INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Extraction and Classification of Named Entities

Diogo Correia de Oliveira

Dissertation for obtaining the Master Degree in

Information Systems and Computer Engineering

Jury

President:	Professora Doutora Maria dos Remédios Vaz Pereira Lopes Cravo
Advisor:	Professor Doutor Nuno João Neves Mamede
Co-advisor:	Professor Doutor Jorge Manuel Baptista
Evaluation jury:	Professor Doutor Bruno Emanuel da Graça Martins

November 2010

Acknowledgements

This dissertation would not have been possible to do without the support of many people. So, first and foremost, I would like to thank my advisor, Professor Nuno Mamede, for all his prompt responses and infinite patience, and my co-advisor, Professor Jorge Baptista, for all his suggestions and knowledge of linguistics, which has always been a key factor in this kind of work. I would also like to thank Vera Cabarrão for having provided an annotated corpus from which it was possible to evaluate this work.

Secondly, I would like to thank all my friends and colleagues that have given me their support over the last six years, either related to this dissertation or not. A special thanks goes out to Bruno Almeida, Francisco Almeida, José Boavida, Filipe Cabecinhas, Filipa Coelho de Sousa, Renato Crisóstomo, Cláudio Diniz, Filipe Ginja, Sérgio Gomes, Carlos Jacinto, João Lemos, Nuno Lopes, Tânia Marques, Rui Martins, João Neves, André Nogueira, João Reis, Marcelo Rolo, João Sales, Daniel Santos, Andreia Silva and Artur Ventura.

Last but not least, I would like to thank my family and my girlfriend, Mariana, for never doubting me even when I doubted myself, and for always supporting me since day one. This dissertation is their work as well.

Lisbon, October 18^{th} 2010

Diogo Correia de Oliveira

Resumo

O Reconhecimento de Entidades Mencionadas (REM) consiste na delimitação precisa e na correcta classificação de expressões linguísticas de natureza variada e com uma forte componente referencial, tais como os nomes de pessoas, locais, organizações, etc., bem como de expressões numéricas e temporais. É uma tarefa-chave na área interdisciplinar do Processamento de Língua Natural (PLN), que mobiliza, por um lado, diversas competências de Engenharia de Sistemas e Computação, e, por outro lado, recorre a conhecimentos de vários ramos da Linguística, e que pode ser enquadrada no domínio mais vasto da Recuperação e Extracção de Informação (IR/IE).

A tarefa de REM tem um importantíssimo papel no desempenho de diferentes módulos de sistemas de PLN (por exemplo, no processamento sintáctico e semântico, em resolução de relações anafóricas e de correferência) mas também como um dos componentes de diversas aplicações do processamento da linguagem (reconhecimento da fala, sumarização ou indexação automática de documentos, tradução automática, entre outras).

Este estudo teve como objectivo central melhorar a performance do módulo de REM do sistema de PLN desenvolvido pelo L^2F/INESC-ID Lisboa (em parceria com a XEROX), relativamente ao desempenho que este teve na campanha de avaliação conjunta do Segundo HAREM (2008), em particular para as categorias `HUMANO`, `LOCAL` e `VALOR`.

Para tal, procedeu-se ao estudo comparativo dos sistemas de REM actualmente existentes para o Português, tendo sido proposto um novo conjunto de directivas de delimitação e classificação, para substituir as da campanha de 2008. Foram introduzidas várias melhorias em diferentes componentes da cadeia de processamento, em particular no analisador sintáctico XIP, responsável a jusante da cadeia pela extracção das entidades mencionadas.

Finalmente, o desempenho do sistema foi avaliado, verificando-se uma melhoria significativa dos resultados.

Abstract

Named Entity Recognition (NER) consists in the precise delimitation and the correct classification of linguistic expressions of a very diverse nature and with a strong referential component, such as names of people, places, and organizations, as well as different types of numeric and temporal expressions. It is an interdisciplinary task, key to Natural Language Processing (NLP), that mobilizes, on the one hand, several skills from the Systems and Computer Engineering domain, and, on the other hand much knowledge from different branches of Linguistics. This task can also be framed in the larger domain of Information Retrieval (IR) and Extraction (IE).

NER has a significant role in the performance of several modules of NLP systems (for example in syntactic parsing and semantic analysis, in anaphora resolution and coreference processing) but also as a key component of many NLP applications (such as speech processing, both in recognition and in synthesis, automatic summarization and document indexation, machine translation, among others).

The central goal of this study consisted in the improvement of the NER module of the NLP system developed at L^2F/INESC-ID Lisboa (in partnership with XEROX). In particular, it aims at improving the performance attained during the Second HAREM joint evaluation campaign (2008), especially for the `HUMAN`, `LOCATION` and `AMOUNT` categories.

To this end, a comparative study of existing Portuguese NER systems was carried out and a new set of delimitation and classification directives has been proposed to replace those used in the 2008 campaign. Several improvements were introduced in the NLP chain, specially in the XIP syntactic parser, the last module of the chain, which is responsible for named entity extraction.

Finally, the system performance has been evaluated, and a general trend of improvement has been confirmed.

Palavras-Chave

Keywords

Palavras-Chave

Reconhecimento de Entidades Mencionadas

Análise sintáctica superficial (“chunking”)

Léxico

Gramáticas locais

Metonímia

Keywords

Named Entity Recognition

Chunking

Lexicon

Local Grammars

Metonymy

Table of Contents

Acknowledgements i

Resumo iii

Abstract v

Palavras-Chave / Keywords vii

List of Figures xiii

List of Tables xv

List of Acronyms xvii

List of Terms xix

1 Introduction 1

1.1 Context . . . 1
1.2 Goals . . . 2
1.3 Thesis Structure . . . 3

2 State of the Art 5

2.1 Context . . . 5
2.2 The CaGE system . . . 6
2.2.1 Overview . . . 6
2.2.2 Functionality . . . 6
2.2.3 Results . . . 7
2.3 The PorTexTO system . . . 7
2.3.1 Overview . . . 7
2.3.2 Functionality . . . 8
2.3.3 Results . . . 9
2.4 The Priberam System . . . 9
2.4.1 Overview . . . 9
2.4.2 Functionality . . . 10

2.4.3 Results . . . 11
2.5 The R3M system . . . 11
2.5.1 Overview . . . 11
2.5.2 Functionality . . . 12
2.5.3 Results . . . 13
2.6 The REMBRANDT system . . . 14
2.6.1 Overview . . . 14
2.6.2 Functionality . . . 14
2.6.3 Results . . . 15
2.7 The REMMA system . . . 16
2.7.1 Overview . . . 16
2.7.2 Functionality . . . 16
2.7.3 Results . . . 18
2.8 The SEI-Geo system . . . 18
2.8.1 Overview . . . 18
2.8.2 Functionality . . . 18
2.8.3 Results . . . 19
2.9 The XIP system . . . 20
2.10 Comparison . . . 20

3 Architecture **25**
3.1 Processing chain . . . 25
3.1.1 Pre-processing . . . 25
3.1.2 Disambiguation . . . 27
3.1.3 Syntactic analysis . . . 29
3.2 The XIP system in the processing chain . . . 29
3.2.1 Chunks and dependencies . . . 30
3.2.2 Custom lexicons, local grammars and disambiguation rules . . . 34
3.3 Improvements . . . 36
3.3.1 Segmentation . . . 36
3.3.2 Consistency . . . 38
3.3.3 Classification Directives . . . 38
3.3.4 AMOUNT category . . . 40
3.3.5 HUMAN category . . . 45
3.3.6 LOCATION category . . . 49
3.3.7 Metonymy . . . 50

4 Evaluation **57**
4.1 Context . . . 57
4.2 Evaluation metrics . . . 58

4.2.1 Golden Collection . . . 58
4.2.2 Cornerstones . . . 58
4.2.3 Evaluation scenarios . . . 60
4.3 Evaluation Results . . . 61
4.3.1 Scenarios without metonymy . . . 62
4.3.2 Scenarios with metonymy . . . 65

5 Conclusions **69**
5.1 Final remarks . . . 69
5.2 Future work . . . 70

Bibliography **73**

A Classification results **79**

B POS categories **81**

C Classification Directives **83**
C.1 The AMOUNT category . . . 83
C.1.1 Delimitation . . . 83
C.1.2 AMOUNT types . . . 84
C.2 The HUMAN category . . . 87
C.2.1 INDIVIDUAL type . . . 88
C.2.2 COLLECTIVE type . . . 90
C.3 The LOCATION category . . . 92
C.3.1 Delimitation . . . 92
C.3.2 LOCATION types . . . 92
C.4 Metonymy . . . 96
C.4.1 Context . . . 96
C.4.2 LOCATION to HUMAN shift . . . 97
C.4.3 HUMAN COLLECTIVE to HUMAN INDIVIDUAL shift . . . 97
C.4.4 HUMAN COLLECTIVE to LOCATION shift . . . 98

List of Figures

2.1 PorTexTO: the architecture of the Annotator module (from Craveiro *et al.* [15]). 8
2.2 R3M: The architecture of the system (from Mota [34]). 12
2.3 REMBRANDT: Wikipedia plays an essential role (from Cardoso [9]). 14
2.4 REMMA: The system's architecture (from Ferreira *et al.* [17]). 16
2.5 SEI-Geo: The system's architecture (from Chaves [11]). 19

3.1 XIP: The processing chain in which the system resides (from Romão [43, Section 3, Figure 3.1]). 25
3.2 XIP: output tree after applying the chunking rules. 32
3.3 XIP: output tree for a complex proper name. 46

4.1 Results: chart from Relaxed ALT, identification, without metonymy. 66
4.2 Results: chart from Relaxed ALT, classification, without metonymy. 67
4.3 Results: chart from Relaxed ALT, identification, with metonymy. 67
4.4 Results: chart from Relaxed ALT, classification, with metonymy. 67

List of Tables

2.1 State of the Art: Comparison (Systems and entities). 21
2.2 State of the Art: Comparison (Systems and Technologies). 22
2.3 State of the Art: Comparison (the global results, regarding the identification and classification tasks). 22

3.1 Processing chain: POS tags (fields and categories). 27
3.2 XIP: examples of features. 30
3.3 Classification directives: differences between the two sets of directives. 39
3.4 Metonymy: list of examples the system is able to handle (all shifts). 55

4.1 Results: evaluation without metonymy, identification task (C. Id: correctly identified; P: precision; R: recall; F: F-measure). 62
4.2 Results: evaluation without metonymy, classification task (Max P: maximum precision; Max R: maximum recal; Max F: maximum F-measure). 64
4.3 Results: evaluation with metonymy, identification task (C. Id: correctly identified; P: precision; R: recall; F: F-measure). 65
4.4 Results: evaluation with metonymy, classification task (Max P: maximum precision; Max R: maximum recall; Max F: maximum F-measure). 66

A.1 State of the Art: Comparison (results from the classification task, strict ALT. P: Precision; R: Recall; F: F-measure). 80

B.1 XIP: list of POS categories (from Mamede *et al.* [28]). 81

List of Acronyms

Acronym	Designation in English	Designation in Portuguese
CaGE	**Ca**pturing **G**eographic **E**ntities	
GC	**G**olden **C**ollection	Colecção Dourada
GKB	**G**eographic **K**nowledge **B**ase	Base de Conhecimento Geográfico
HAREM	Named Entities Recognition Evaluation	**A**valiação do **R**econhecimento de **E**ntidades **M**encionadas
HMM(s)	**H**idden **M**arkov **M**odel(s)	Modelo(s) oculto(s) de Markov
HTTP	**H**yper**t**ext **T**ransfer/**T**ransport **P**rotocol	
ID(R)	**I**mmediate **D**ependency (**R**ules)	(Regras de) Dependência Imediata
IE	**I**nformation **E**xtraction	Extracção de Informação
IMDB	**I**nternet **M**ovie **D**ata**b**ase	
IP	**I**nternet **P**rotocol	
IR	**I**nformation **R**etrieval	Recuperação de Informação
L^2F	Spoken Language Systems Laboratory	**L**aboratório de Sistemas de **L**íngua **F**alada
LP(R)	**L**inear **P**recedence (**R**ules)	(Regras de) Precedência Linear
MARv	**M**orphosyntactic **A**mbiguity **R**esol**v**er	
MT	**M**achine **T**ranslation	Tradução automática
NE(s)	**N**amed **E**ntity(ies)	Entidade(s) Mencionada(s)
NER	**N**amed **E**ntities **R**ecognition	Reconhecimento de Entidades Mencionadas
NLP	**N**atural **L**anguage **P**rocessing	Processamento de Língua Natural
NP	**N**oun **P**hrase	Sintagma nominal
PorTexTO	**Por**tuguese **T**emporal **Ex**pressions **To**ol	
POS	**P**art **o**f **S**peech	
PP	**P**repositional **P**hrase	Sintagma preposicional

QA	**Q**uestion **A**nswering	Sistemas de Pergunta e Resposta
REMBRANDT	Named Entities Recognition Based on Relations and Detailed Text Analysis	**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **An**álise **D**etalhada do **T**exto
REMMA	MedAlert's Named Entities Recognition	**R**econhecimento de **E**ntidades **M**encionadas do **M**ed**A**lert
RuDriCo	**Ru**le-**Dri**ven **Co**nverter	
SASKIA	**S**PARQL **A**PI **S**ervice for **K**nowledge and **I**nformation **A**ccess	
SEI-Geo	Extraction, Annotation and Integration System for Geographic Knowledge	**S**istema de **E**xtracção, Anotação e **I**ntegração de Conhecimento **Geo**gráfico
TRE	**T**ree **R**egular **E**xpression	Expressão Regular de Árvore
XIP	**X**erox **I**ncremental **P**arser	
XML	E**x**tensible **M**arkup **L**anguage	
YACC	**Y**et **A**nother **C**ompiler **C**ompiler	

List of Terms

Term	Meaning
Anthroponym	The name of a person (e.g. “John”, “Sophie”).
Corpus	A collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.
F-measure	An evaluation measure that combines *Precision* and *Recall*.
Gentilic	Designation for a resident of a locality, which is derived from the name of that particular locality (e.g. “lisboeta”, “nova-iorquino”); also the designation for the nationality of a person (e.g. “Portuguese”, “Norwegian”).
Hydronym	The name of a body of water (e.g. “Atlantic”, “Pacific”).
Metonymy	A figure of speech that designates the substitution of a noun for another noun, usually the two having a part-whole relation between them (e.g. “suit” for “business executive” or “tracks” for “horse races”).
Oronym	The name of landform, such as valleys, mountains, hills, ridges or plateaus (e.g. “Mount Everest”).
Precision	An evaluation measure that considers the proportion of correct answers provided by a system over the set of answers *given* by the same system.
Recall	An evaluation measure that considers the proportion of correct answers provided by a system over the set of *all possible correct* answers.
Regex(p)	A Regular Expression (pl. Regexps): also called a pattern, a regexp is an expression that describes a set of strings. They are used to give a concise description of a set, without having to list all its elements. For example, `[0-9]` means “any digit from 0 to 9” (accepted input: `0`, `1`, . . . , `9`, etc), and `[0-9]+` means “any digit from 0 to 9, repeated one or more times” (accepted input: `0`, `123`, `275698`, etc).
Toponym	A general description for a place name (e.g. “Lisbon”, “Portugal”).

Chapter 1

Introduction

1.1 Context

BEING able to program a computer to fully understand our language has been an unfulfilled dream for many years in the scientific community. The computer science field of Natural Language Processing (NLP) is vast and is concerned with how computers interact with human beings, in particular by means of human (natural) languages. Named Entity Recognition (NER) is the NLP task that focuses on locating and classifying entities in a given text. For example, consider the following sentence:

> "James Cameron's latest movie premiered in December 2009 in the USA and its budget exceeded 200 million dollars."

A system capable of performing NER should be able to *identify* four named entities (NEs) in this example: (a) *James Cameron*; (b) *December 2009*; (c) *USA*, and (d) *200 million dollars*. Furthermore, it should be able to *classify* them according to established criteria. A possible scenario is to classify *James Cameron* as `HUMAN`, *December 2009* as `TIME`, *USA* as `LOCATION` and *200 million dollars* as `AMOUNT`. These categories are often subcategorized; in this case, the most obvious subcategories would be `PERSON`, `DATE`, `COUNTRY` and `CURRENCY`, respectively.

Due to the ambiguity that pervades natural language, the classification directives must be precise and unambiguous, which is not always easy to achieve. It is part of the objectives of this dissertation to present a set of directives for the classification of some NER categories in Portuguese texts.

NER is important because it is one of the first steps towards extracting meaning out of texts, which substantially helps several other NLP tasks, such as automatic summarization, information retrieval (IR), machine translation (MT), question answering (QA) and speech recognition, just to name a few.

The most common approaches to NER are grammar-based techniques and statistical models. Typically, systems that use the former obtain better precision but at the cost of a lower recall[1]. Moreover, these systems are often handcrafted by a team of computer engineers and computational linguists for a long time, in what is a slow, costly and time-consuming process. On the other hand, statistical NER

[1]The definition of "recall", as well as "precision" and "F-measure" can be seen in the List of Terms section, on page xix.

systems usually require a large amount of manually annotated training data. The task of annotating corpora is also a costly and time-consuming process. Unfortunately, even the best NER systems are fragile, because often they may have been built for one specific domain, so they typically do not perform well when transposed to other domains (Poibeau & Kosseim [41]).

Because of this trade-off, semi-supervised machine learning techniques have been put to use (Mota [34]), but results indicate that as costly and as time-consuming as they may be, more traditional methods such as using manual rules and lexicons are still absolutely necessary, especially to guarantee a higher recall.

The overall results of the evaluation of this work were satisfactory. These can be consulted in Chapter 4. Even if the results can not be directly compared with those from the Second HAREM, since the directives are different, it is possible to say that the main objective of this thesis has been achieved: results seem to show a general trend of improvement.

1.2 Goals

This thesis continues the development of a NER system, created in 2007 as a collaboration between the L^2F (Laboratório de Sistemas de Língua Falada, INESC-ID, Lisboa, Portugal) and the XRCE (Xerox Research Centre Europe, Grenoble, France).

This system already identified and classified NEs from several different categories, such as `AMOUNT`, `LOCATION`, `EVENT`, `PERSON` and `ORGANIZATION`. In 2008, it took part in the Second HAREM evaluation campaign (Carvalho *et al.* [10]), having obtained encouraging results: it was the third best system in terms of F-measure on both the identification and classification tasks, and the best one regarding the classification and normalization of temporal expressions. However, according to Hagège *et al.* [19], the recall had been lower than expected and could have been much improved, because there had been little time to add more lexical entries. Moreover, the system needed new rules and the existing ones needed further development.

Therefore, the main goals of this thesis are:

- To improve the system by adding more lexical entries, by correcting rules and by adding new rules, specifically in the `AMOUNT`, `HUMAN` and `LOCATION` categories;
- To establish a new way of presenting metonymy (see Section 3.3.7) and also to improve the capacity of the system for capturing these kinds of entities;
- To contribute to the creation of a new set of directives for Portuguese texts (see Appendix C), thereby replacing the ones that were used in the Second HAREM evaluation campaign;
- To evaluate the work by using evaluation metrics such as precision and recall.

1.3 Thesis Structure

The remainder of this document is structured as follows:

- Chapter 2 presents a comparative study of eight systems that took part in the second NER Portuguese evaluation campaign held in 2008;
- Chapter 3 describes one of these systems in more detail, presenting its main characteristics in Sections 3.1 and 3.2. Section 3.3 presents the improvements that have been introduced in the system during this study;
- Chapter 4 presents all data related to the evaluation of this work;
- Chapter 5 presents the conclusions.

Chapter 2

State of the Art

2.1 Context

The HAREM evaluation campaign brought together some systems that are involved in the NER task regarding the Portuguese language. Their common goal is to correctly identify and classify entities in any given text. Typically, these texts are gathered from newspapers, web articles and many other sources and are manually marked. In this campaign, the compilation resulted in 1.040 documents (15.737 paragraphs, 670.610 words). The Golden Collection (GC) is an annotated subset of these documents, which is used for the participants for evaluation. It follows the annotation guidelines of HAREM and is comprised of 129 documents (2.274 paragraphs, 147.991 words), approximately 12% of the original set of texts. The annotation process is usually difficult due to differences of opinion among the people who work on it. According to Carvalho *et al.* [10, Section 1.4.2, Table 1.1], there were 121 cases of annotations that raised doubts, as well as 14 cases of disagreement on the category of an entity, out of a total of 7.836 NEs.

However, not all participants were interested in the same categories. For instance, on the one hand, the CaGE system (which will be analyzed in Section 2.2), dealt only with entity recognition and classification in `PERSON`, `TIME`, `ORGANIZATION` and `LOCATION` categories (Martins [29]). The PorTexTO system (Section 2.3), on the other hand, dealt only with `TIME` category (Craveiro *et al.* [15]). Therefore, one of the general problems to be addressed is the evaluation metrics: if one wants to compare these two systems, what metrics should be used? Should they only be compared regarding the `TIME` category, since it is the one they have in common, or should they be compared using other methods?

