Segment
b330 fluency and rhythm of speech functions	voz arrastada	---	---	História Clínica
b110 consciousness functions	consciente	---	entrada	Exame Físico
b114 orientation functions	orientado	---	entrada	Exame Físico
b130 energy and drive functions	colaborante	pouco	entrada	Exame Físico
b215 functions of structures adjoining the eye	pupilas isocóricas e reactivas	---	---	Exame Físico
b320 articulation functions	disartrico	---	---	Exame Físico
b750 motor reflex functions	lateralização motora	sem	---	Exame Físico
b420 blood pressure functions	corado	---	---	Exame Físico
b545 water, mineral and electrolyte balance functions	hidratado	---	---	Exame Físico
b550 thermoregulatory functions	apirético	---	---	Exame Físico
b440 respiration functions	eupneico	---	---	Exame Físico
b410 heart functions	AC	arritmica	---	Exame Físico
b440 respiration functions	AP	sem alteracoes	---	Exame Físico
b110 consciousness functions	consciente	---	data de alta	Evolução
b114 orientation functions	orientado	---	data de alta	Evolução
b130 energy and drive functions	colaborante	---	data de alta	Evolução
b320 articulation functions	disartrico	---	---	Evolução
b420 blood pressure functions	corado	---	---	Evolução
b545 water, mineral and electrolyte balance functions	hidratado	---	---	Evolução
b550 thermoregulatory functions	apirético	---	---	Evolução
b420 blood pressure functions	normotenso	---	---	Evolução

Impairments of Body Structures

List of Body Structures	PDL text	Qualifiers	Time	Segment
s320 structure of mouth	desvio da comissura labial [direita]	ligeiro	entrada	Exame Físico
s1 structures of the nervous system	edemas [periféricos]	sem	entrada	Exame Físico
s320 structure of mouth	desvio da comissura labial [direita]	ligeiro	data de alta	Evolução
s1 structures of the nervous system	edemas [periféricos]	sem	data de alta	Evolução

Impairments of Activities and Participation

List of Activities and Participation	PDL text	Qualifiers	Time	Segment
d5 self-care	autonomo	---	---	História Clínica
d510 washing oneself	independente na sua higiene	---	---	História Clínica

Impairments of Environmental Functions

List of Environmental Functions	PDL text	Qualifiers	Time	Segment
e580 health services, systems and policies	s.u.	---	---	História Clínica
e110 products or substances for personal consumption	aas	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	enoxaparina	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	ranitidina	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	amiodarona	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	furosemido	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	nitratos transdermicos	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	carvedilol	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	digoxina	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	enalapril	---	---	Terapêutica Efetuada
e110 products or substances for personal consumption	amoxicilina	---	---	Terapêutica Efetuada
e355 health professionals	terapeuta da fala	---	---	Evolução
e580 health services, systems and policies	consulta externa	---	---	Evolução
e155 design, construction and building products and technology of buildings for private use	domicilio	---	---	Destino
e110 products or substances for personal consumption	ticlopidina	excepto	---	Destino
e110 products or substances for personal consumption	clopidogrel	---	---	Destino
e110 products or substances for personal consumption	amoxicilina	---	---	Destino
e110 products or substances for personal consumption	ac	---	---	Destino
e580 health services, systems and policies	consulta externa de terapia da fala	---	---	Destino