The main purpose of this chapter is to provide a detailed view of the systems that took part in the campaign and to compare them. This comparison is not obvious because one needs to define metrics that can be coherently used in order to compare systems that are interested in capturing different things. Section 2.10 presents this comparison.

2.2 The CaGE system

2.2.1 Overview

The CaGE system deals with the recognition and disambiguation of geographic NEs as a means to map them to actual geographic information, such as latitude and longitude coordinates. Having this kind of information available may prove useful in the area of IR, especially if one is interested in singling out data according to its geographic characteristics (Martins [30]).

2.2.2 Functionality

In order to fully identify and disambiguate geographic entities, the CaGE system relies on external information resources. In particular, it uses dictionaries that help improving the identification task, and a geographic almanac in which every reference is assigned to an unique identifier; this almanac improves the disambiguation task.

The dictionaries were built with the help of lexical resources, such as: (a) names of people listed on IMDB; (b) lists of time periods and common first names extracted from Wikipedia, and (c) a list of names belonging to the geographic almanac that was used in the DIGMAP project (Borbinha *et al.* [8]). More can be found in Martins [29].

For `LOCATION` entities, the system also uses the so-called exceptions dictionary; the idea is that it should contain those entities that despite being geographic in nature, are known to be most commonly used in other senses.

The geographic almanac is responsible for mapping names to geographic concepts; naturally, this is a many-to-many relationship, because several names may correspond to several concepts. Moreover, the almanac also defines a geographic area, using latitude and longitude coordinates, as well as an inclusion hierarchy chain between concepts. In order to understand why this chain is important, one must understand the meaning of geographic scope.

According to Martins [30, Chapter 1, page 3]: "Scopes represent the most probable location to which a document's content might be concerned with. The idea is to disambiguate the different place-names in the text, afterwards combining them into an encompassing geographical region. We refer to this region as the geographic scope of the document."

The inclusion hierarchy chain is important to the extent that in the final stage of processing, the system assigns a geographic scope to the document as a whole. This assignment is made through a combination of all the geographic references found in the text, which is put in practice by an algorithm that uses an inclusion hierarchy chain.

The whole processing can be divided into four steps, which are summarized below. For a more detailed explanation, consult Martins [29], Martins *et al.* [31].

1. Identification of NEs: the text is broken down into atoms. The system will ignore any named entity (NE) whose length is over six words;

2. Classification and partial disambiguation of NEs: for all entities that have multiple mappings, the system uses disambiguation rules in order to determine the category and type of the entity;

3. Complete disambiguation of geographic and temporal entities: `LOCATION` entities are searched for in the DIGMAP almanac in order to find their geographical concept;

4. Assignment of geographic and temporal scopes to the documents: the system assigns both a geographic scope and a temporal scope to the whole document.

2.2.3 Results

One could say that this system does not share the same criteria and goals with those of the HAREM evaluation campaign. The latter only deals with the identification and classification of NEs but does not tackle the complete disambiguation problem that arises from expressions that deal with `TIME` or `LOCATION`.

However, in reality, processing a geographic reference is similar to identifying an entity that describes a `LOCATION`. Therefore, the CaGE system was able to play a role as one of the participants, not only in the `LOCATION` category, but also in the `PERSON`, `ORGANIZATION` and `TIME` categories. The inclusion of these three improved the system because geographic references are often ambiguous regarding entities of other categories, for example, Paris Hilton the person and the Hilton Hotel in Paris.

Probably due to the differences between the goals of the CaGE system and those of the campaign, results were modest: it was the fifth best system (out of eight) both in the identification and classification scenarios, having reached a 0.4340 F-measure result in the former and 0.3419 in the latter. These results can be further consulted in Section 2.10 and Appendix A.

According to Martins [29], one of the most detrimental aspects is that the system does not try to determine the role of an entity in the text. For instance, consider the sentence "The UK announced that ...". According to the evaluation metrics of the campaign, in this case UK should be marked as a `PERSON` entity. However, the CaGE system marks UK as a `LOCATION` entity, because it does not try to determine the role of UK in the sentence.

2.3 The PorTexTO system

2.3.1 Overview

The PorTexTO system focuses on the identification of temporal expressions embedded in documents, which can later be used to sort the results provided by some IR system. The idea to use temporal data associated to documents appeared as an alternative to popular sorting methods, such as sorting by popularity. According to Alonso *et al.* [1], there are four application areas (surely among many others) that could benefit from using temporal information: ad-hoc retrieval, hit-list clustering, exploratory search and presentation.

2.3.2 Functionality

Unlike the system described in the previous section, PorTexTO was specifically created for the HAREM evaluation campaign. It was written from scratch and its main requirements were simplicity and low processing time (Craveiro *et al.* [15]). Before the main processing stages, the input text is separated into sentences by a module written in Perl[1]. This separation is an indication of how the system processes the text: not term by term, but sentence by sentence.

The system is comprised of two modules: the Annotator and the Co-occurrences Processor. Let us start by analyzing the former.

Figure 2.1 presents a general view of the architecture of the module. The processing is comprised of four stages. First, the module processes the documents sentence by sentence and for each one of these, it tries to determine if the sentence may contain temporal expressions; this is achieved by looking for numerical terms or for at least one temporal keyword among those defined in the Temporal Keywords file. All the sentences that do not contain at least one temporal expression are excluded and will not be processed.

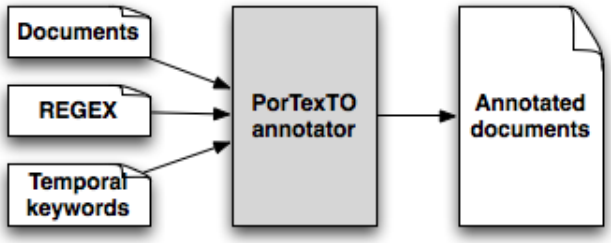

Figure 2.1: PorTexTO: the architecture of the Annotator module (from Craveiro *et al.* [15]).

Afterwards, for those sentences that were not excluded, the system applies rules to determine the existence of dates (complete or incomplete), hours, years, weekdays and months, abbreviated or not. Finally, at the end of this stage, the expressions within these sentences are marked as candidate expressions and will later be used by the REGEX component of the Annotator: a file whose contents are temporal patterns, that is, regular expressions (regexps).

The REGEX component is essential because it stores the temporal patterns that will be compared to the sentences obtained from the previous stages. A typical pattern may say: "match everything that starts with *on*, followed by a date". This simple pattern would match sentences such as "The ship's scientific mission ended *on April 30, 2002*".

Finally, the sentence is annotated according to the HAREM annotation rules, which can be found in Hagège *et al.* [18].

Let us now focus our attention on the Co-occurrences Processor module. Simply put, this module is responsible for creating the REGEX component that is used as input in the Annotator module. It is only executed when there are not enough patterns to continue a normal execution. The processing is done in five stages.

[1] http://www.perl.org/

First, the module builds a list of expressions considering N words before and/or after the Temporal Reference Word (TRW). For instance, consider the TRW "year". Some possible expressions to be included on the list are "Last year" or "The following year of 2010". Subsequently, the module aggregates the temporal expressions included on the list, so that expressions such as "Last year" and "Next year" become "Last|Next year" and then it sorts this list by descending order of occurrences.

The fourth stage, and the most difficult one according to the authors, consists of a manual analysis in order to exclude those expressions that despite having a temporal unit, are not really temporal expressions and do not make sense. These situations are very uncertain in nature and ultimately result in an inability to decide whether an expression is temporal. Nevertheless, this analysis is necessary because the common sense factor is hard to implement.

In the fifth and final stage, the module creates the regexps that define the temporal patterns used in the REGEX component of the Annotator module.

2.3.3 Results

There are some known limitations involving this system. For example, it does not correctly classify composed temporal expressions such as "on the *10th* day of last *month*", which should be classified as one single temporal expression. However, PorTexTO considers it to be two separate temporal expressions.

Despite the known limitations of the system and the lack of time to implement and test everything, it exceeded expectations and produced very encouraging results. Furthermore, it was able to annotate 675.000 words spread across 33.000 lines of text in about 3 minutes and 20 seconds, which is a very good indication of how fast the system processes data, thus meeting one of the main requirements that were established in the first place.

Globally, PorTexTO was the worst system (eighth position, out of eight) both in the identification and classification scenarios, having only reached a disappointing 0.1592 F-measure result in the former and 0.1562 in the latter. However, it is important to notice that this system focuses exclusively on the `TIME` category, and regarding it, PorTexTO obtained good results, having only been surpassed by XIP. For a more detailed description, consult Section 2.10 and Appendix A.

According to the authors, there are four improvements that could be implemented in the future. On the one hand, regarding temporal patterns, they should consider more than two words before/after the Temporal Reference Word and they should have more than one temporal unit. The process by which they are created should also be more automatized. On the other hand, the system should also support other languages, which is not a difficult task because the modules are language independent.

2.4 The Priberam System

2.4.1 Overview

For some time now, Priberam[2] has been developing a NER system that is built on their platform for linguistic development. This platform includes a set of linguistic resources (a lexicon, an ontology and

[2] http://www.priberam.pt

a grammar) and software tools that cover proofing, text processing and IR tools.

The system was not written specifically for the HAREM evaluation campaign and actually, it was already in use as an independent module in some Priberam products, mainly FliP[3], as well as an automatic answering system (Amaral *et al.* [4]) and also an IR system, specifically a search engine (Amaral *et al.* [2]).

This section begins by briefly describing the platform upon which the system was built, because its characteristics are very important in order to understand the latter. Subsequently, it describes the system in detail and concludes with a short and general view of the results it achieved.

2.4.2 Functionality

This system takes advantage of the linguistic resources that are part of the platform used by Priberam. First, the platform uses a lexicon with morphosyntactic and semantic classification. Each lexical unit may have several meanings (or "senses", as the authors call them), which are embedded within the lexicon.

Second, each and every entry of the lexicon maps to one or more levels of a multilingual ontology (Amaral *et al.* [5]), which is structured by conceptual proximity relations: this means that the ontology is organized in a way that things that might appear together are combined in the same category. Moreover, the ontology also considers semantic relations such as hyperonymy and hyponymy (for example, *color* is a hyperonym of *red* and *red* is a hyponym of *color*). Each entry in the ontology is a structure with six fields: Portuguese word, part of speech, sense index, ontological domain, English word and French word. This structure enables multilingual translations through the ontology domain. In order to better understand how this is done, consider the brief examples for the Portuguese word "porco", which can be further analyzed in Amaral *et al.* [3, Section 2.2, example 4]:

{porco, N, 1, 26.3.2.9, pig, cochon}	[animal]
{porco, N, 2, 28.1.2.7, pig, porceau}	[dirty person]
{porco, A, 1, 28.1.2.6, dirty, cochon}	[dirty]

Finally, the platform has a grammar that was built using a descriptive language. This language was created specifically to give linguists the chance to describe a grammar in a way that software might handle it. The language is closely related to the one used by YACC (Johnson [20]), that is, a grammar is described by writing its rules. It is out of the scope of this document to provide a detailed analysis of this grammar, but more information can be found in Amaral *et al.* [3, Section 2.3].

Having briefly described the main components of the platform upon which the Priberam system was built, let us now focus our attention on the system itself. In order to proceed to the NE identification, the system begins by simply inheriting the semantic and morphological values that exist in the lexicon. This approach is obvious in nature but naive: it is also important to analyze the context that surrounds the entity. This issue has already been analyzed in this document, as other systems also deal with it.

[3]FliP (http://www.flip.pt) is a set of tools for the Portuguese language and it includes two correctors (syntactic and orthographic), a dictionary of synonyms and a module that deals with hyphenation.

Consequently, the system uses contextual rules. These rules are well defined within the platform and they provide a lot more features than simply detecting NEs; for example, they also perform morphological disambiguation and detection of fixed expressions. Regarding NER, the system uses the contextual rules in order to add semantic and/or morphological values to individual units or to sequences of units. It tries to find a sequence of two or more proper nouns and recognize them as a single token, which will be classified according to the criteria established in the lexicon for each element of the sequence. For example, "Luís Vaz de Camões" will be classified as an anthroponym and "rio de São Domingos" as a hydronym. The latter is because the system detects a series of proper nouns following a common name such as "rio" (river), so it is able to infer that the NE is referring to a body of water.

A particular aspect that needs to be pointed out is that, according to the authors, the system is able to identify NEs whose elements are in a different language of that established in the lexicon, although the classification itself will be ignored in many cases if the context surrounding the NE is not good enough to provide a semantic value.

2.4.3 Results

This system participated in a full scenario, that is, it tried to identify and classify all NEs from all possible categories. The results were very positive: globally, it was the best system both in the identification and classification scenarios, having reached a 0.7110 F-measure result in the former and 0.5711 in the latter. It was also the best system in numerous other aspects, which can be consulted in Appendix A. However, it was not the best one in any scenario regarding precision. The system clearly has better results when identifying; it needs further development of the semantic classification of NEs.

There was a particular category, `TIME`, which had results really below what was expected. This happened because the rules of the campaign for detecting and classifying `TIME` entities are almost completely incompatible with the rules that exist in the system.

According to Amaral *et al.* [5], the system needs improvement to detect metonymy cases, as well as improvement in the rules that detect and classify entities of some categories (e.g. `EVENT`, `THING` and `ABSTRACTION`). Furthermore, they also believe that as long as the ontology keeps improving, the system itself will improve along with it.

2.5 The R3M system

2.5.1 Overview

The R3M (Mota [34]) is a NER system that was built in order to only identify and classify people, organizations and locations, due to time limitations.

The system was built in a flexible way, so that not only future categories could be easily added, but also recognition of relations between entities. The author decided to adopt a semi-supervised learning approach, using as little as possible any manually written linguistic resources, since they are usually time-consuming, or expensive to obtain.

This section begins by describing the architecture of the system in detail, covering its five main stages and their substages. Afterwards, it analyzes the contribution of the system to the campaign by commenting on the results. Finally, it presents possible improvements in the future.

2.5.2 Functionality

The R3M system is an improvement on an already developed system (Mota [35]), which in turn was inspired by the one suggested in Collins & Singer [13]. It has a modular and sequential architecture, which clearly distinguishes the identification and classification tasks. It is comprised of five main stages (represented by the black boxes in Figure 2.2), which occur in two distinct phases: training and testing (represented by the blue dashed boxes).

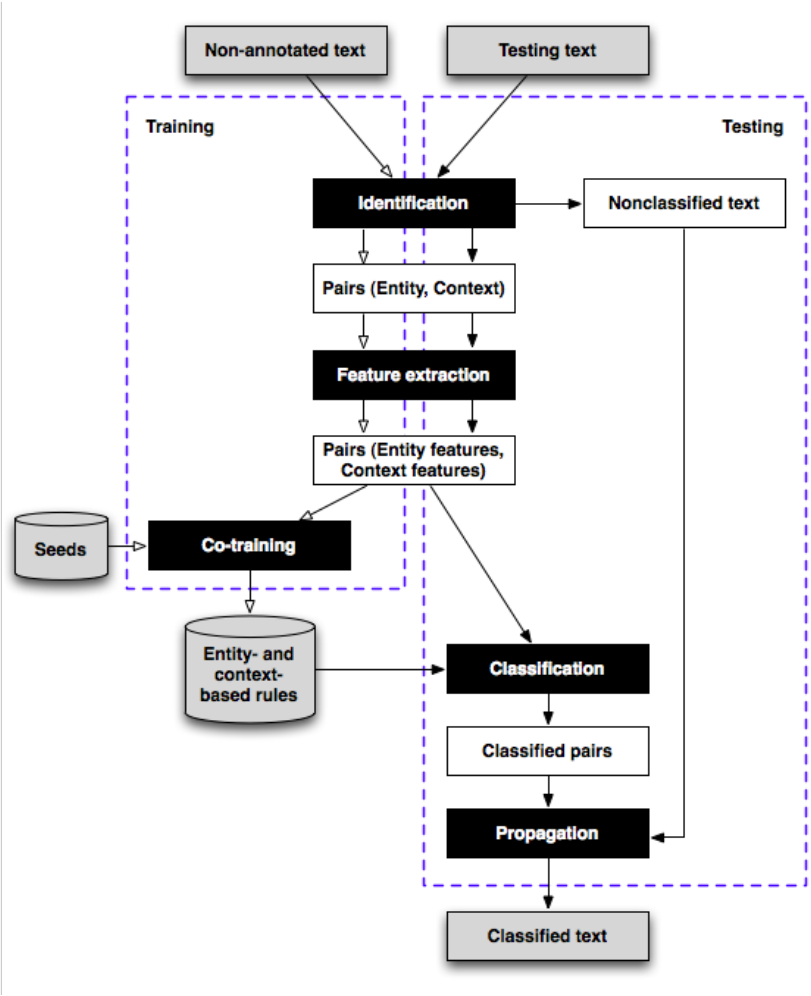

Figure 2.2: R3M: The architecture of the system (from Mota [34]).

The identification module is responsible for identifying candidate NEs and the context in which they appear. Due to these dual purposes, the module is comprised of two submodules to perform these tasks.

The identification submodule splits a sentence into atoms, and applies rules that either identify or eliminate a candidate. For example, the system has a rule that excludes any entity that represents a month of the year, because it is not interested in identifying `TIME` entities.

The second submodule identifies the context in which a candidate appears by matching a finite set of rules, such as: "if the candidate has a left context, then its left limit must be either an article, an empty word, a preposition or a two atom sequence separated by hyphen (-) or slash (/)" (Mota [see 34, page 185]).

Consider the following example: "The pictures greatly impressed the Prime-Minister". The module is able to infer that the entity is *Prime-Minister* and that its left context is *The pictures greatly impressed the*. Moreover, in cases like "The National Aeronautics and Space Administration (NASA) committee said that ...", the module is able to infer that the context is not "NASA", but what comes after that.

The feature extraction module receives a list of pairs `(entity,context)` as input and generates a new list of pairs with entity- and context-related features. These features include data such as the entity itself and its length, as well as the whole context and its type (left or right), among others.

The classification module receives the list generated by the feature extraction module and computes its classification through a co-training algorithm that can be consulted in Mota [34, page 188]. As far as this module is concerned, a classification rule is a triple (x, y, z) that represents the probability of observing the category y when the entity has the feature x. z is the precision of the rule. The entity is classified according to the rule with the highest precision among those that are relevant to it.

The co-training module receives the list of pairs generated by the feature extraction module and incrementally infers the classification rules mentioned in the previous paragraph, using a semi-supervised approach, according to the co-training algorithm. This essentially means that the module is able to learn new rules from nonclassified entity-context pairs. Moreover, these rules are used to reclassify the entity-context pairs.

The final major component of the R3M architecture is the propagation module. Essentially, it was built in order to increase the recall of the system. It detects entities that the classification module was not able to classify and it simply assigns the most frequent classification to the entity. As it does not use any other kind of information, the module inevitably decreases the precision of the system; it is however a trade-off that the author considered important enough to uphold.

2.5.3 Results

According to the author, the module that learns the classification rules had last minute problems and, due to lack of time, they were not solved. Therefore, the R3M system entered the campaign only in an entity identification scenario.

Globally, R3M was the second best system (out of eight) along with REMBRANDT, having reached a 0.6828 F-measure result in the identification task. More information on these results can be found in Section 2.10.

To conclude this analysis, Mota [34] believes that the system needs improvements in: (a) the creation of context restrictions; (b) detecting the context; (c) extending the seeds; (d) and finally, it would greatly benefit from a text selection module before the training phase. This module would be responsible for selecting relevant annotated sentences, instead of simply increasing the number of sentences in the training set, thus aiming at improving the result of the classifier.

2.6 The REMBRANDT system

2.6.1 Overview

The REMBRANDT[4] system performs both NER and detection of relations between entities, the latter being out of the scope of this document. It was designed to detect all different kinds of entities occurring in Portuguese texts and it uses Wikipedia[5] as a source for knowledge.

This section begins with a description of the architecture of REMBRANDT, especially focusing on its processing stages. In order to fully understand these, it is equally important to understand the architecture of SASKIA, hence this will also be an important aspect to consider. Finally, there will be a brief overview of the results in the HAREM evaluation campaign.

2.6.2 Functionality

Figure 2.3 depicts the process behind REMBRANDT. The communication between REMBRANDT and Wikipedia is made through an interface known as SASKIA. According to Cardoso [9], it simplifies navigation tasks in the categories' structure, links and redirections of Wikipedia, in order to extract knowledge.

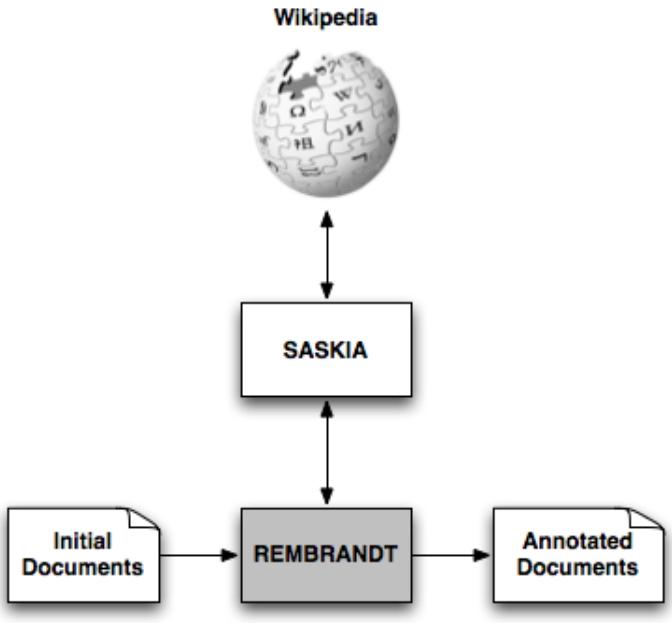

Figure 2.3: REMBRANDT: Wikipedia plays an essential role (from Cardoso [9]).

REMBRANDT supports several file formats, including plain text, XML and HTML, and is implemented in Java. The documents are processed in three stages: (a) recognition of numeric expressions and generation of NE candidates; (b) NE classification, and (c) repetition of stage (b) for NEs without classification.

[4]REMBRANDT can be used for free at http://xldb.di.fc.ul.pt/Rembrandt
[5]http://pt.wikipedia.org/wiki/

The first stage begins by dividing the text into smaller units, sentences, with the help of a module written in Perl (the same used in PorTeXTO, see Section 2.3.2). Then it identifies any numeric expression, such as isolated numbers (for example, 3, three and third), using a set of rules. These numbers will then help a second set of rules, responsible for the identification of temporal and amount expressions, such as "March 3" or "three Euros". Any sequence of words that has at least one capitalized letter and/or a number is a NE candidate.

REMBRANDT uses a dual strategy regarding NE classification. First, SASKIA classifies every NE candidate, but then they are reclassified, twice, using grammatical rules. These two approaches are intertwined and complement each other: SASKIA is able to classify an entity without losing track of what other meanings it might have (this is done with the help of the disambiguation pages of Wikipedia, as will be explained later on in this section), but at the same time the grammatical rules are able to consider other factors besides the entity itself, mainly the context surrounding it, which will in turn help SASKIA classify an entity according to its context.

Wikipedia often generates static HTML dumps[6] of every language, that is, a copy of all pages from all Wikipedia wikis, in HTML form. SASKIA was originally developed to work with the page file of Portuguese, which was around 1.4 Gigabytes in March 2008 (Cardoso [9]). This file was processed in a few hours.

The classification process done by SASKIA is divided in three steps. First, it associates a given entity to a Wikipedia page, trying to produce an exact match between the name of the entity and the title of the page. If there is such a page, it collects the categories of the page and then the association is finished; it can proceed to the next step. Otherwise, it will try to find the most connected page using hyperlinks.

Second, for each category found in the previous step, SASKIA analyzes its type and visits related pages, extracting more categories as it goes. This is done following a limited depth-first search approach, namely it will keep visiting related pages until it reaches four levels of depth. SASKIA is able to distinguish between several category types, but these are out of the scope of this document. For more information, consult Cardoso [9, page 201].

Finally, SASKIA applies a set of grammatical rules to the categories that were obtained in the previous step. Their goal is to extract a meaning and a geographic reference, in case it exists.

The grammatical rules represent sentence patterns that indicate whether an entity is present, and they can establish actions to execute when they are successfully applied. REMBRANDT applies these values to all sentences, sequentially and one at a time.

One particular aspect that should be noted is that all successful rules are immediately executed, which allows for newly created entities to be available at once. More information regarding these rules can be found in Cardoso [9, Section 11.4].

2.6.3 Results

REMBRANDT participated in a full scenario regarding both the identification and classification of NEs. Globally, it was the second best system (out of eight) both in the identification and classification scenar-

[6]These can be accessed by the general public at http://static.wikipedia.org/

ios, having reached a 0.6828 F-measure result in the former (the same result as R3M) and 0.5674 in the latter. For a more detailed comparison to the other systems, please refer to Section 2.10.

The system further distinguished itself in numerous other aspects, for example: it was the best system to classify AMOUNT entities, having reached the highest precision, recall and F-measure in that category; it obtained the highest F-measure of all systems in the classification of PERSON, WORK, LOCATION and ORGANIZATION entities; and it was the most precise system in the classification of PERSON and WORK entities. All these results can be consulted in Appendix A.

2.7 The REMMA system

2.7.1 Overview

REMMA is a NER system that, like REMBRANDT (see Section 2.6), uses Wikipedia as an external knowledge source. It was developed under the MedAlert project. According to Ferreira *et al.* [17], the objective of MedAlert is to use Information Extraction (IE) techniques on medical texts in order to automatically deduce irregularities or doubts that may arise from a particular decision made by a doctor, nurse or any health official.

This section of the document presents the architecture of REMMA, along with the methods it uses to identify and classify entities. Then, the results of the campaign are shown and briefly discussed.

2.7.2 Functionality

One of this system's main goals is to study how important is the use of an external source such as Wikipedia to the NER task and, more concretely, whether its use could produce better results in the campaign than those obtained by "traditional" methods, such as systems based on manually written rules and lists of words. Their implementation is time-consuming and they are even more difficult to maintain, especially the lists, because they frequently overlap each other.

Figure 2.4 presents an overview of REMMA's architecture. Reader, Pre-processing and Finalizer are all done using tools that are built in the platform upon which REMMA is implemented. Documents are read one by one and separated into sentences, which are then separated into atoms. All morphosyntactic categories are obtained with the TreeTagger analyzer (Schmid [46]), a tool for annotating text with part-of-speech and lemma information.

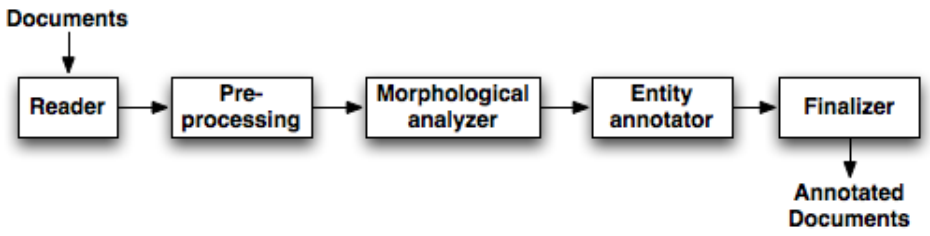

Figure 2.4: REMMA: The system's architecture (from Ferreira *et al.* [17]).

The Entity Annotator module is the more interesting one. REMMA uses a dual strategy regarding entity classification. One the one hand, it uses manually written rules and lists of words; on the other hand, it also classifies entities using information extracted from Wikipedia. The interesting part is that REMMA is able to use both strategies separately or together, thus providing an excellent means to compare results with or without Wikipedia's influence.

The annotation process is comprised of four stages. First, all sets of terms that begin with a capitalized letter are found and marked as NE candidates.

Second, the system combines a set of context rules with several lists of words in order to classify NE candidates. The lists had already been manually created and they span several topics, such as first names, Portuguese cities and diseases. The rules were also manually created and they are based on the context in which the expression is referred to, covering several semantic classes, such as locations, jobs, types of organizations, among others. This stage of processing includes several annotator modules that deal with the following entities: `PERSON`, `LOCATION`, `ORGANIZATION`, `EVENT`, `THING`, `ABSTRACTION` and `WORK`.

Third (optionally), REMMA classifies entities with Wikipedia's help. If this approach is used in conjunction with the previous one (using contextual rules), then the only entities that will be classified are the ones which have not been already classified. Otherwise, if used individually, it will classify all entities.

The method used to search for information in Wikipedia is based on the assumption that typically a Wikipedia article begins with a summary paragraph in which the entity's semantic category occurs. For example, if one searches for "Instituto Superior Técnico" in Wikipedia, the article begins with: "Instituto Superior Técnico is a Portuguese faculty of engineering and part of the Universidade Técnica de Lisboa. It is a public institution of university higher education ..."

REMMA begins by concatenating each entity's term with the underscore symbol, because that is how Wikipedia presents the URL for a given entity, and retrieves the corresponding article. Then, the article's introductory sentence is analyzed and REMMA tries to retrieve the entity's semantic category from it. In the previous example, the keyword *institution* is the one REMMA will use to classify Instituto Superior Técnico.

In the fourth and final stage, the system processes `VALUE` and `TIME` entities separately. It identifies sets of terms which contain at least one number or that belong to a predefined keyword list. For example, for `TIME` entities this list contains keywords such as *Easter*, *Spring*, *Summer* or weekdays; for `VALUE` entities, this list contains keywords such as *Kg*, *GB*, *Euros*, *Dollars*, among others.

A final aspect worth mentioning is that REMMA expands annotations already set to an entity in case it is preceded by a word beginning with a non-capitalized letter. Typically, this word must be in the campaign's directives. For example, consider the sentence: "When the ex-president José Sarney said that his greatest mission was to lead the country to elections", *ex-president José Sarney* would be classified as a `PERSON` entity (Ferreira *et al.* [17, page 223, example 12.10]).

2.7.3 Results

REMMA participated in the campaign with three runs: only with rule-based classification; only with Wikipedia-based classification, and both. The goal was to determine whether the results could be improved by using Wikipedia as an external knowledge source. It turns out that when used all by itself, Wikipedia provides the worst results of the three methods: the F-measure fell approximately 0.12 relative to the run where both methods were applied (this was actually the one that provided the best results).

Globally, REMMA was the fourth best system (out of eight) both in the identification and classification scenarios, having reached a 0.5515 F-measure result in the former and 0.4526 in the latter. It also distinguished itself for being the most precise system in the classification of `ABSTRACTION`, `ORGANIZATION` and `THING` categories.

To sum up, the results suggest that extracting semantic categories from Wikipedia is useful, even if the method is quite simple. However, using such methods without combining them with more "traditional" ones can be detrimental and should be avoided, unless it is possible to take better advantage of Wikipedia's internal structure, such as the redirect links, disambiguation pages and so forth, but that approach has not been applied in this system. Even so, REMMA is very precise, practically as precise as REMBRANDT (see Section 2.6.3), which is more important for medical reports than a high recall, since the latter could indicate "noisy" results. For more information about these results and a comparison with other systems, please refer to Section 2.10 and Appendix A.

2.8 The SEI-Geo system

2.8.1 Overview

SEI-Geo (Chaves [11]) is a NER system that only deals with the identification and classification of `LOCATION` entities and is essentially based on patterns and geographic ontologies (from this point on known as geo-ontologies).

This section describes SEI-Geo's architecture, mainly one of its essential modules: the extractor and annotator of geographic information, and how it is inserted in the Geographic Knowledge Base (GKB) (Chaves *et al.* [12]), as well as the results it obtained in the HAREM evaluation campaign.

2.8.2 Functionality

SEI-Geo is integrated in GKB, a geographic knowledge management system that acts as a repository for integrating this type of knowledge from multiple sources. It supports several languages and has tools for generating ontologies. A high-level observation of this system is as follows: on one side, GKB is able to extract geographic knowledge from structured knowledge sources; on the other side, it receives other geographic knowledge in the form of trees, which has already been manipulated by SEI-Geo (originally, this knowledge was not structured). GKB combines these two and using several tools, it is able to construct geo-ontologies. Figure 2.5 presents an overview of the extractor and annotator of geographic

information.

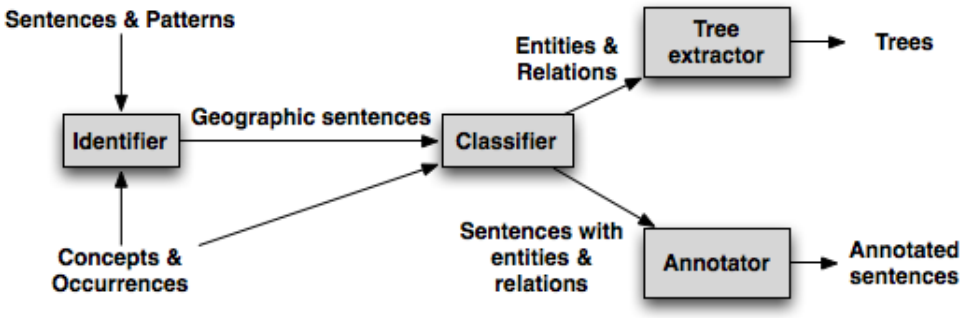

Figure 2.5: SEI-Geo: The system's architecture (from Chaves [11]).

Although the figure is self-explanatory, there are three things that should be noted. First, the classifier consults the geo-ontologies in order to perform the disambiguation and identify semantic relations; second, the tree extractor only constructs a tree with at least two NEs and one relation. There is not a predefined maximum number regarding these two. Finally, the annotator is capable of annotating texts in a format required by an application. In this case, sentences are annotated according to the campaign's rules.

SEI-Geo's geo-ontologies provide lists of names and concepts. They allow, among other things, the exploration of relations between entities and they provide a more detailed entity classification. SEI-Geo uses two geo-ontologies: Geo-Net-PT, Portugal's complete geo-ontology, which contains more than 400 thousand entities and is available online[7], and WGO, the World's Geographic Ontology, which contains names, concepts and relations, all related to countries, cities, oceans and mountains, among others.

2.8.3 Results

SEI-Geo participated with four runs: the first one only used Geo-Net-PT; the second one only used WGO; the third and fourth ones used both geo-ontologies and the difference between them has to do with the relation recognition task, which is out of the scope of this document.

The first run obtained the worst results: by only using Geo-Net-PT, the system significantly lowered its precision and recall. Runs 3 and 4 were the best ones and provided above average results regarding precision, recall and F-measure. SEI-Geo's most prominent aspect is its precision: it obtained the best result of all systems in 4 out of 5 scenarios, with values ranging from 0.86 to 0.91, in both identification and classification tasks. However, the system suffers from a low recall, mainly due to its simplicity – there is no syntactic analysis and the set of patterns is very limited.

Globally, SEI-Geo was the sixth best system (out of eight) both in the identification and classification scenarios, having reached a 0.2359 F-measure result in the former and 0.2017 in the latter. All these results can be further consulted in Section 2.10 and Appendix A.

[7] http://xldb.fc.ul.pt/wiki/Geo-Net-PT_02

2.9 The XIP system

Unlike other sections so far, this section of the document only provides a brief description of the XIP system, which is needed to contextualize the following section, that is, the comparison between systems. Section 3 provides a more complete description of XIP and also the improvements that have been made to it in the course of this thesis.

XIP is a rule compiler that allows text parsing both at syntactic and semantic levels and it is used by the system to identify and classify entities. Although "the XIP system" is commonly used throughout this document, the system itself is not (only) XIP. In reality, XIP is a tool that is used in the final stage of a processing chain, as depicted in Section 3, Figure 3.1.

The chain is comprised of three stages: (a) pre-processing; (b) disambiguation (rule-driven and statistical), and (c) syntactic analysis. There are also converter modules between each stage. All existing rules are manually written and the system uses lexicons, lists of words and grammars as external knowledge sources.

While building this system, the authors aimed at identifying and classifying all NEs, except `THING` and `ABSTRACTION`. Also, the authors invested a lot of time and effort in the development of the `TIME` category, which proved to be worthwhile since XIP obtained the best results in it, considering the three usual evaluation metrics: precision, recall and F-measure. XIP was also the most precise system in the `EVENT` category, having also obtained the highest F-measure in it. Globally, it was the third best system both in the identification and classification scenarios, having reached a 0.6121 F-measure result in the former and 0.5445 in the latter. All these results can be further consulted in the next section and in Appendix A.

2.10 Comparison

The systems that have been analyzed so far are the ones that participated in the Second HAREM evaluation campaign, which took place in 2008. The ones that participated in the First HAREM are not part of this document since Romão [43] and Loureiro [25] already provide a comparison between them.

As previously stated, it is not easy to introduce metrics in order to compare systems that are interested in different things. In this comparison, three criteria were considered: (a) the different kinds of entities the systems deal with; (b) the technology they use in order to identify and classify entities, and (c) the results they present.

Table 2.1 presents a list of the systems and of the entities they consider, in what concerns the identification and classification tasks. Entries marked with a check mark (✓) represent entities both identified and classified by a particular system, whereas an asterisk (*) represents entities solely identified. Empty entries represent entities not considered.

Notice the diversity: Priberam, REMBRANDT and REMMA try to identify and classify all entities, whereas PorTexTO and SEI-Geo only deal with `TIME` and `LOCATION` categories, respectively. Moreover, there are cases of systems, such as R3M, that perform entity identification on a set of entities and classification on a different one. In this case, R3M did not classify anything due to last minute problems

System	Identification and Classification tasks								
	Abstr.	Amount	Event	Location	Org.	Person	Thing	Time	Work
CaGE				✓	✓	✓		✓	
PorTexTO								✓	
Priberam	✓	✓	✓	✓	✓	✓	✓	✓	✓
R3M				*	*	*			
REMBRANDT	✓	✓	✓	✓	✓	✓	✓	✓	✓
REMMA	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEI-Geo				✓					
XIP		✓	✓	✓	✓	✓		✓	✓

Table 2.1: State of the Art: Comparison (Systems and entities).

regarding its classification module (see Section 2.5.3).

It is also important to retain that the purpose behind the creation of each system varies greatly. Some systems, mainly CaGE, REMBRANDT and SEI-Geo, were created as a result of their authors' thesis, which means that the system's scope is connected to the thesis. CaGE and SEI-Geo are mainly concerned with the extraction of geographic knowledge, thus it is conceivable that they are more inclined to deal with `LOCATION` entities, rather than trying to identify and classify all possible ones, regardless of their category. Even so, CaGE is able to identify and classify more entities than just `LOCATION` ones, as it is explained in Section 2.2.3. Other systems, such as Priberam, despite not having been explicitly created for the campaign (in fact, the system already existed and was in use – see Section 2.4.1), had to be adapted in order to follow the campaign's rules and to broaden its results.

All these systems were created for different purposes, even though they all share the intention of performing NER. As a result, this variety of objectives among the systems is partially responsible for some modest results. In more extreme cases there is even a complete lack of agreement between the system's principles and HAREM's rules (see Section 2.4.3).

The second important criterion is the technology behind each system. Table 2.2 presents a list of the systems and the technology they use in order to perform NER.

As easily noticed, one aspect they all share is the use of manual rules. As time-consuming as they may be, they are still nowadays one of the most effective methods of writing disambiguation and context capture rules, and patterns. Even systems that use other knowledge sources, such as Wikipedia or ontologies, still need to rely, on some level, on manual rules in order to increase precision and recall. REMMA, for example, specifically tried to determine if the use of Wikipedia alone would improve the overall results and reached the conclusion that it did not. A hybrid solution is far more effective.

The third and final criterion to consider in this comparison is the actual results. Table 2.3 presents an overview of the results. Table A.1, in Appendix A, presents the classification results by category.

There are cases in which different text segmentation possibilities occur and, in those cases, the alternatives are represented inside a `<ALT></ALT>` tag body, being each alternative separated by a "`|`"

System	Technologies		
	Manual rules	Automatic rules	Other sources
CaGE	✓		Dictionaries, Gazetteer
PorTexTO	✓		
Priberam	✓		Ontology, Lexicon, Grammar
R3M	✓	✓	
REMBRANDT	✓		Wikipedia
REMMA	✓		Lists of words, Wikipedia
SEI-Geo	✓		Ontologies
XIP	✓		Lexicons, Lists of words, Grammars

Table 2.2: State of the Art: Comparison (Systems and Technologies).

System	Identification			Classification		
	Precision	Recall	F-measure	Precision	Recall	F-measure
CaGE	0.5108	0.3773	0.4340	0.4499	0.2757	0.3419
PorTexTO	0.7003	0.0898	0.1592	0.6790	0.0882	0.1562
Priberam	0.6994	**0.7229**	**0.7110**	0.6417	**0.5146**	**0.5711**
R3M	0.7644	0.6170	0.6828			
REMBRANDT	0.7577	0.6214	0.6828	0.6497	0.5036	0.5674
REMMA	0.7083	0.4516	0.5515	0.6050	0.3615	0.4526
SEI-Geo	**0.8963**	0.1358	0.2359	**0.7485**	0.1166	0.2017
XIP	0.7214	0.5315	0.6121	0.6566	0.4652	0.5445

Table 2.3: State of the Art: Comparison (the global results, regarding the identification and classification tasks).

symbol. HAREM provides two kinds of scenarios: "strict" or "relaxed" ALT evaluation scenarios. The former takes into account all alternatives, whereas the latter only chooses the element which maximizes the classification result. The results shown in Tables 2.3 and A.1 are from the strict ALT evaluation scenario. The ones from the relaxed scenario can be consulted in Mota *et al.* [36, Tables I.17, I.18].

HAREM's evaluation directives (Oliveira *et al.* [37]) are comprised of global (all entities are considered) and selective scenarios (each one combines some of the entities). Globally, Priberam is the best system both in the identification and classification tasks, having reached the highest F-measure result in both (approximately 71% and 57%, respectively). It is also the only system that was able to achieve a higher recall than precision in the identification task.

Priberam has been involved in this area for twenty years and has a lot of experience. Their platform for linguistic development is greatly developed and it provides a large set of different products, some of which are mentioned in Section 2.4.1. When compared to the other systems, Priberam clearly distinguishes itself as a more mature system, especially due to the platform and to the research and

development that has been put into practice over the last twenty years. Therefore, one can extrapolate that these are the determinant factors that boost Priberam into the first place in the podium.

However, it is also important to notice that Priberam is not the most precise system among the participants. That award goes to SEI-Geo, which was able to identify 90% of all entities and classify 75%, even though it only deals with LOCATION entities. For this reason, SEI-Geo's recall is as low as 12%, which inevitably pushes this system to the bottom of the table.

A more realistic comparison is to consider Priberam, REMBRANDT and REMMA, whose common attempt to identify and classify all kinds of entities is an ambitious one. Although in this global scenario both REMBRANDT and REMMA are more precise than Priberam in the identification task (1 to 6% more precise), this difference is almost residual when compared to their recall differences in the same task: Priberam is 10 to 27% better.

Typically the presence of a low recall can be explained by two reasons: (a) the system is built to deal with a small set of all possible entities, or (b) some parts of the system are underdeveloped, either from lack of time or resources. PorTexTO, SEI-Geo and CaGE fall into the former, whereas REMMA and XIP fall into the latter. In XIP's case, there was not enough time to add as many lexical entries as desired. Consequently, XIP's results are very diverse.

On the one hand, XIP is the best system to classify temporal expressions, mainly because a lot of effort was put into improving the results on that particular category. On the other hand, the AMOUNT category proves to be one of XIP's weaknesses, for the F-measure was 42%, which is very low for a category whose excellent results are typical: Bick [7] reached a 95–97% F-measure during the First HAREM. Therefore, XIP can surely be very improved in this area, as well as in others (see Section 3.3), although already being a good NER system. Globally, it was the third best, being only surpassed by Priberam, R3M and REMBRANDT (these two tied in second place), and at a low percentage distance from them (7–9% in identification and 2–3% in classification).

Chapter 3

Architecture

THIS chapter of the document is comprised of three parts. First, in Section 3.1, the processing chain in which XIP is inserted is described; second, Section 3.2 presents a detailed description of the XIP system, covering chunking and dependency rules, as well as custom lexicons and local grammars; finally, Section 3.3 presents the improvements that have been made in the course of this thesis.

3.1 Processing chain

L^2F has developed a processing chain that is comprised of several modules (see Figure 3.1). The chain is divided in three main stages:

- Pre-processing;
- Disambiguation (rule-driven and statistical);
- Syntactic analysis.

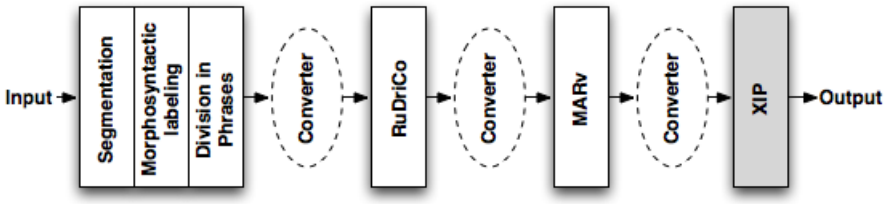

Figure 3.1: XIP: The processing chain in which the system resides (from Romão [43, Section 3, Figure 3.1]).

3.1.1 Pre-processing

The pre-processing is comprised of three modules. The first one, the segmentation module, is mainly responsible for dividing the input into individual segments, also known as "tokens" (as a result, it is

sometimes referred to as the "tokenizer"). For example, consider the sentence "O Diogo foi ao Japão" (Diogo went to Japan) given as input to the segmentation module. In this case, the output would be:

```
word[0]:  |O|
word[1]:  |Diogo|
word[2]:  |foi|
word[3]:  |ao|
word[4]:  |Japão|
word[5]:  |.|
```

Besides this, the module is also responsible for the early identification of certain types of NEs, namely: email addresses, ordinal numbers (e.g. 3º, 42ª), numbers with "." and "," (e.g. 12.345,67), IP and HTTP addresses, integers (e.g. 12345), several abbreviations with "." (e.g. "a.c.", "V.Exa."), sequences of interrogation and exclamation marks, as well as ellipsis (e.g. ???, !!!, ?!?!, ...), punctuation marks (e.g. ! ? . , : ; ( ) [ ] -), symbols (e.g. «, », #, $, %, &, +, *, <, >, =, @), Roman numerals (e.g. LI, MMM, XIV) and also words, such as "alface" (lettuce) and "fim-de-semana" (weekend).

According to Mamede [27], one of the problems regarding this module was that it did not perform the identification of numbers written in full, such as "duzentos e trinta e cinco" (two hundred and thirty-five). This, as explained in Section 3.3.1, has been corrected and the identification is now fully operational.

Afterwards, the segmentation module's output tokens are tagged with POS (part of speech) labels, such as "noun", "verb" or "adjective", among others (see Table 3.1 for a complete list). There are thirteen categories (Table 3.1's rows) and the information is encoded in ten fields (Table 3.1's columns): category (CAT), subcategory (SCT), mood (MOD), tense (TEN), person (PER), number (NUM), gender (GEN), degree (DEG), case (CAS) and formation (FOR). No category uses all ten fields.

Consider again the sentence presented above. At this stage, the output would be:

```
word[0]:  |O|      POS->[o]Td...sm... [o]Pp..3sm.as
word[1]:  |Diogo|  POS->[Diogo]Np...sm...
word[2]:  |foi|    POS->[ser]V.is3s=... [ir]V.is3s=...
word[3]:  |ao|     POS->[ao]S...sm..f
word[4]:  |Japão|  POS->[Japão]Np...sm...
word[5]:  |.|      POS->[.]O.........
```

As easily noticed, at this point each token is assigned a POS tag and each tag has a corresponding code. For example, "Diogo" and "Japão" are "Np...sm...", which means that they are proper nouns, singular number and male gender. Whenever a token is ambiguous and might belong to several categories, they are all listed: this is what happens with the token "foi", which might mean "ser" (to be), as in "Ele foi Rei" (He was King) or it might mean "ir" (to go), as in "Ele foi ao Japão" (He went to Japan).

This tagging is performed by the Palavroso system (Medeiros [32]), which is very old and needs replacement or improvement. According to Mamede [27], Palavroso suffers from several problems:

- The lemmas are not adequate for parsing;

Fields	CAT	SCT	MOD	TEN	PER	NUM	GEN	DEG	CAS	FOR
Noun	1	2				6	7			
Verb	1		3	4	5	6	7			
Adjective	1					6	7	8		
Pronoun	1	2			5	6	7		9	10
Article	1	2				6	7			
Adverb	1							8		
Preposition	1					6	7			10
Conjunction	1	2								
Number	1	2				6	7			
Interjection	1									
Passive marker	1									
Residual	1	2								
Punctuation	1									

Table 3.1: Processing chain: POS tags (fields and categories).

- Guessed words cannot be distinguished from the ones that are in the dictionary;
- Verbs and adverbs are not subcategorized;
- It is hard to insert new features;
- It is hard to control the output.

The final step of the pre-processing stage is the text division into sentences. In order to build a sentence, the system matches sequences that end either with ".", "!" or "?". There are, however, two exceptions to this rule:

- All registered abbreviations (e.g. N.A.S.A.);
- If any of the following symbols or any lower case letter is found after an ellipsis: "»", ")", "]", "}".

Finally, the output is converted to XML by the first converter module seen in Figure 3.1.

3.1.2 Disambiguation

The next stage of the processing chain is the disambiguation process, which is comprised of two steps:

- Rule-driven morphosyntactic disambiguation, performed by RuDriCo (Pardal [39]);
- Statistical disambiguation, performed by MARv (Ribeiro *et al.* [42]).

3.1.2.1 RuDriCo

According to Pardal [39], RuDriCo's main goal is to provide for an adjustment of the results produced by a morphological analyzer to the specific needs of each parser. In order to achieve this, it modifies the segmentation that is done by the former. For example, it might contract expressions provided by the morphological analyzer, such as "ex-" and "aluno", into one segment: "ex-aluno"; or it can perform the opposite and expand expressions such as "nas" into two segments: "em" and "as". This will depend on what the parser might need.

Altering the segmentation is also useful for performing tasks such as recognition of numbers and dates. The ability to modify the segmentation is achieved through declarative rules, which are based on the concept of pattern matching. RuDriCo can also be used to solve (or introduce) morphosyntactic ambiguities. By the time RuDriCo is executed along the processing chain, it performs all of the mentioned tasks, and more: it also corrects some of Palavroso's output (e.g. "sida") and it modifies the lemmas of the pronouns, adverbs, articles, etc (e.g. "quaisquer").

According to Mamede [27], RuDriCo had two major problems:

- The rules lacked expressiveness (e.g. lack of operators such as negation or disjunction, and the ability to change only the lemma);
- The programming was not very efficient (e.g. it dealt mostly with strings and not integers).

Currently, RuDriCo 2.0 is in use and it is the result of several improvements to RuDriCo, which have been made by Diniz [16] in the context of his Master Thesis.

Finally, the output is converted by the second converter module seen in Figure 3.1.

3.1.2.2 MARv

MARv's main goal is to analyze the labels that were attributed to each token in the previous step of the processing chain, and then choose the most likely label for each one. In order to achieve this, it employs the statistical model known as Hidden Markov Model (HMM). Without getting into much detail, for it is out of the scope of this document to provide a detailed description of HMMs, a HMM is a very important machine learning model in speech and language processing. According to Jurafsky & Martin [21, see Ch. 6, Secs. 6.1 & 6.2], in order to properly define a HMM, first one needs to introduce the Markov chain, sometimes called the observed Markov model.

A Markov chain is a special case of a weighted automaton in which the input sequence uniquely determines which states the automaton will go through. Because it cannot represent inherently ambiguous problems, a Markov chain is only useful for assigning probabilities to unambiguous sequences, that is, when we need to compute a probability for a sequence of events that can be observed in the world. However, in many cases events may not be directly observable.

In this case in particular, POS tags are not observable: what we see are words, or "tokens", and we need to infer the correct tags from the word sequence. So we say that the tags are "hidden" – because they are not observed. Hence, a HMM allows us to talk about both observed events (like words that we see in the input) and hidden events (like POS tags).

There are many algorithms to compute the likelihood of a particular observation sequence. MARv uses the Viterbi algorithm, which can be analyzed in Jurafsky & Martin [21, see Ch. 6, Sec. 6.4].

Currently, the processing chain is using MARv 3.0, which is faster than the previous version and it also stores the deprecated tags. Still, according to Mamede [27], MARv suffers from several problems that need to be overcome in the future:

- The word's category and subcategory are the only criteria used in its choice;
- It does not choose a lemma;
- It does not choose the verb's tense;
- The training corpus needs to be increased (it currently has 250.000 words).

Finally, the output is converted by the third (and final) converter module seen in Figure 3.1.

3.1.3 Syntactic analysis

The third and final stage of the processing chain is the syntactic analysis performed by XIP. This is where the identification and classification of NEs occurs, and where the major work done for this thesis has taken place. XIP is a language-independent parser that takes textual input and provides linguistic information about it. XIP can modify and enrich lexical entries, construct chunks and other types of groupings, and build dependency relationships (Xerox [49]). The next section provides a detailed description of XIP and its main characteristics.

To conclude this section, it is worth mentioning that each of the stages above can be parameterized; in particular, XIP allows a parameterization using local grammars and lexicon files, which are analyzed below.

3.2 The XIP system in the processing chain

XIP receives as input the converted data from MARv and is able to handle it in order to perform several tasks, namely:

- Calculation of chunks and dependencies;
- Adding lexical, syntactic and semantic information;
- Applying morphosyntactic disambiguation rules;
- Applying local grammars;

The fundamental data representation unit in XIP is the node. A node has a category, feature-value pairs and "brother" nodes. For example, the node below represents the noun "Diogo" and it has several features that are used as a means to express its properties. In this case, the features have the following meaning: "Diogo" is a noun that represents a human, an individual male (feature `masc`); the node also has features to describe its number (singular, `sg`) and the fact that it is spelled with an upper case initial letter (feature `maj`):

```
Diogo: noun[human, individual, proper, firstname, people, sg, masc, maj]
```

Every node category and every feature must be declared in declaration files. Furthermore, features must be declared with their domain of possible values. They are an extremely important part of XIP, as they describe the properties of nodes. Features, by themselves, do not exist; they are always associated with a value, hence the so-called feature-value pair.

Moreover, features can be instantiated (operator =), tested (operator :), or deleted (operator =~) within all types of rules (Aït-Mokhtar *et al.* [6]). While instantiation and deletion are all about setting/removing values to/from features, testing consists of checking whether a specific value is set to a specific feature:

Type	Example	Explanation
Instantiated	`[gender = fem]`	The value "fem" is set to the feature "gender"
Tested	`[gender:fem]`	Does the feature "gender" have the value "fem" ?
	`[gender:~]`	The feature "gender" should not be instantiated on the node
	`[gender:~fem]`	The feature "gender" should not have the value "fem"
Deleted	`[acc =~]`	The feature "acc" is cleared of all values on the node

Table 3.2: XIP: examples of features.

3.2.1 Chunks and dependencies

3.2.1.1 Chunking rules

Chunking is the process by which sequences of categories are grouped into structures; this process is achieved through chunking rules. There are two types of chunking rules (Xerox [49]):

- Immediate dependency and linear precedence rules (ID/LP rules);
- Sequence rules.

The first important aspect about chunking rules is that each one must be defined in a specific layer. This layer is represented by an integer number, ranging from 1 to 300. Below is an example of how to define two rules in two different layers:

```
1>  NP = (art;?[dem]), ?[indef1].  // layer 1
2>  NP = (art;?[dem]), ?[poss].    // layer 2
```

Layers are processed sequentially from the first one to the last. Each layer can contain only one type of chunking rule.

ID/LP rules are significantly different from sequence rules. While ID rules describe unordered sets of nodes and LP rules work with ID rules to establish some order between the categories, sequence rules describe an ordered sequence of nodes. The syntax of an ID rule is:

```
layer> node-name -> list-of-lexical-nodes.
```

Consider the following example of an ID rule:

```
1> NP -> det, noun, adj.
```

Assuming that `det`, `noun` and `adj` are categories that have already been declared (see Table B.1 in Appendix B for a complete list of the possible POS categories), this rule is interpreted as follows: "whenever there is a sequence of a determiner, noun and adjective, regardless of the order in which they appear, create a Noun Phrase (NP) node". Obviously, this rule applies to more expressions than those desirable, e.g. "o carro preto" (the car black), "o preto carro" (the black car), "preto carro o" (black car the) and "carro preto o" (car black the)[1]. This is where LP rules come in. By being associated with ID rules, they can apply to a particular layer or be treated as a general constraint throughout the XIP grammar. They have the following syntax:

```
layer> [set-of-features] < [set-of-features].
```

Consider the following example:

```
1>  [det:+]   < [noun:+].
1>  [noun:+]  < [adj:+].
```

Thus, by stating that a determiner must precede a noun only in layer 1, and that a noun must precede an adjective also only in layer 1, the system is now setting constraints in this layer, which means that expressions such as "o preto carro" (the black car) will no longer be allowed. However, "o carro preto" (the car black) will[2].

It is also possible to use parentheses to express optional categories, and an asterisk to indicate that zero or more instances of a category are accepted. The following rule states that the determiner is optional and that as many adjectives as possible are accepted:

```
1> NP -> (det), noun, adj*.
```

Taking into account both LP rules established above, the following expressions are accepted: "carro" (car), "carro preto" (car black), "o carro preto" (the car black), "o carro preto bonito" (the car black beautiful).

Finally, it is worth mentioning that these rules can be further constrained with contexts. For example:

```
1> NP -> |det, ?*| noun, adj |?*, verb|.
```

Simple enough, this rule states that a determiner must be on the left of the set of categories, and that a verb must be on the right. By applying this rule on a sentence such as "o carro preto andou na estrada" (the black car went on the road), we obtain the following chunk:

```
NP[o carro preto].
```

Hence, although they help constraining a rule even further, contexts are not "saved" inside a node.

The other kind of chunking rules, sequence rules, though conceptually different because they describe an ordered sequence of nodes, are almost equal to the ID/LP rules in terms of syntax. There are, however, some differences and additions:

[1] A word-for-word translation is provided only to illustrate syntactic phenomena. Its acceptability in English is irrelevant.

[2] Naturally, these are just examples of ID/LP rules. The actual grammatical rules governing the relative position of adjectives and nouns are much more complex.

- Sequence rules do not use the `->` operator. Instead, they use the `=` operator, which matches the shortest possible sequence. In order to match the longest possible sequence, the `@=` operator is used;
- There is an operator for applying negation (`~`) and another for applying disjunction (`;`);
- Unlike ID/LP rules, the question mark (`?`) can be used to represent any category on the right side of a rule;
- Sequence rules can use variables.

The following sequence rule matches expressions like "alguns rapazes/uns rapazes" (some boys), "nenhum rapaz" (no boy), "muitos rapazes" (many boys) or "cinco rapazes" (five boys):

```
1> NP @= ?[indef2];?[q3];num, (AP;adj;pastpart), noun.
```

Finally, consider again the example from Section 3.1.1, "O Diogo foi ao Japão." (Diogo went to Japan). At this stage, after the pre-processing and disambiguation, and also after applying the chunking rules, the system presents the following output tree:

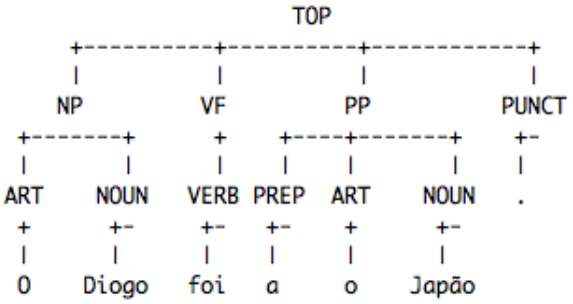

Figure 3.2: XIP: output tree after applying the chunking rules.

3.2.1.2 Dependency rules

Being able to extract dependencies between nodes is very important because it can provide us with a richer, deeper understanding of the texts.

Dependency rules take the sequences of constituent nodes identified by the chunking rules and identify relationships between them (Xerox [49]). This section presents a brief overview of their syntax, operators, and some examples.

A dependency rule presents the following syntax:

```
|pattern| if <condition> <dependency_terms>.
```

In order to understand what the `pattern` is, first it is essential to understand what is a Tree Regular Expression (TRE). A TRE is a special type of regular expression that is used in XIP in order to establish connections between distant nodes. In particular, TREs explore the inner structure of subnodes through the use of the braces characters (`{}`). The following example states that a NP node's inner structure must be examined in order to see if it is "made of" a determiner and a noun:

```
NP{det, noun}.
```

TREs support the use of several operators, namely:

- The semicolon (`;`) operator is used to indicate disjunction;
- The asterisk (`*`) operator is used to indicate "zero or more";
- The question mark (`?`) operator is used to indicate "any";
- The circumflex (`^`) operator is used to explore subnodes for a category.

Hence, and returning to the dependency rules, the `pattern` contains a TRE that describes the structural properties of parts of the input tree. The `condition` is any Boolean expression supported by XIP (with the appropriate syntax), and the `dependency_terms` are the consequent of the rule.

The first dependency rules to be executed are the ones that establish the relationships between the nodes, as seen in the next example:

```
| NP#1{?*, #2[last]} |
    HEAD(#2, #1)
```

This rule identifies HEAD relations (see below), for example "a bela rapariga" (the beautiful girl) ⇒ HEAD(rapariga, a bela rapariga).

As already stated, the main goal of the dependency rules is to establish relationships between the nodes. Coming back to our usual example, the following output is the current result of applying these rules to the sentence "O Diogo foi ao Japão." (Diogo went to Japan):

```
MAIN(foi)
HEAD(Diogo,O Diogo)
HEAD(Japão,a o Japão)
HEAD(foi,foi)
DETD(Diogo,O)
DETD(Japão,o)
PREPD(Japão,a)
VDOMAIN(foi,foi)
MOD_POST(foi,Japão)
SUBJ_PRE(foi,Diogo)
NE_INDIVIDUAL_PEOPLE(Diogo)
NE_LOCAL_COUNTRY_ADMIN_AREA(Japão)
```

The last two indicate that two NEs have been captured and classified in this sentence: "Diogo" has been captured and classified as `HUMAN INDIVIDUAL PERSON` and "Japão" (Japan) has been captured and classified as `LOCATION CREATED COUNTRY`. The tags `NE_INDIVIDUAL_PEOPLE` and `NE_LOCAL_COUNTRY_ADMIN_AREA` are merely used to see that the NEs have been classified. The final XML tags are created afterwards, as the final step of the whole process.

The other dependencies listed above cover a wide range of binary relationships such as:

- The relation between the nucleus of some chunk and the chunk itself (`HEAD`);
- The relation between a nominal head and a determiner (`DETD`);
- The relation between the head of a Prepositional Phrase (PP) and the preposition (`PREPD`);
- Among many others.

To see a complete list and a detailed description of all dependency relationships as of July 2009, please refer to Mamede *et al.* [28, see Sec. 2] and Santos [44].

Now consider the following example of another kind of dependency rule (aimed at classifying NEs):

```
| #1{?*, num[quant,sports_results]} |
  if (~NE[quant,sports_results](#1))
    NE[quant=+,sports_results=+](#1)
```

This rule uses a variable, represented by `#1`, which is assigned to the top node, because it is placed before the first brace (`{`). This variable could have been placed inside the braces structure, assigned (for example) to the node `num`. This rule states that if a node is "made of" any category followed by a number (with two features that determine whether it is a sports result), and if this node has not yet been classified as a NE with these features, then one wants to add them to the top node in order to classify it as `AMOUNT SPORTS_RESULT`. Please notice that it is the top node that is classified, because the variable is assigned to it; if it had been placed next to the node `num`, for example, then only this subnode would have been classified.

Notice also the usage of the negation operator (`~`) inside the conditional statement. XIP's syntax for these conditional statements also allows the operators `&` for conjunction and `|` for disjunction. Parentheses are also used to group statements and establish a clearer precedence, as in most programming languages.

3.2.2 Custom lexicons, local grammars and disambiguation rules

3.2.2.1 Lexicons

XIP allows the definition of custom lexicons (lexicon files), which add new features that are not stored in the standard lexicon. Having a rich vocabulary in the system can be very beneficial for improving its recall.

In XIP, a lexicon file begins by simply stating `Vocabulary:`, which tells the XIP engine that the file contains a custom lexicon. Only afterwards come the actual additions to the vocabulary.

The lexical rules attempt to provide a more precise interpretation of the tokens associated with a node (Xerox [49]). They have the following syntax (the parts of the rule contained in parentheses are optional):

```
lemma(: POS([features])) (+)= (POS)[features].
```

Some examples of lexical rules follow:

```
$US              =  noun[meas=+, curr=+].
eleitor:  noun +=  [human=+].
google           +=  verb[intransitive=+].
```

The first two examples show how to add new features to existing words (in this case, they are both nouns). In the first case, the features `meas` (measure) and `curr` (currency) are added to `$US`; in the second case, the `human` feature is added to `eleitor` (elector). In the third case, however, `google` is given the additional reading of `verb`.

3.2.2.2 Local grammars

Local grammars are text files that contain chunking rules and each file may contain ID/LP and sequence rules. Essentially, we use different local grammar files to capture desirable sequences of nodes and to attribute features to them. We employ a division based on different categories of NEs. For example, whereas the file `LGLocation` is aimed at capturing sequences of nodes related to the `LOCATION` category, the file `LGPeople` will capture sequences of nodes related to the `INDIVIDUAL` type (`HUMAN` category).

After the pre-processing and disambiguation stages, XIP receives its input sentence(s) and tries to match it/them to the rules in the local grammars' files. They are run sequentially through a predefined order in a configuration file.

As an example, consider the following sequence rule belonging to the local grammar responsible for dealing with `LOCATION` NEs:

```
1> noun[location=+, admin_area=+] = ?[lemma:novo,maj];?[lemma:nova,maj],
                                    noun[location,maj].
```

This rule is responsible for matching expressions such as "Novo México" (New Mexico), "Nova Zelândia" (New Zealand) or "Nova Escócia" (New Scotland), and then it creates a `noun` node with two feature-value pairs (`location` and `admin_area`). Notice how the `?` and `;` operators were used in order to capture either "Novo" or "Nova".

3.2.2.3 Disambiguation rules

To conclude this section, it is also important to state that XIP allows the definition of disambiguation rules. The general syntax for a disambiguation rule is (Xerox [49]):

```
layer> readings_filter = |left_context| selected_readings |right_context|.
```

Like chunking rules (see Section 3.2.1.1), disambiguation rules also employ the concept of layer and contexts. The left side of a disambiguation rule contains the `readings_filter`. This filter specifies a subset of categories and features that can be associated with a word. The list can be constrained with features and the filter applies when it matches a subset of the complete ambiguity class of a word. Finally, the `selected_readings` portion of a disambiguation rule gives the selected interpretation(s) of the word.

There are four main operators used in disambiguation rules:

- The <> operator: it is used to define specific features associated with a category;
- The `[]` operator: it is used to refer to the complete set of features for a category;
- The % operator: it restricts the interpretation of a word to one solution;
- The <* operator: when used, one specifies that each reading must bear the features listed immediately after.

Consider the example below:

```
1> ?<maj:+,start:~> = ?<proper:+> .
```

This rule states that upper case words (other than at the beginning of a sentence) must be a proper name.

3.3 Improvements

This section describes the improvements that have been made to the XIP system in the course of this thesis. Each subsection below (from 3.3.1 to 3.3.7) is related to a specific task.

3.3.1 Segmentation

As already explained in Section 3.1, XIP is inserted in a long processing chain (Figure 3.1). This task was intended to be applied in the chain's first module, the Segmentation module. At this stage (among other things), there is a script written in Perl that is responsible for early identification of standard NEs, namely: XML tags, email addresses, ordinal numbers, currency units with symbols (e.g. $), HTTP and IP addresses, abbreviations, acronyms, symbols, punctuation marks, and others. Numbers written in full, such as "trezentos e quarenta e dois" (three hundred and forty-two), however, were not being captured at this stage, but delayed to a later stage, in which a specific set of rules joined the various tokens to form a single one, which was the number itself.

By capturing them as early as the Segmentation stage, not only would the system benefit from the fact that Perl has an extremely efficient and fast regexp matching mechanism, which rapidly detects most of the cases, but it would also free subsequent processes from this task, allowing them to deal with other situations.

The script has been improved in order to support the detection of numbers written in full, ranging from 0 (zero) to 999.999.999 (novecentos e noventa e nove milhões, novecentos e noventa e nove mil, novecentos e noventa e nove). As mentioned before, this module had to be integrated with the "tokenizer" script of the processing chain. Because of this, certain rules had to be followed. In particular, the script analyzes a sentence word by word and it uses a sliding window technique, which means that no regexp could impose the "end of string" meta-character $. Also, all regexps work by detecting the word in the beginning of the string, using the meta-character ^.

The script now successfully detects all numbers written in full (case-insensitive), and with different writing styles. For example, the number 1.101 (one thousand one hundred and one) is read: "mil cento

e um", but it can be written as mentioned, or "mil, cento e um", or "mil e cento e um", or "mil, e cento e um", and so on. Notice, however, that some expressions such as "um mil" (one thousand), which are not used in Portugal but are used in Brazil, are not accepted by the script. In this case, the script would consider two separate numbers: "um" (one) and "mil" (thousand). Moreover, the script is strict in some cases, such as the number 1.123 (one thousand one hundred and twenty-three): "mil cento e vinte e três". In this case, it is mandatory to write the number as mentioned (the last "e" is mandatory), or else the script will split the string in two number words. The script always matches the longest, most correct sequence. Therefore, if it detects "mil cento e vinte três" (with the last connective missing), it will match "mil cento e vinte" (1.120), which is the longest, most correct sequence. In this sort of task, even if one cannot work under the assumption that the text is always correctly written, it is however necessary to constraint the expressive power of the tokenizing rules in order to achieve higher linguistic adequacy.

Nevertheless, careful attention has been taken in order to prevent the script from detecting wrong number exceptions: it is harder to create a program that does not match anything unwanted, than it is to create one that matches everything that is desired. Misspelled numbers are not detected, nor are badly formed ones, such as "vinte e trinta" (twenty and thirty) or "duzentos milhões e seiscentos milhões e trezentos mil e quarenta e dois" (two hundred million and six hundred million and three hundred thousand and forty-two); once again, the script matches the longest, most correct sequences, which in this case are "vinte" (twenty) and "duzentos milhões e seiscentos" (two hundred million and six hundred), respectively. In reality, it is easily noticeable that each of these two cases represent two quantities: 20 and 30 in the former, 200.000.000 and 600.300.042 in the latter. However, since the script was integrated with the existing "tokenizer" script, after a first successful match the subsequent word is analyzed, thus both quantities will be correctly detected.

Furthermore, the script also detects digits followed by either "mil" (thousand), "milhões" (millions) and "mil milhões" (thousand millions), or by "dezenas" (tens) and "centenas" (hundreds). For example, the following numbers are all accepted: "2 mil", "142.876 milhões", "0,5 mil milhões", "duas centenas", "cento e uma mil, trezentas e quarenta e duas dezenas", "3,14 mil milhões de centenas", etc.

Finally, two remarks:

- First, the script only matches numbers that are written as such. This means that coincidences such as the number "dez" (ten) being part of another Portuguese word (e.g. "dezenas") are dealt with and are not detected. For a word to be considered a number, it must have a terminator character, which has been defined as any character that is not a letter. So, "dez.", "dez…", "dez,", "dez-" and "dez!", among many others, are all correctly matched[3];

- Second, in order to guarantee that all numbers were being detected, a second script was used: this one converts numbers (digits) to their corresponding words, and as a result we were able to create a text file with all the numbers written in full, one per line. By running our script against this file, we were able to match each line.

These improvements resulted in a trade-off. Later stages benefited from this, as their computational

[3]The abbreviation "Dez." or "dez." for December is only matched if the context so allows, i.e., "10 Dez 2010".

load was reduced (in particular, RuDriCo was the greatest beneficiary); however, the price to pay turned out to be that the "tokenizer" script is now a little more complex and consequently the Segmentation stage now takes a little more time to process.

3.3.2 Consistency

Since XIP is inserted in a long processing chain, errors may appear due to inconsistencies between the several modules. One of the most common arises when a particular rule in XIP is expecting a lemma for a token, but that lemma has been changed along the chain and so the rule can not be matched.

Lemmas are often changed as a result of improvements in RuDriCo. For example, at one point in time XIP had a rule to capture "Presidente da República" (President of the Republic) token by token:

```
1> noun[people=+,position=+] @= ?[lemma:presidente,maj], prep[lemma:de],
                                art, ?[lemma:república,maj].
```

However, if RuDriCo's rules are improved and "Presidente da República" may now be captured as an unique token once the disambiguation process is over, then the rule above will not be matched and, as a result, it will have to be changed into the following:

```
1> ? @= ?[lemma:"Presidente da República", people=+, position=+].
```

Notice how this new rule does not create a noun node, but instead it just adds the `people` and `position` features to an already existing noun node.

According to Diniz [16], RuDriCo 2.0 currently has 28.733 contraction rules, which is quite a contrast to the 3.096 contraction rules it had before (there was an increase by a factor of 9,3 in a period of one year and a half). Consequently, this task has been extremely important in order to guarantee not only that XIP's rules can keep up with the development of RuDriCo, but also that the overall results are not compromised.

3.3.3 Classification Directives

Having participated in the Second HAREM, L^2F accepted the proposed classification directives even though other solutions could have been adopted. The main divergences concerned the `TIME` expressions, which justified the proposal of a specific set of directives for that campaign (Hagège *et al.* [18]), and some of the categories, in which the system chose not to participate at all (see Table 2.1).

For the development of the processing chain and of its NER modules, a new set of NE classification directives was developed, different from, although inspired by, the Second HAREM evaluation campaign. Table 3.3 shows an overview of the changes that have been introduced in the new set of directives.

Two major modifications were made in the directives:

- the general reformulation regarding the `PERSON` and `ORGANIZATION` categories, and some minor adjustments involving the exclusion of some types and inclusion of others in some of the existing categories;

Second HAREM			Thesis		
CATEGORY	TYPE	SUBTYPE	CATEGORY	TYPE	SUBTYPE
Person	Individual Position GroupPosition GroupMember Member GroupInd People		Human	Individual	Person Position
Organization	Administration Company Institution			Collective	Administration Institution Group
Location	Human	Country Division Region Construction Street	Location	Created	Country Division Region Construction Street
	Physical	Watermass Watercourse Relief Planet Island Region Other		Physical	Watermass Watercourse Relief Planet Island NaturalRegion
	Virtual	Site Work SocialCom Other		Virtual	Site Documents
Amount	Quantity Currency Classification		Amount	Quantity Currency Classification SportsResults	

Table 3.3: Classification directives: differences between the two sets of directives.

- the inclusion of metonymy.

Unlike other ontologies, which place the distinction between PERSON and ORGANIZATION at the very top, we have decided to create a HUMAN category at the very top, for several reasons of linguistic nature. Since the strategies used for the identification and classification of NEs result from the linguistic properties they present in the texts, the categories must also result from those properties.

Language itself clearly distinguishes human from non-human expressions, but the distinctions between a singular person (HUMAN INDIVIDUAL) and a collective (HUMAN COLLECTIVE) are much fuzzier. These are merely marked by some specialized predicates and, even then, not always in a sufficiently clear manner. For example, verbs like "filiar-se em" as in "Ele filiou-se no Partido Social Democrata" (he joined the Social Democratic Party), or "aderir a", as in "Ele aderiu à Ordem dos Engenheiros" (he joined the Order of Engineers), have a collective in their second argument position (complement) and cannot admit an individual in that position. Similarly, there are verbs that only admit individuals as their argument, e.g. "beijar" (to kiss).

However, regarding the verbs with human arguments, it is very common to observe collective NEs in those positions, because there is often a metonymical transference from the members of the collective to the name that designates the collective as a whole. As a result, we can have sentences like "A IBM anda a namorar a Google há muito tempo" (IBM has been courting Google for a long time).

Moreover, a HUMAN category is clearly broader in scope than the PERSON and ORGANIZATION categories. For these reasons, we have decided that it would be more adequate to have a top HUMAN category and, within it, to distinguish between PEOPLE and COLLECTIVE.

Finally, regarding metonymy, the campaign's directives included at least three types under the PERSON category (namely: MEMBER, GROUPMEMBER and PEOPLE) that were being used to capture metonymy. These types have been excluded and now metonymy is treated in a very different way (see Section 3.3.7 and Appendix C.4).

For the remaining categories changes were minor, as can be seen in Table 3.3.

The following sections contain detailed descriptions of the improvements that have been made to the system during this thesis, with respect to the directives' categories.

3.3.4 AMOUNT category

The AMOUNT category is very important for the NER task because amounts are very common in all sorts of texts and they constitute relevant information in many IE/IR applications. From simple, isolated numbers (e.g. cardinal and ordinal numbers; integers and fractional numbers; percentages and intervals; roman numerals, etc.), to monetary amounts (e.g. 100 euros, 42 dollars, etc.) and other expressions, this category encompasses many NEs (approximately 5% of all annotated entities in the Golden Collection of the Second HAREM were AMOUNT). Unfortunately, and for reasons that have already been explained in Section 2.10, the XIP system obtained below average results in this category during the Second HAREM evaluation campaign (see Table A.1 in Appendix A). Therefore, it was important to improve the system in this category, not only in terms of precision, but also for its recall.

The four following sections explain the improvements that have been made to the system regarding this category's types: QUANTITY (Section 3.3.4.1), CURRENCY (Section 3.3.4.2), CLASSIFICATION (Section 3.3.4.3) and SPORTS RESULTS (Section 3.3.4.4).

3.3.4.1 QUANTITY type

In an effort to improve the system's recall regarding this category, and in particular this type, there has been a major restructuring of the lexical files, having added a total of 212 lexical entries in the following areas:

- Units of frequency (62 entries: multiples of Hertz and their abbreviations, e.g. "MHz", "megahertz", "kHz");
- Units of volume (15 entries, e.g. "quilolitro", "kl", "decilitro", "dl");
- Units of length (58 entries, multiples of meter with both Portuguese spellings (Portugal and Brazil), e.g. "quilómetro", "quilômetro", "km", "milimícron");
- Units of mass (54 entries: multiples of gram and their abbreviations, e.g. "decagrama", "dag", "tonelada", "nanograma");
- Prefixes (19 entries, e.g. "yotta", "zetta", "tera", "giga", "micro", "mili", "femto", "yocto");
- Other missing abbreviations (3 entries: "min", "mseg" and "kilo", for "minuto", "milisegundo" and "quilograma", respectively);
- Power to weight ratio (kg/cv).

Notice, however, that, if isolated in the text, none of these lexical entries will be captured and/or classified as QUANTITY. In order for that to happen, they must be found together with a numeric value (digit or not), for that is the definition of a QUANTITY NE. Furthermore, rules have been modified in order to change the delimitation of this type of NE. According to the Second HAREM's directives, the sentence "o barco estava a 50 milhas da costa" (the boat was at 50 miles from the coast) should indicate one NE: "50 milhas" (50 miles). However, following the trend initiated with the time expressions' directives, in the new set of directives the superficial parsing (chunking) is kept in the delimitation of the NE. Therefore, in this sentence the system must indicate the following NE: "a 50 milhas" (at 50 miles), i.e., if the quantified expression is in a PP, then the preposition introducing the PP must also be included.

This change in the delimitation of QUANTITY NEs was implemented by a dependency rule:

```
| PP#1[quant]{?*, num, ?[meas, time_meas:~, curr:~]} |
   if ( ~NE(#1) )
      NE[quant=+,num=+](#1)
```

Essentially, this rule states that if a node is a PP which consists of something followed by a number and by a unit of measure (as long as it is not a time or currency measure), then the whole node should be marked as an AMOUNT QUANTITY NE.

Appendix C.1.2.1 presents some examples of correct and incorrect annotations for this type.

3.3.4.2 CURRENCY type

After having run the system with a significant amount of corpora, a common error was noticed when capturing and classifying currency units. As an example, consider the following (incorrect) annotation:

- <EM CATEG="VALOR" TIPO="QUANTIDADE">**1**</EM> dinar tunisino valorizou-se face a <EM CATEG="VALOR" TIPO="QUANTIDADE">**1**</EM> dinar da Argélia.

The system did not recognize "dinar tunisino" (Tunisian dinar) and "dinar da Argélia" (dinar from Algeria) as the strings involving names of countries and their derived (gentilic) adjectives: because of this, it captured and classified "1", an otherwise isolated number, as an AMOUNT (type QUANTITY) named entity.

In this case, the first step for improving the system has been to complete the list of currency nouns using Wikipedia[4], which provided a very complete list of existing currencies. The system had already been thoroughly built with a complete list of all three letter codes for currencies from all over the world, but it lacked more complex expressions such as those presented in the example above.

A total of 177 lexical entries have been added to the lexical file responsible for dealing with currencies. Note, however, that each entry admits many, but equivalent, different designations: for example, the entry "dólar americano" (american dollar) includes also the plural "dólares americanos" (american dollars), "dólar(es) dos Estados Unidos" (dollar(s) of the United States) and "dólar(es) dos Estados Unidos da América" (dollar(s) of the United States of America).

The different designations have been added because of alternate derivations that form the gentilics from certain countries' names. As a result, some currencies such as the Bahraini dinar support up to 14 different spellings (e.g. "dinar baremense", "dinar baremês", "dinar bareinita", "dinar do Barém", "dinar do Bahrain").

The total amount of new currency expressions supported by the system is currently 1.031.

Consider again the example above. After having added all these lexical entries, the system now produces the following output (which is not yet the intended result):

- <EM CATEG="VALOR" TIPO="MOEDA">**1 dinar tunisino**</EM> valorizou-se face a <EM CATEG="VALOR" TIPO="MOEDA">**1 dinar da Argélia**</EM>.

According to the classification directives, the delimitation of AMOUNT NEs (see Appendix C.1.1) must include the preposition introducing a PP (in this case, "face a" should be included in the second NE).

Therefore, a dependency rule has been created to mark PP expressions such as this one: it follows the pattern PREP + NUM + NOUN[currency,measure], i.e., a preposition followed by a number, followed by a noun with currency and measure features. In this case, PREP is "face a" (against), NUM is "1" and NOUN is "dinar da Argélia". The final result is presented below.

- <EM CATEG="VALOR" TIPO="MOEDA">**1 dinar tunisino**</EM> valorizou-se <EM CATEG="VALOR" TIPO="MOEDA">**face a 1 dinar da Argélia**</EM>.

[4] http://pt.wikipedia.org/wiki/ISO_4217

Finally, a small number of corrections have been made to the rules that capture intervals, such as "o portátil custa entre 800 e 1000 euros" (the laptop costs between 800 and 1000 euros), and rules have also been created for capturing leaps, such as "o preço subiu de 20 para 40 euros" (the price rose from 20 to 40 euros).

In the first case, the system was not always correctly capturing the interval, especially if the number had been written with spaces on it (e.g. 20 300 instead of 20.300 or 20300); it was also not setting the `interval` feature to the expression. Furthermore, it is common to see sentences like "a taxa de desemprego situa-se entre os 9 e os 10%" (the unemployment rate is between the 9 and the 10%), which, because of the definite articles, was not being captured as an interval. That has also been corrected.

In the case of leap values, the system lacked the ability to capture and classify expressions such as "the price of something changed from X euros to Y euros", so a rule has been created to include these patterns. Even though this problem has been discovered in the context of solving the currency issues, the rule has a much broader scope and it will match other units besides time units. In particular, it will have impact on time measures, which are the base for time NER.

Appendix C.1.2.2 presents some examples of correct and incorrect annotations for this type.

3.3.4.3 CLASSIFICATION type

`CLASSIFICATION` is a domain-specific type of the `AMOUNT` category because it deals almost only with expressions related to sporting events. In particular, this type encompasses a small number of NEs, which were already being captured and correctly classified by the system under the `AMOUNT` category but were not detached as a type of their own. This type includes expressions like "ele chegou em primeiro lugar" (he arrived in fist place), "ela ficou na primeira posição" (she was in first position), "eles foram os segundos classificados" (they were the runners-up), etc. The constitution of a separate type intended to isolate a class of expressions with a clear-cut structure, lexical features and syntactic behavior, clearly different from other `AMOUNT` types.

Besides the major modification of excluding results of sporting events from this type (more about this can be seen in Appendix C.1.2.4), the only modifications that have been made to this type are:

- Instead of capturing expressions such as "primeiro lugar" (first place) by means of a rule that matches an ordinal number followed by "lugar", the system now captures these expressions, and others, such as "em primeiro lugar" (in first place), "em primeira posição" (in first position), or "o primeiro classificado" (the first ranked), by means of a generic lemma. So, for example, the lemma for "em primeiro/segundo/terceiro/etc lugar" is `"em enésimo lugar"` (in N^{th} place)[5]. Similarly, the lemma for "em primeira/segunda/terceira/etc posição" is `"em enésimo posição"` (in N^{th} position). As a result, there has been a major restructuring of the rules that capture these expressions. Rules have also been created to capture and classify expressions involving some specialized adverbs such as "chegámos em primeiro lugar *ex-equo*" (we arrived in first place *ex-equo*).

[5]The idiom "em enésimo lugar", however, is not captured by the rule, since its meaning ("in the last place, very badly classified") is not compositional.

- Similarly to what has been done to QUANTITY and CURRENCY, this type has also suffered changes in the delimitation criteria for NEs. Prepositions introducing a PP are now also included in the NE, as in (for example) the sentence "Ele chegou <EM>em 1º lugar</EM>" (he arrived <EM>in 1^{st} place</EM>).

Special care has been taken, however, to not hinder precision with too broad rules, so that these types of expressions are not captured blindly and marked imprecisely as AMOUNT (type CLASSIFICATION). Consider, for example, the following sentence: "Em primeiro lugar, gostaria de dizer que está calor lá fora" (First, I would like to say that it is hot outside); in this case, "em primeiro lugar" (first) is a sentential modifier[6] and it should not be marked as CLASSIFICATION. For this reason, these expressions are only captured and marked when preceded or followed by certain verbs, particularly those related to sporting events. Below are some examples of expressions that are accepted by the system (the NE is presented in its lemma):

- Estar/ficar/chegar/partir **em enésimo lugar / em enésima posição** (ex-equo);
- Começar/iniciar/acabar/terminar/continuar/colocado/classificado **em enésimo lugar / em enésima posição**;
- Encontrar-se **em enésimo lugar / em enésima posição**;
- **Em enésimo lugar / Em enésima posição**, está/ficou ...
- Obter/assegurar/manter/ganhar/alcançar/ocupar **o enésimo lugar / a enésima posição**;
- Descer/Subir **para o enésimo lugar / para a enésima posição**;
- Descender/Ascender/Subir **ao enésimo lugar / à enésima posição**;
- **O enésimo lugar / A enésima posição** coube a/vai para ...
- Ser **o enésimo classificado**;
- **O enésimo classificado** foi ...

Appendix C.1.2.3 presents some examples of correct and incorrect annotations for this type.

3.3.4.4 SPORTS RESULTS type

HAREM's directives did not provide a clear separation between what is a sports result and what is not, because they were all being marked as CLASSIFICATION. While this is not entirely incorrect, because the CLASSIFICATION type includes expressions such as "em primeiro lugar" (in first place), which is somehow related to a competition of some sort, sports (as an activity) was deemed as an important enough domain so to have a specific type, considered separately. Moreover, the NEs of some sporting results have specific formats, which are completely unrelated to any other AMOUNT NE (e.g. football

[6] It makes part of the large class of sentential-modifying adverbs (Costa [14], Molinier & Lévrier [33], Palma [38]).

scores and tennis sets, presented in the form `3-0`, `5-3` and the like), therefore should be considered separately.

Consequently, a new type has been created under the `AMOUNT` category. Like `CLASSIFICATION`, `SPORTS RESULTS` is a domain-specific NE type. This task required several modifications in at least 4 different parts of the system: the list of features, the grammar rules, the dependency rules and the XML tags. In order to create this new type, it was necessary to gain a better understanding of how the system works, especially in terms of the processing flow.

Besides the rules that have been created in order to establish this new type, there are now a total of 14 grammar rules that deal with expressions such as the ones listed below. Notice that words between parentheses are optional, and as such, these rules apply to many sports: "ganhar 2-0" (winning 2-0) is not bound to any sport in particular.

- Ganhar (ao Real Madrid) (por) **2-0**;
- Ganhar (ao Real Madrid) **por 2 (bolas) a 0**;
- Vencer/Derrotar (o Real Madrid) (por) **2-0**;
- Vencer/Derrotar (o Real Madrid) **por 2 (bolas) a 0**;
- Perder (com o Real Madrid) (por) **2-0**;
- Perder (com o Real Madrid) **por 2 (bolas) a 0**;
- Empatar (com o Real Madrid) (**a**/**por**) **2-2**/**2 igual**;
- Empatar (com o Real Madrid) **a 2 bolas**;
- O empate (com o Real Madrid) **a**/**por 1-1**/**1 bola**;
- Ganhar/Vencer/Derrotar (ao / o Real Madrid) **por uma diferença de 2 bolas** / **2 golos**;
- Perder (com o Real Madrid) **por uma diferença de 2 bolas** / **2 golos**;
- Fazer (**o**) **2-0**;
- O jogo estava/acabou **2-2**/**2 igual**.

Appendix C.1.2.4 presents some examples of correct and incorrect annotations for this type.

3.3.5 HUMAN category

As a consequence of the new set of directives, the `HUMAN` category has been created (see Appendix C.2), encompassing the `PERSON` and `ORGANIZATION` categories of the HAREM's directives. The reasons for this shift have been presented above, in Section 3.3.3. The scope of this new category has been greatly improved, not only at the lexical level (which is important to have high recall), but also at the rule level (chunking and dependency rules). The following sections present the improvements that have been made in this category: first in the `INDIVIDUAL` type (Section 3.3.5.1), and also in the `COLLECTIVE` type (Section 3.3.5.2).

3.3.5.1 INDIVIDUAL type

The improvements that have been made to the `INDIVIDUAL` type are threefold:

- The restructuring of the way the names of people are captured and classified;
- The change in the delimitation of some NEs;
- The insertion of a large number of new lexical entries;

One of the system's major problems regarding this type was the imprecise capture of people's names, in particular complex ones, such as "José António dos Santos Faria de Macedo e Castro Domingues". The strategy that was being used was to mark the start and end of a name through the use of features: `start_people` and `end_people`, respectively. This strategy, however, was somewhat limited because more complex names have prepositions followed by articles ("de", "da", "das", "do", "dos") and conjunctions ("e") and, as a result, the resulting tree would consist of several separate nodes (e.g. "José António", "dos Santos Faria", "de Macedo", "Castro Domingues") instead of one node. Since XIP does not allow the partial classification of a node, but only of the whole node itself, these names would be classified separately, and not as an unique name.

In order to correct this, we have improved the strategy: instead of simply marking the start and end of a name, we now also mark the places where a name continues: the prepositions and conjunctions. We have done this by adding a third feature: `cont_people` (from "continue" people). Consequently, we are now able to produce a single node that represents the whole name, even if it is long and complex (see Figure 3.3).

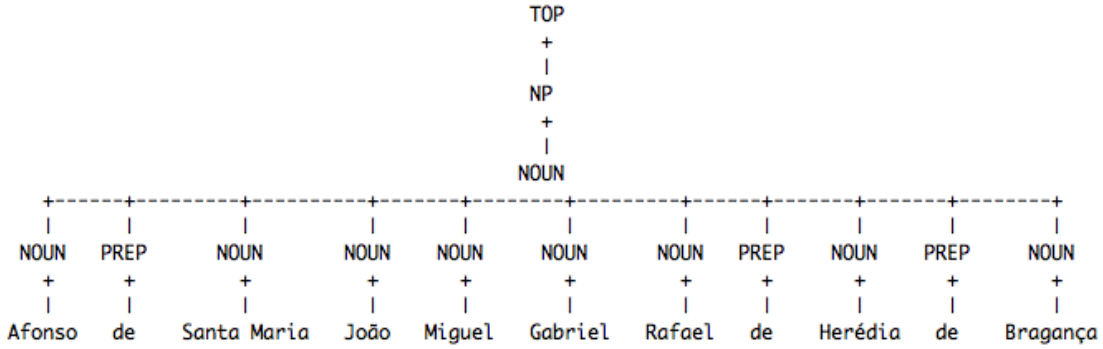

Figure 3.3: XIP: output tree for a complex proper name.

This process has been done in two steps:

1. We created a rule in the local grammar file that is responsible for this type of NE. This rule does not create a node; instead, it captures desirable sequences of nodes and sets the features where it is appropriate. For example, this rule might capture the name "Diogo Correia de Oliveira" and set the `start_people` feature to "Diogo", the `end_people` to "Correia" ("Diogo Correia" may be the person's full name), the `cont_people` feature to "de" and finally the `end_people` feature to "Oliveira";

2. At a later stage, another file is responsible for detecting the features and for creating the final node.

By doing this separately, we ensure that the process is flexible. Imagine that we want to capture the name "Filipa de Lencastre": while the first rule guarantees that the features are well set, the second rule creates the node. However, if we wanted to capture "Rainha D. Filipa de Lencastre", the first rule still applies, but there is a different rule to create the node: one that, in this case, contemplates jobs, positions or titles, such as "Rainha" (Queen) and "Duquesa" (Duchess), among others.

Similarly to what has been done to the delimitation of `QUANTITY` NEs (see Section 3.3.4.1), the delimitation of `HUMAN INDIVIDUAL` NEs has also been altered through dependency rules. Among the most important changes is the delimitation of kinship degrees: for example, "tio João" (uncle John) is now a NE, instead of simply "João" (John). These changes can be further consulted by referring to Appendix C.2.1.

Although aimed at processing Portuguese texts, the system also needs to deal with some foreign NEs, mainly people's names, because they are frequently used in Portuguese texts in any subject domain: art, sports, politics, economy, etc.

Before this task began, the system had only 2.973 lexical entries, each one representing an individual proper name. These names were almost only first names, since there were only 207 (approximately 7%) surnames on the list. Moreover, these surnames were only of Portuguese (141) and Chinese (66) origins. Although the system already had several foreign names listed as lexical entries, it was not enough to capture and correctly classify most foreign NEs. After some quick tests, we detected several "known" NEs that were not being captured, such as "John Wayne", "Tom Cruise", "Robert Redford", among others.

By having searched online for lists of common first names and surnames of different origins, namely Portuguese (including Brazillian), Spanish, French, German, Italian, English (including Welsh, Scottish, Irish, American and Canadian), Dutch, Belgian, Danish, Norwegian, Swedish, Finnish, Polish, Jewish, Chinese, Japanese and Muslim, a total of 3.272 *new* lexical entries were inserted, having increased the total amount of lexical entries from 2.973 to 6.245 (an improvement by a factor of 2,1).

Each lexical entry is assigned either the `firstname` or the `lastname` feature (or both, in cases of names that work as first names and surnames, such as "Vincent", "Stewart" or "Nelson", among many others). This task was important because the rules that exist in order to capture and classify people's names use the `firstname` and `lastname` features as a means to detect which nouns belong to a name and which do not. Therefore, by having merely added these entries, these rules now capture twice as many names as before.

Besides adding these entries to the lexicon files, it was also necessary to add them to the Palavroso's dictionary, otherwise they would not be recognized in the early stages of the processing chain. Whenever Palavroso finds an unknown word, it tries to guess if it is a noun, a verb, an adjective or an adverb using different heuristics. So, if a proper noun word is not in Palavroso's dictionary, it is likely that it will it be incorrectly classified.

By having added such a large number of lexical entries to XIP and, simultaneously, to Palavroso's dictionary, the system is now able to detect more than 6000 first names and surnames from 23 different

languages.

An obvious problem that results from operating under these terms, i.e., from having to create exhaustive lists of words, is that many of them are bound to be ambiguous. In this case, extreme care was necessary not to include surnames like "Cork" (a city in the Republic of Ireland) or "Thames" (the English river), among many others.

3.3.5.2 COLLECTIVE type

The HUMAN COLLECTIVE type corresponds roughly to the ORGANIZATION category that already existed in HAREM's directives. Those made the distinction between companies and institutions as types of organizations for- and non-profit, respectively. However, to make this (linguistically unmotivated) distinction, even if relevant for many practical purposes, and to do so at such an early stage of the IR/IE task, may be imprudent. Consequently, the COMPANY type was eliminated and the INSTITUTION type was expanded. According to this new set of directives, there are no distinctions between companies and institutions: they are all institutions and the system is not concerned about whether they are for- or non-profit. Appendix C.2.2.3 presents several examples regarding INSTITUTION NEs.

The ADMINISTRATION and INSTITUTION subtypes have been improved in three ways:

- All rules concerning them have been updated in order to comply with the new set of directives;
- Corrections have been made at the level of classification of NEs;
- New rules have been created to capture more NEs.

The first step towards eliminating the COMPANY type and converting it into INSTITUTION was to change its features in the appropriate places. For example, the following rule captured NEs such as "Grupo Sonae" (Sonae Group):

```
1> NOUN[org=+, company=+] @= ?[lemma:grupo, maj], ?+[maj].
```

Since organizations are now collective NEs, and companies are now institutions, this rule has been changed to:

```
1> NOUN[collective=+, institution=+] @= ?[lemma:grupo, maj], ?+[maj].
```

However, this is not enough. Afterwards, the rules that classify NEs had to be adapted:

```
| NP{?*, noun#1[collective, institution]}
   ; PP{?*, noun#1[collective, institution]} |
      if (~NE[collective,institution](#1))
         NE[collective=+, institution=+](#1)
```

On another level, the files of the local grammars that are responsible for treating these types have been scrutinized in order to detect cases of misclassification. This has proven to be important because there were in fact several classification errors, such as the one below, which classified "governo de Portugal (ou outro local qualquer)" (government of Portugal (or any other location)) as HUMAN COLLECTIVE INSTITUTION, when the correct classification is HUMAN COLLECTIVE ADMINISTRATION:

```
1> NOUN[org=+, institution=+] @= ?[lemma:governo, maj], prep[lemma:de],
                                 (art), ?[location,maj].
```

Rules have also been created to capture NEs that were not being captured yet, such as the names of known Portuguese radio stations (which are now institutions): Rádio Regional (de *local*); Rádio Comercial (da Linha); Antena 1/2/3; M80 (*local*); 97 FM Pombal; Romântica FM, Cidade FM, Best Rock FM, Horizonte Açores FM.

Obviously, these merely serve as examples and the actual rules capture many more cases. Consider, for example, the rule used to match the last example; it matches any expression comprised of at least one upper case word followed by "FM" (or "fm"):

```
1> NOUN[collective=+, institution=+] @= ?+[maj], ?[surface:FM];
                                        ?[surface:fm].
```

The same rationale was used to capture "97 FM Pombal":

```
1> NOUN[collective=+, institution=+] @= num[dig,frac:~,time:~],
                                       ?[surface:FM];?[surface:fm],
                                       ?*[location,maj].
```

Let us now focus on the third subtype of this category: `GROUP`. The purpose of this subtype, for the moment, is to include non-descriptive proper names, such as the names of musical groups, e.g. "U2", "Queen", etc. In the future, however, it is expectable that this subtype will grow and it will encompass more NEs besides these domain-specific names.

The names of musical groups are nouns that cannot be generalized through grammar rules, nor can they be guessed by some other means. With no possibility of searching online for the meaning of a word, the only way to capture them is through lexicon files (gazetteers), i.e. through a comprehensive list that contains the names of musical groups. With the help of two websites that are dedicated to this domain[7], a total of 7.621 *new* lexical entries were created, which contain musical groups from all over the world, covering all genres.

The only NEs of this subtype that are currently captured and classified by the system are the names of musical groups. There are not yet any grammatical rules to capture these NEs using, for example, the context of a sentence.

3.3.6 LOCATION category

The `LOCATION` category has been improved both in terms of vocabulary and in terms of grammatical rules. The vocabulary that has been added or corrected is mainly related to the `CREATED` and `PHYSICAL` types. Below there is a non-exhaustive list of examples of vocabulary that has been added or corrected within this category:

- Abbreviations of types of streets and other address elements that did not exist and have been added, like "pç." ("praça", square) and "R." ("Rua", Street);

[7] http://www.sing365.com/ and http://pt.wikipedia.org/wiki/Anexo:Lista_de_bandas_de_Portugal

- Words that had the wrong features and have been corrected: many oronyms like "montanha" (mountain), "planície" (plain), "planalto" (plateau) and many hydronyms like "estuário" (estuary), "foz" (mouth), "delta" (id.), but also words that describe administrative areas, like "aldeia" (village), "bairro" (neighborhood), "concelho" (county), and general buildings, like "aeroporto" (airport), "estação" (station) and "porto" (harbor);
- Many countries have also been added, for example: "Emirados Árabes Unidos" (United Arab Emirates), "Ilhas Salomão" (Solomon Islands), "República da Serra Leoa" (Republic of Sierra Leone), "Djibouti" (id.) and "El Salvador" (id.), as well as cities, such as "Lewes" and "Totnes" (in England) and "Los Angeles" (in the USA), just to name a few;
- A list of the most famous rivers around the world has been added, having obtained a total of 133 new lexical entries;
- Finally, many entries have been added representing alternative spellings for countries and cities, like "Bahrein" and "Bareine" (for Bahrain), "Bangladexe" (for Bangladesh), "Burúndi" (for Burundi), "Quatar" and "Katar" (for Qatar), "Iémene" and "Iémen" (for Yemen), and also "Kuweit", "Kuaite", "Coveite", "Couaite", "Cuaite" (for Kuwait), among many others.

With respect to the grammatical rules, most of them were inconsistent with the new set of rules provided by RuDriCo, so they have all been corrected, as explained in Section 3.3.2. Moreover, several new rules have been created in order to capture new NEs. The following are mere examples and do not represent the full extent of the new rules:

- "O nordeste brasileiro" (the Brazillian Northeast), "o sul de França" (the south of France), "o deserto do Sahara" (the Sahara desert);
- "Oceano Glaciar Árctico/Antárctico" (Arctic/Antarctic Ocean), "Oceano Austral" (Southern Ocean);
- "Campo dos Mártires da Pátria", "Campo Pequeno", "Campo Grande", "Campo de Ourique" (all of them are locations in Lisbon, Portugal);
- "O Decreto-Lei n.º 35/10" (The Decree-Law number 35/10), "o Regulamento geral" (the general Regulation);

Finally, the delimitation of some NEs had to be changed according to the new set of directives. For example, NEs like "cidade de Lisboa" (city of Lisbon), "rio Tejo" (Tagus river), "ilha da Madeira" (Madeira island) and "arquipélago dos Açores" (Azores archipelago), among others, now include the noun (and preposition/article) introducing the entity, unlike what happened in HAREM.

3.3.7 Metonymy

3.3.7.1 Introduction

Metonymy (literally, "change of name") is a figure of speech that designates the substitution of a noun for another noun, usually the two having a part-whole (or "metonymical") relation between them (Laus-

berg [24]). For example, "suit" for "business executive" or "tracks" for "horse races". This figure is related to (and sometimes distinguished from) the figures of "antonomasia" ("different name") and "synecdoche"[8] ("simultaneous understanding"), but here we will conflate those distinctions in the general definition given above.

Hence, for example, in a sentence such as "Portugal votou a favor desta moção na ONU" (Portugal voted in favor of this motion at the UN) the noun "Portugal" does not designate the country as a geographical location, but the human collective, the member state of the international organization. On the other hand, in "Portugal não conseguiu chegar aos quartos de final no Mundial de Futebol de 2010" (Portugal did not manage to get to the quarter finals in the 2010 Football World Cup), the interpretation of "Portugal" is now different, for the name of the country is used instead of the team of players representing that same country. On both cases, the name "Portugal" is not being used as a geographical entity, as we could find in "Vivo em Portugal desde 1986" (I live in Portugal since 1986).

Metonymy plays a very important role in natural language in general (Lakoff & Johnson [23]; Stefanowitsch & Gries [48]), especially from a cognitive and discourse point of view, and has tremendous impact in NER as well as in other NLP tasks. While humans can easily grasp metonymy and the sense(s) it conveys because of their linguistic and extra-linguistic knowledge, computers still lack the ability to capture the metonymical relation between an overt, explicit word and another unsaid, underlying noun.

Many studies have been dedicated to metonymy in the scope of NLP (Kiyota *et al.* [22]; Shutova [47]; Peirsman [40]), to cite a few, and here we will focus on the issues pertaining to NER in order to present a (partial) solution to the problem integrated in the L^2F's NLP chain (Mamede [26]).

3.3.7.2 Formalizing the metonymy relation

As metonymy hinges on a semantic relation between two nouns (an explicit, overt denomination, and an implicit, hidden name), it is only natural that the NER task explicitly takes into consideration these two poles of the figure of speech. We adopt a very pragmatic approach to the issue, considering that for each denomination it is possible to define a basic, literal (non metonymical) classification. Thus, "Portugal" is considered, before anything else, the name of the country, and a country is a particular (sub)type of a `LOCATION` NE, more precisely `LOCATION CREATED COUNTRY`, having, among other distributional properties, the ability to be a locative complement, replaceable by interrogative pronouns such as "where?" and the like. When used metonymically, its new distribution corresponds to another class of NEs (say `HUMAN COLLECTIVE ADMINISTRATION`). In the NE tag, the two classes are then presented and the metonymical class is introduced by the `MET-CAT` tag: see Appendix C.4 for examples.

3.3.7.3 Capturing metonymy

In order to (partially) capture metonymy, first it is necessary to identify the distributional shift from the literal to the figurative context of the NE. Next, it is also necessary to determine if the new context is such that a metonymical relation can be inferred between the two denominations; alternatively, if that

[8] See http://en.wikipedia.org/wiki/Synecdoche for a succinct overview.

information is available, it is necessary to determine if the two denominations are usually metonymically related.

Hence, in a sentence such as: "Portugal come muito peixe" (Portugal eats a lot of fish), the subject of the verb is constraint so that ordinarily locatives could not appear in that position. Thus, it is possible to calculate that the NE is being used figuratively (not as a LOCATION) and to attribute it the features of the new context. However, as "comer" (to eat) allows for both human and non-human subjects, extra-linguistic knowledge must be invoked, that is, a general rule that attributes the HUMAN feature to a COUNTRY NE:

- <EM MET-CAT="HUMANO COLECTIVO" CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAÍS">**Portugal**</EM> come muito peixe.

The metonymical classification, however, cannot always be deepened much further unless more information is available.

In another case, "Um GNR passou uma multa ao João" (A GNR=policeman gave João a ticket) there is a shift of the name of an institution (HUMAN COLLECTIVE INSTITUTION) to a HUMAN INDIVIDUAL context:

- Um <EM MET-CAT="HUMANO INDIVIDUAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**GNR**</EM> passou uma multa ao João.

It is not possible to further classify the metonymical use of the NE, since it does not comply with the definitions of any of the HUMAN INDIVIDUAL subtypes.

So far, distributional information is scarce in the system. Lists of human-activity verbs have been compiled and the system currently has a total number of 665 verbs of this kind. Also, at this stage the system is only capable of dealing with three kinds of distributional shifts: LOCATION to HUMAN, HUMAN COLLECTIVE to HUMAN INDIVIDUAL and HUMAN COLLECTIVE to LOCATION. For more information on these, please refer to Appendix C.4.

3.3.7.4 Improvements in metonymy

As already stated in Section 3.3.7.2, one of the improvements that has been made in the course of this thesis has been to enhance the detection and classification of metonymy and to provide a clearer way of presenting it through the inclusion of the MET-CAT tag. Before this was implemented, there were many cases of metonymy that were being captured, but they were spread over several categories. As a result, metonymy passed almost unnoticed in the system, because it was not being explicitly represented.

Several things had to be implemented in order to improve metonymy in this system, namely:

- Features had to be created to support the new cases of metonymy;
- New rules for marking entities had to be created;
- The list of XML tags had to be updated.

Currently, the rules that exist in the system to treat metonymy with the MET-CAT tag are based on the relations between the various elements of the sentence (relations-based rules). Unlike syntax-based rules[9], relations-based rules grasp the (deeper) connection between the sentence's elements. Consider, for example, the sentence: "o Brasil, irritado com Portugal, não votou a favor da moção" (Brazil, angry with Portugal, did not vote in favor of the motion). In this case, it is important to capture and treat the SUBJ relation between "Brasil" (Brazil) and "votar" (to vote), because a syntax-based rule will surely not match this sentence, due to the separation that exists between the subject and the verb. The unpredictability of how the sentence is constructed makes it impractical to treat metonymy in this way.

Before this thesis, most of the existing rules were based on the relations between the elements of the sentence, but as already stated, they were spread over several categories and were not being treated evenly, thus making metonymy hard to detect. Moreover, all NEs were being classified according to the HAREM's directives.

Therefore, on the one hand, we have tried to convert the rules that already existed in order to comply with the new set of directives; on the other hand, we have created new rules for dealing with metonymy, which were syntax-based in the beginning, but that have been gradually converted into relations-based rules.

In order to better explain the work that has been done, consider the following examples, which will be used throughout the remainder of this section. They represent the three shifts that were treated in this thesis: LOCATION to HUMAN COLLECTIVE, HUMAN COLLECTIVE to HUMAN INDIVIDUAL and HUMAN COLLECTIVE to LOCATION, respectively:

- "Portugal constatou que a crise estava forte" (Portugal realized that the crisis was strong);
- "O GNR afirmou que eu ia ser multado por excesso de velocidade" (The GNR=policeman said that I was going to get a ticket for speeding);
- "Esse artigo foi publicado no Diário de Notícias" (That article was published in the Diário de Notícias).

In the first example, the name of a country is used with a verb that typically selects a human subject (countries do not realize things, people do). As such, in this case "Portugal" refers to the Portuguese people and there needs to be a feature to represent this shift in category: let us call it met-country-to-coll (country to collective). In order to create a dependency rule to treat this case (and many more, for that is the advantage of using relations-based rules instead of syntax-based rules), the most important thing to note is that this will happen every time a LOCATION NE is used as subject of a human-activity verb:

```
if ( ^NE[location,admin_area,country](#1) & SUBJ(?[human_activity],#1) )
                        NE[met-country-to-coll=+](#1)
```

At this point, it is prudent to make an aside to explain exactly how this rule operates. Essentially, we start by saying that if there is a LOCATION CREATED COUNTRY NE, then we mark it as variable number one: #1. Afterwards, there needs to be a subject (SUBJ) relation between a human-activity verb

[9]The interpretation of syntax, in this case, is related to the immediate constituents of the sentence.

(`?[human_activity]`) and the NE itself. Finally, in case this happens, we set the metonymical feature to the dependency. It is important to distinguish features that are set to dependencies from features that are set to nodes. In this case, we are dealing with the former. As already stated in Section 3.3.7.2, "Portugal" is, before anything else, a `COUNTRY`, so we do not want to lose that information. The node itself will keep the features that mark it as a `COUNTRY`, and it will be classified as a `COUNTRY` NE (i.e., there will be a NE dependency with the `location`, `admin_area` and `country` features). Afterwards, we simply add another feature (the metonymical feature) to the NE dependency[10], which is what will change the entity's final classification.

These rules offer flexibility, since in order to adapt the rule above to divisions (cities), regions (continents), or to any other type, we only need to replace the features for the appropriate ones. At this stage, and regarding the `LOCATION` category, the system can only handle shifts from `LOCATION CREATED COUNTRY`, `LOCATION CREATED DIVISION` and `LOCATION CREATED REGION` to `HUMAN COLLECTIVE`.

The second example is very similar to the first one, with the exception that the shift is now within the same category (`HUMAN`). The concept is, however, the same: whenever there is a `HUMAN COLLECTIVE` NE that is subject of a human-activity verb, then we mark the NE with a metonymical feature that represents the shift:

```
if ( ^NE[collective,institution](#1) & SUBJ(?[human_activity],#1) )
                        NE[met-inst-to-ind=+](#1)
```

Currently, the system can only handle two shifts from `HUMAN COLLECTIVE` to `HUMAN INDIVIDUAL`, which are from `HUMAN COLLECTIVE ADMINISTRATION` and `HUMAN COLLECTIVE INSTITUTION`.

The third and final example represents a shift from `HUMAN COLLECTIVE` to `LOCATION`:

```
if ( ^NE[collective,institution](#1) & ( MOD[post](?[lemma:publicar],#1) |
    MOD[post](?[lemma:escrever],#1) | MOD[post](?[lemma:redigir],#1) |
        MOD[post](?[lemma:presente],#1) ) & PREPD(#1,?[lemma:em]) )
                        NE[met-inst-to-loc=+](#1)
```

In this case, we used the disjunction operator (`|`) to represent several possibilities, namely: "esse artigo foi *publicado* no Diário de Notícias" (that article was *published* in Diário de Notícias), "esse artigo foi *escrito* no Diário de Notícias" (that article was *written* in Diário de Notícias), etc. The idea, however, is simple: whenever one of these verbs precedes the preposition "em", which in turn precedes a `HUMAN COLLECTIVE` NE, then the NE is being used (metonymically) as a location (in this case, a place for information, such as a newspaper).

Obviously, these three rules merely represent examples of how metonymy is being treated in the system. The actual extent of cases the system is able to cover is much larger. Table 3.4 presents a non-exhaustive list of examples the system is able to handle. The words in bold are the ones that convey a metonymical sense, and words in parentheses are optional.

[10]The circumflex operator is placed immediately before the dependency that is to be altered.

Shift	Examples
LOCATION to COLLECTIVE	**Lisboa** invadiu o Porto. **Montenegro** obteve a independência da **Sérvia** em 2006. José Sócrates lidera **Portugal**. A ameaça da **Coreia do Norte**. **Portugal** ficou chateado com a morte de António Feio. O acordo assinado entre **Portugal** e **Espanha**. As relações diplomáticas entre os **EUA** e o **Iraque** estão pelas ruas da amargura. Cristiano Ronaldo conquistou a admiração da **Inglaterra**.
COLLECTIVE to INDIVIDUAL	O **GNR** constatou que eu ia em excesso de velocidade. A **Associação de Futebol de Lisboa** deliberou ... O **Governo de Portugal** confirmou ...
COLLECTIVE to LOCATION	Esse artigo foi publicado no **Diário de Notícias**. Li um artigo no **Jornal de Notícias**. No ano passado estive na **Google**. Ele apareceu na **SIC**. A minha música preferida passou na **M80**.

Table 3.4: Metonymy: list of examples the system is able to handle (all shifts).

Chapter 4

Evaluation

THIS chapter presents the evaluation of the system's performance in the NER task for the categories here studied. Section 4.1 serves as a contextualization of the problem; Section 4.2 presents a detailed view of all the evaluation metrics that have been used, and, finally, Section 4.3 presents the results that have been obtained.

4.1 Context

The typical evaluation methodology followed in this kind of NLP task consists of processing a large amount of non annotated corpus with the system that is to be evaluated, and afterwards compare the answers with the (previously) annotated version of the same corpus. This annotation is usually performed by hand, over a significant period of time, by one or more linguists, who use a set of directives to guide their work.

From the beginning, this study intended to use HAREM's evaluation programs not only because the organization provides them freely on their website[1], but also because we had already used them in the past, so we were familiar with their use. However, since they had been written with an established set of directives in mind, these programs had to be slightly altered in order to support the new set of directives. In particular, they had to be adapted to the new definitions for the category `HUMAN`, the new type (`SPORTS_RESULTS`, under the `AMOUNT` category) and subtype (`GROUP`, under the category `HUMAN`, `COLLECTIVE` type), and finally to the annotations of metonymy.

The evaluation corpus (henceforth called "Golden Collection" (GC)) is the same that was used in the HAREM evaluation campaign (see Section 2.1), but it was re-annotated following the new set of directives, instead of the directives of HAREM. For all intents and purposes, this new GC has two versions: one in which there are entities marked for metonymy, and another in which metonymy is left out. This was meant to evaluate the impact that metonymy induces in our system.

[1] They can be downloaded at `http://www.linguateca.pt`, Section "Avaliação Conjunta", "HAREM".

4.2 Evaluation metrics

4.2.1 Golden Collection

The GC is a XML file comprised of several documents, where each document is written between <DOC> </DOC> tags. XIP is able to process these tags as if they were comments in a programming language, i.e., they are not evaluated. Each document pertains to a specific text domain or topic; for example, there may be documents with texts related to sports and other documents with texts related to politics. They are used to organize the texts within the GC file. Below the documents, the next immediate unit of organization in the hierarchy of the texts is the paragraph. Each paragraph is delimited by the <P> </P> tags and it may contain one or more sentences, and each sentence may contain (or not) one or more named entities.

The GC consists of 129 documents, 2.274 paragraphs and 5.569 named entities (out of which 283 are metonymy NEs).

Another important element of the GC is the <ALT> </ALT> tags, which represents alternative (ALT) segmentations for a specific NE. When this happens, the alternatives are expressed with the following syntax:

```
<ALT> NE segmentation 1 | NE segmentation 2  |... </ALT>
```

The original GC used in the HAREM evaluation campaign contained 411 ALTs. These were reconsidered while the GC was annotated and most of them were removed, which was a decision of the linguist who annotated the corpus. The only ALTs that currently remain are the ones involving the categories relevant to this study, namely to the AMOUNT, HUMAN and LOCATION categories, which add up to a total of 49 ALTs.

4.2.2 Cornerstones

Before analyzing the results, let it first be defined the most important concepts involved in the evaluation; namely:

- Identification of NEs;
- Classification of NEs;
- The possible states for a NE: correct, missing or spurious;
- Precision, Recall and F-measure.

By assessing the identification, one is interested in knowing how good a system is at locating NEs in the text, especially by indicating *where* they start and *where* they end. So, for example, in the sentence "O bilhete custou 10 euros" (The ticket cost 10 euros), the identification of NEs is concerned with capturing (and only capturing) "10 euros":

O bilhete custou <EM>**10 euros**</EM>**.**

A very different concept is the classification of NEs. By asking "how good is a system at classifying entities?", one is interested in knowing how well the system is able to place a specific NE into the right category. Obviously, the "right" category depends on established criteria, i.e., the classification directives. In the previous example, and according to the classification directives here used (see Appendix C.1.2.2), all monetary quantities must be classified as "VALOR" (AMOUNT) "MOEDA" (CURRENCY), so in this case the expected output is:

```
O bilhete custou <EM CATEG="VALOR" TIPO="MOEDA">10 euros</EM>.
```

As far as the evaluation programs are concerned, a named entity can be either correct, missing, or spurious. A NE is *correct* if it is equal to the one in the GC: for identification purposes, this means that the NE must contain the same elements; for classification purposes, this means that it was accorded the same category, type and subtype. A NE is *missing* if there is a NE in the GC but the system fails to correctly detect any of its elements. Finally, a NE is *spurious* if the system marks a NE that does not exist in the GC.

It is important to stress that, unlike what happened in the First HAREM evaluation campaign, in this work (and in the Second HAREM evaluation campaign as well) there are no partially correct entities. Even if a system produces a NE that partially matches the one in the GC, the NE will be discarded and marked as spurious. The corresponding NE as it appears in the GC will be marked as missing. This decision makes it harder for a system to obtain better results, because instead of earning scores for partially finding an entity, it penalizes the system for not being 100% precise and may largely increase the number of spurious named entities.

The three metrics that have been used both in the HAREM evaluation campaign and in this thesis are *precision*, *recall* and *F-measure*[2]. *Precision* is a measure of the system's response quality and it measures the amount of right answers among all answers given by the system. *Recall*, on the other hand, measures the percentage of solutions (in this case, contained in the GC) that a system can retrieve; or, in other words, the amount of right answers among all possible answers (not only those given by the system). Finally, *F-measure* combines precision and recall according to the following mathematical formula:

$$\text{F-measure}\ (\%) = \frac{2 \times \ \text{Precision} \times \ \text{Recall}}{\text{Precision} \ + \ \text{Recall}} \times 100.$$

As an example, if a system has a precision of 90% but a recall of only 50%, its F-measure result is calculated as:

$$\text{F-measure} \ = \frac{2 \times 0.9 \times 0.5}{0.9 + 0.5} \times 100 \approx 64.286\%.$$

Precision and recall, however, are not calculated in the same way for identification and classification purposes. Although the idea is the same, the values used in one are different from the ones used in the other. Regarding the identification task, they are both calculated by directly using the number of entities that were obtained: for example, if there are 5000 named entities in the GC and the system is able to identify 4000, out of which 3000 are correct, then the precision and recall are:

[2]In the HAREM campaigns, two more metrics were used, over-generation and sub-generation. As these correspond to the inverse of precision and recall, respectively, we considered them redundant and did not use them here.

$$\text{Precision} = \frac{3000}{4000} \times 100 = 75\% \qquad \text{Recall} = \frac{3000}{5000} \times 100 = 60\%.$$

For classification purposes, though, the calculation is more complex. Given two NEs, one produced by the system and the other present in the GC, the classification's main objective is to compare all three attributes (category, type and subtype) between the two entities. The annotation scheme that was adopted in HAREM defines a four level hierarchy: identification of the NE (by simply marking it with the `<EM>` `</EM>` tags), and filling the remaining three attributes.

It is enough to correctly identify the entity to receive one point. The total score is obtained by adding this value to the one obtained in the classification. Each attribute has a specific weight: the default values are 1.0 for the category, 0.5 for the type and 0.25 for the subtype. This means that we consider the category to be the most important attribute to classify correctly. Immediately follows the type and finally the subtype[3]. There is a final weight associated to an entity: it is 1.0 for correct entities and 0.0 for missing or spurious ones. In case of an ALT with N alternatives, the weight for each NE inside the ALT is $\frac{1}{N}$ (see Section 4.2.3 for a detailed view of the different kinds of evaluation scenarios there are).

To conclude, consider this example: after all calculations, the maximum possible score in the GC for classification is 10,000, and the maximum possible score for the system's classification is 9000. However, the actual classification of the system is 6000. This means that:

$$\text{Precision} = \frac{6000}{9000} \times 100 = 66.7\% \qquad \text{Recall} = \frac{6000}{10000} \times 100 = 60\%.$$

4.2.3 Evaluation scenarios

4.2.3.1 Global and Selective scenarios

There are two types of scenarios in the evaluation of the NER task: global and selective scenarios. Whereas the former consists of evaluating all categories, the latter consists of evaluating a subset of the categories. As already stated in Section 2.10, selective scenarios have the disadvantage of not allowing the direct comparison between systems that participated in different scenarios. However, in this particular case this will not be a problem because we are not going to compare our results to the results of other systems. Moreover, we cannot even compare these results to the ones obtained in HAREM because the directives have been changed and the GC was re-annotated accordingly, and also because the set of categories of the selective scenarios adopted in the Second HAREM is different from the categories under study here.

HAREM's evaluation programs allow the systems to test different scenarios because they provide a way for filtering the entities. The following syntax is used:

```
CATEGORY1(TYPE1{SUBTYPE1,SUBTYPE2,...};TYPE2{...};...):CATEGORY2(...)
```

In order to only evaluate the `AMOUNT` category (and its types), for example, we indicate:

```
VALOR(QUANTIDADE;MOEDA;CLASSIFICACAO;RESULTADOS_DESPORTIVOS)
```

[3]These values can be parameterized.

Similarly, to only evaluate the HUMAN category, INDIVIDUAL type, PERSON subtype, we indicate:

HUMANO(INDIVIDUAL{PESSOA})

The evaluation programs remove all entities not pertaining to the ones desired and (obviously) decrease the total amount of named entities used for comparison, so that the results are not compromised.

In this study, the following scenarios have been defined:

- Without metonymy:
  - Selective: only AMOUNT;
  - Selective: only HUMAN;
  - Selective: only LOCATION;
  - Global: AMOUNT, HUMAN and LOCATION;
- With metonymy:
  - Selective: only metonymy, i.e., the evaluation will be solely focused on the entities that have been marked for metonymy;
  - Global: AMOUNT, HUMAN and LOCATION.

In this way, one will clearly see how well the system performs in each category and also how strong is the impact of metonymy in the results.

4.2.3.2 Strict ALT and Relaxed ALT

As already stated in Section 4.2.1, the ALT tag is used to annotate all possible segmentations of a particular named entity. There are two ways of evaluating them:

- Strict ALT evaluation: each alternative inside the ALT tags is accounted for. In this case, the system will only have the maximum possible score if it shows all alternatives;
- Relaxed ALT evaluation: only the ALT's element that maximizes the system's classification is chosen. This option usually entails better results than the previous one.

It may happen that Strict and Relaxed ALT scenarios produce the exact same results for a particular category, because there may not be any alternative segmentation for any entity of that particular category.

4.3 Evaluation Results

This section of the document presents the results obtained by the system during the evaluation of the NER task. Sections 4.3.1 and 4.3.2 show the results for the evaluation without and with metonymy, respectively, both for the identification and the classification tasks.

4.3.1 Scenarios without metonymy

Table 4.1, below, shows the results for the identification task. Notice that the results of the AMOUNT category are equal in both scenarios. As already stated in Section 4.2.3.2, this is because there are no alternative segmentations in any entity of this particular category. Overall, the existence of only 49 ALTs induced very little difference in the results. The largest difference (adding up to only 1.2%) can be found in the HUMAN category.

		Total	Identified	C. Id	Spurious	Missing	P	R	F
Strict ALT	AMOUNT	546	852	367	485	179	**0.431**	**0.672**	**0.525**
	HUMAN	3046	2503	1565	938	1481	**0.625**	**0.514**	**0.564**
	LOCATION	1649.5	1249.5	1034.5	215	615	**0.828**	**0.627**	**0.714**
	GLOBAL	5204.5	4597.5	3051.5	1546	2153	**0.664**	**0.586**	**0.623**
Relaxed ALT	AMOUNT	546	852	367	485	179	**0.431**	**0.672**	**0.525**
	HUMAN	3047	2471	1573	898	1474	**0.637**	**0.516**	**0.570**
	LOCATION	1651	1236	1035	201	616	**0.837**	**0.627**	**0.717**
	GLOBAL	5215	4557	3067	1490	2148	**0.673**	**0.588**	**0.628**

Table 4.1: Results: evaluation without metonymy, identification task (C. Id: correctly identified; P: precision; R: recall; F: F-measure).

The AMOUNT category is the only one in which XIP identified more entities than those present in the GC (out of 546 entities in the GC, XIP identified 852: 367 were correct, but 485 were spurious). The reasons for this large number of spurious NEs differ in nature.

Some entities were just plainly misidentified. In the sentence "O vírus H5N1" (the H5N1 virus) the alphanumeric designation of the pathogen had not been previously contemplated in the tokenizer. Because of this, the system identified 4 separate tokens, and two of them were wrongly marked as AMOUNT QUANTITY. Because this is a domain-specific type of alphanumeric word form, and the general-purpose grammar only lists commonly occurring tokens of this type, the shortcoming of the tokenizer is corrected by updating the list of this type of word forms.

A more significant origin for apparently spurious entities involves nominal-numeral determiners. These types of numerals – e.g. "centena" (one hundred), "milhar" (one thousand), "milhão" (million), "dúzia" (dozen), and the like, are a particular set of determiners that, very much like measure units, operate on a noun by way of a linking preposition "de" (of): "Centenas de pessoas vieram aqui" (Hundreds of people came here). Ordinary numerals have, on the other hand, an adjective-like, prenominal syntactic behavior: "300 pessoas vieram aqui" (300 people came here).

In the L^2F/XIP grammar for Portuguese, a quantifier dependency is already extracted between these nominal-numeral determiners and the head of the immediately subsequent PP:

```
QUANTD(pessoas, centenas)
```

This enables the system to correctly identify the distributional subject (or other high order depen-

dency) of the main verb in the sentence (or clause), i.e. the "real" subject of "vieram" (came) is "pessoas" (people) and not "centenas" (hundred).

In the directives (see Section C.1.1), the delimitation of the AMOUNT NEs is defined as an entire NP or PP including the head noun even if this noun is not a measure unit. Therefore, and strictly speaking, the NE should consist of the chunk (NP or PP) whose head is the nominal-numeral determiner:

<EM>centenas</EM> de pessoas ... (1)

However, since the QUANTD dependency is already being extracted, a more sophisticated delimitation could also be envisaged, so that the NE would include the immediately subsequent PP:

<EM>centenas de pessoas</EM> ... (2)

Unfortunately, XIP can only extract NEs out of single nodes, i.e. it cannot extract the same NE out of two distinct nodes. So, at this stage, it is not possible to obtain that delimitation in a straightforward manner.

If one counts the delimitation illustrated in (1) as correct, the precision of the system regarding the identification task for NEs of the AMOUNT category improves from 43.1% to 51.3%, while the recall improves from 67.2% to 80.1%. The resulting F-measure then improves from 52.5% to 62.5%.

In contrast, the HUMAN category suffers from the inverse problem: a large number of missed NEs (recall = 51.6% in Relaxed ALT), but also an important quantity of spurious, i.e., misidentified NEs (36.3%).

These results, however, must be interpreted under the light of two main considerations: on the one hand, the very broad scope of this category, which includes not only proper names of people, organizations, institutions, groups, etc., but also job or office designations, titles and the like; on the other hand, its scope has been made broader than originally defined by the Second HAREM directives, since, under the new classification guidelines, each subtype/type was considerably enlarged, but also because HUMAN now encapsulates both PEOPLE and ORGANIZATION categories from the Second HAREM.

Because of the lexical basis of many of these types/subtypes, and in spite of the large effort put in the task of increasing the size of the lexical resources that deal with them, results are still under the thresholds attained in the Second HAREM by some systems. This low recall problem must therefore be addressed most urgently in the near future, even though it is well known that this category (and its types) are among the most difficult and challenging of the NER task.

The LOCATION category produced the best results in the identification task, with 82.8–83.7% precision and 62.7% recall, thus yielding an F-measure of 71.4–71.7%.

While results seem to show an improvement of the system in the identification of NEs of this category since its participation in the Second HAREM (even if they can not be directly compared), the cause for this still low recall can be found in the certain types of missed entities, such as region names like Badakhshan (Afghanistan) or general constructions like "Pavilhão das Grandes Indústrias" (Big Industries Pavilion), "Teatro Académico de Gil Vicente" (Gil Vicente's Academic Theater), etc. In contrast, XIP produced a small number of spurious NEs and after analyzing the results, it is possible to conclude that the majority of them are induced by delimitation errors caused by other categories. Again, the issue

is not whether XIP finds entities that do not exist, but that it lacks precision at delimiting them. Consequently, those entities are marked as spurious, e.g. "Associação Académica de Coimbra" (Coimbra's Academic Association) is marked as:

```
Associação Académica de <EM CATEG="LOCAL" TIPO="CRIADO"
                        SUBTIPO="DIVISAO">Coimbra</EM>
```

In reality, the whole expression should have been identified as HUMAN COLLECTIVE INSTITUTION. If it had, then by only evaluating the LOCATION category, this entity would have been removed from the scenario. Since it was not identified, this caused an identification error in the LOCATION category. In other words, an improvement in the delimitation of the HUMAN category will inevitably result in an improvement in the recall of the LOCATION category.

Finally, the GLOBAL scenario produced balanced results, with an overall precision of 66.4% in the Strict ALT scenario and 67.3% in the Relaxed ALT scenario. The recall was 58.6% and 58.8% respectively and the final F-measure was 62.3% and 62.8%. While not directly comparable, these results apparently indicate an improvement trend in the performance of the system since the Second HAREM. In order to improve these results in the future, there must be an effort to improve the general precision of the system and particularly the recall in the HUMAN category, which are the factors that currently decrease the overall score of the system. Table 4.2 below shows the results for the classification task without metonymy. These results are very similar to the ones obtained in the identification task without metonymy.

		GC's max	System's max	System's score	Max P	Max R	Max F
Strict ALT	AMOUNT	747.75	1171.5	500.125	**0.427**	**0.669**	**0.521**
	HUMAN	4236.750	3469.896	2138.771	**0.616**	**0.505**	**0.555**
	LOCATION	2520.416	1901.464	1556.958	**0.819**	**0.618**	**0.704**
	GLOBAL	10921.051	9597.125	6250.388	**0.651**	**0.572**	**0.609**
Relaxed ALT	AMOUNT	747.75	1171.5	500.125	**0.427**	**0.669**	**0.521**
	HUMAN	4238.163	3425.875	2149.792	**0.628**	**0.507**	**0.561**
	LOCATION	2522.716	1880.758	1557.725	**0.828**	**0.617**	**0.707**
	GLOBAL	10944.034	9513.384	6282.975	**0.660**	**0.574**	**0.614**

Table 4.2: Results: evaluation without metonymy, classification task (Max P: maximum precision; Max R: maximum recal; Max F: maximum F-measure).

The AMOUNT category remains the one with most problems, which in this case is shown by the system's maximum score (1171.5 out of a possible 747.75). This is a direct consequence of the large number of spurious entities: with more entities to analyze, the system's maximum score increases to the point where it may even exceed the GC's maximum possible score, which in this case does. As a consequence, the precision is also drastically reduced (42.7%).

Globally, the results are satisfactory, especially regarding recall: 57.2% in the Strict ALT scenario and 57.4% in the Relaxed ALT scenario. This means that out of 10 possible solutions in the GC, XIP is able

to retrieve almost 6. Simultaneously, out of 10 given answers, almost 7 are correct (65.1% and 66.0% of precision). The factors that continue to diminish the final results are the precision of the AMOUNT category and the recall of the HUMAN category.

4.3.2 Scenarios with metonymy

Table 4.3 below shows the system's results in the identification task for two scenarios with metonymy: in the first one, only the metonymical annotations are considered; in the second, the system is analyzed globally, with all categories considered, including metonymy.

		Total	Identified	C. Id	Spurious	Missing	P	R	F
S. ALT	Metonymy	283	109	41	68	242	**0.376**	**0.145**	**0.209**
	GLOBAL	5490.5	4707.5	3094.5	1613	2396	**0.657**	**0.564**	**0.607**
R. ALT	Metonymy	283	109	41	68	242	**0.376**	**0.145**	**0.209**
	GLOBAL	5501	4667	3110	1557	2391	**0.666**	**0.565**	**0.612**

Table 4.3: Results: evaluation with metonymy, identification task (C. Id: correctly identified; P: precision; R: recall; F: F-measure).

Once again, there are no significant differences between the Strict and Relaxed ALT scenarios with regard to the Metonymy evaluation. The results are poor, but it was to be expected. As already explained in Section 3.3.7.4, the most important aspect of metonymy that we wanted to address in this study was to treat it evenly, i.e., to represent it in a clear and unambiguous way. This goal has been achieved by the inclusion of the MET-CAT tag. Improving metonymy as a whole is a different issue altogether. Although some situations have been covered by the addition of new rules, and also by the increase on the number of human-activity verbs, there is still much work to be done. It is a hard subject to work with and it requires a lot of time, which unfortunately was not possible to spend.

Globally, the identification task with metonymy presents (as expected) worse results than without metonymy, even if the difference seems small, because of the limited number of cases affected by metonymy and here treated. Everything needs to be improved: from the amount of spurious entities (68 out of 109, which results in a 37.6% precision) to the amount of missed entities (242 out of 283, which results in a recall of only 14.5%), there is still much room for improvement.

Table 4.4 below shows the results of the system in the classification task with metonymy. Like in the previous table, this task presents the same problems: a very low recall (15.6%), but a higher precision (40.0%).

These numbers show that metonymy, for the moment, is still not a positive factor in the performance of the system. Its impact is very small because of the limited number of metonymic NEs present in the GC. However, it is likely that the inclusion of other types of metonymic shifts in the "tool kit" of tropologic phenomena dealt with by the system may enable it to achieve higher accuracy in the classification task and other NER dependent tasks, such as semantic analysis.

		GC's max	System's max	System's score	Max P	Max R	Max F
S. ALT	Metonymy	491.063	191.438	76.563	**0.400**	**0.156**	**0.224**
	GLOBAL	11093.373	9732.019	6301.937	**0.648**	**0.568**	**0.605**
R. ALT	Metonymy	491.063	191.438	76.563	**0.400**	**0.156**	**0.224**
	GLOBAL	11116.615	9646.748	6334.897	**0.657**	**0.570**	**0.610**

Table 4.4: Results: evaluation with metonymy, classification task (Max P: maximum precision; Max R: maximum recall; Max F: maximum F-measure).

To conclude this evaluation, below are four charts that sum up the most important results. As a matter of convenience to the reader, the results were rounded to the units.

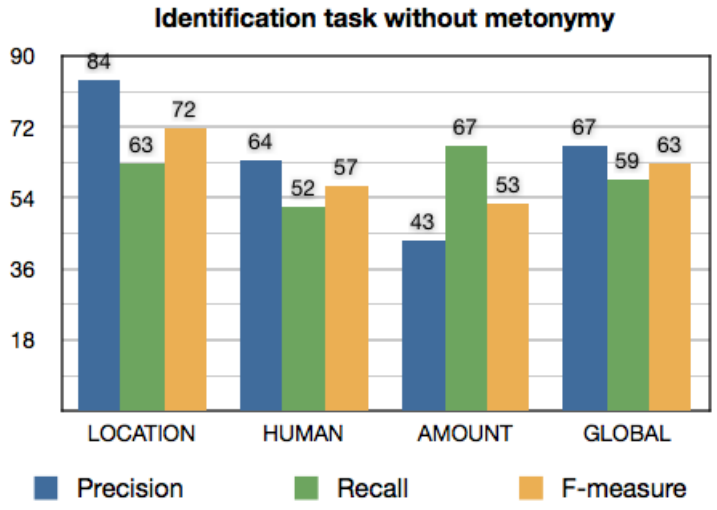

Figure 4.1: Results: chart from Relaxed ALT, identification, without metonymy.

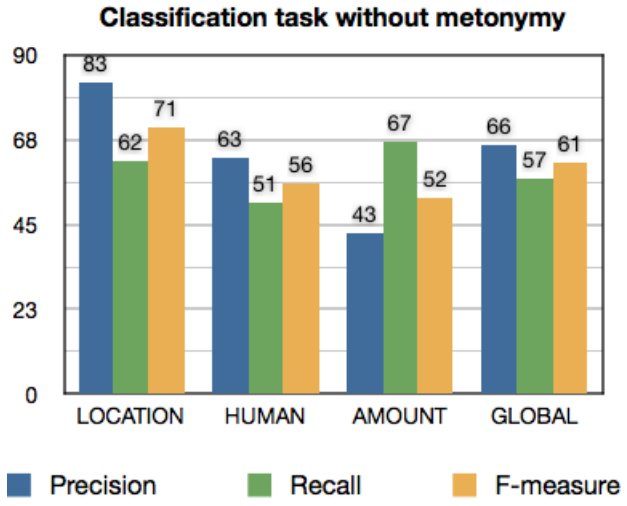

Figure 4.2: Results: chart from Relaxed ALT, classification, without metonymy.

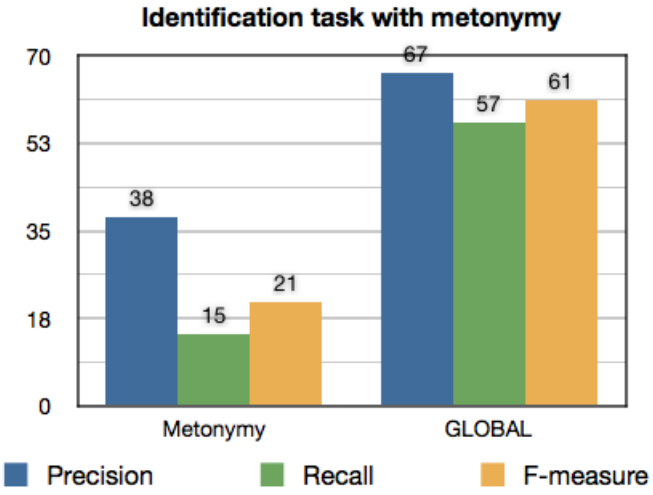

Figure 4.3: Results: chart from Relaxed ALT, identification, with metonymy.

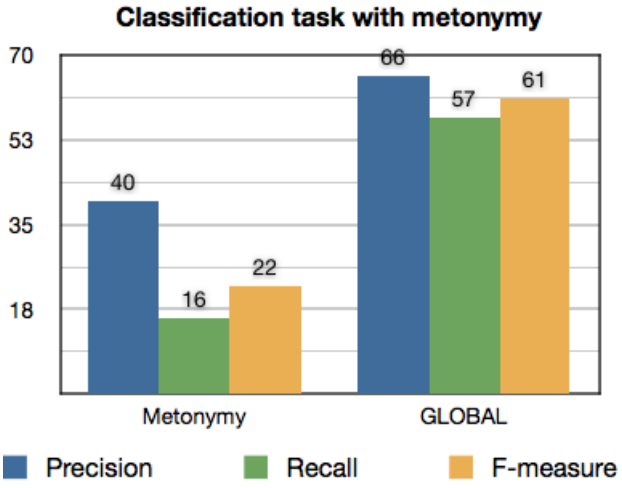

Figure 4.4: Results: chart from Relaxed ALT, classification, with metonymy.

Chapter 5

Conclusions

THIS final chapter presents a brief summary of the main aspects of this study, along with some final remarks, in order to conclude with prospective work, deemed as most urgent, to be implemented in the near future.

5.1 Final remarks

Named Entity Recognition is one of the most known tasks in Natural Language Processing. This study aimed at improving the performance of a NLP system developed at L^2F/INESC-ID Lisboa, by developing the NER modules responsible for the identification and classification of NEs from the HUMAN (INDIVIDUAL and COLLECTIVE), LOCATION and AMOUNT categories.

Chapter 2 carried out a comparison of the 8 systems that took part in the Second HAREM evaluation campaign. This evaluation initiative was a collective effort of the Portuguese NLP community aimed at assessing the performance of those different systems in the complex task of identifying and subsequently classifying named entities in Portuguese texts.

Different ways of addressing the NER task were shown: from automatic, statistically-based approaches, to rule-based systems. An overview of the main linguistic resources for the NER task was done, namely ontologies, grammars, lexicons, among others. Each system was described in detail and the results were compared.

In Chapter 3, one of these systems was analyzed in more detail: XIP, a language-independent parser, which takes textual input and provides linguistic information about it. The NLP chain, in which XIP is inserted, was described, covering its three main stages: pre-processing, disambiguation and syntactic rules. Then, the XIP's main characteristics were detailed: features, chunking rules, dependency rules, disambiguation rules, lexicons and local grammars.

The last section of Chapter 3 presented the improvements made to XIP during this study. In particular, this section showed how each category was improved either by adding more lexical entries or by correcting/adding chunking/dependency rules.

The need for developing a new set of classification directives was discussed in some depth. In fact, these new guidelines, specially developed during this study with the aim of replacing the Second

HAREM directives, constitute a major contribution of this thesis and are presented in full in Appendix C.

Finally, Chapter 4 presented the evaluation of the NLP chain in the NER task after all these improvements were introduced. First, the main concepts involved in NER evaluation were briefly presented: the difference between identification and classification; the three possible states of a named entity after processing (correct, missing or spurious); and finally, the three evaluation measures: precision, recall and F-measure.

The evaluation itself was organized according to different scenarios in order to assess precisely the performance of the system for each category independently and to compare the impact of metonymy in the NER task as a whole; a global scenario was also considered; furthermore, in each scenario both the identification and the classification were tested separately.

The overall results were satisfactory, particularly for the LOCATION category, where above 70% F-measure was attained, both in the identification and classification tasks.

Even if the results can not be directly compared with those from the Second HAREM, since the directives are different (and, consequently, even though the corpus is the same, it is annotated differently), it is possible to say that the main objective of this thesis has been achieved: results seem to show a general trend of improvement.

5.2 Future work

In the following lines, different venues for future work are presented so that the NER module of the L^2F/XIP system might still be further improved:

- The AMOUNT category must still receive special attention, particularly towards a higher precision. Existing rules must be refined and, most likely, new rules must be devised in order to avoid an excessive number of spurious NEs;

- The pre-processing script that deals with number expressions still requires some work in order to support the detection of fractional numbers such as "quatro/4 e meio" (four and a half). Also, the standard nominal-numeral determiners, e.g. "dezenas" (dozens), are currently supported only up to 999 999; larger number expressions, such as "dezenas de milhão" (dozens of million), for example, are not yet detected by this particular script. The delimitation of NEs involving these types of nominal-numeral determiners should be made less syntactically dependent (i.e. chunking) and more semantically oriented, in order to capture as NE, in a similar way, both adjective-like and nominal-numerals. For the time being, "300/trezentos livros" (three hundred books) and "3/três centenas de livros" (three hundreds of books) are captured differently, the latter leaving out the "real", distributional head of the complex syntagma;

- The HUMAN category, being of such a broad scope, still requires a much larger lexical coverage, especially in the COLLECTIVE type: names of institutions, companies and the like are urgently needed to be added to the lexicons of the system in order to significantly improve its recall;

- The feature propagation module still only works for `HUMAN INDIVIDUAL PERSON` named entities. Ideally, this method should have a broader scope and be applied to the remaining categories;
- Metonymy is now clearly and consistently expressed throughout the system by way of a XML tag. Most of the already existing rules have been adapted and many new rules have been created in order to allow the system a smooth integration of this powerful rhethorical device, so very common in language and with such a profound impact in the NER task. However, much still needs improvement. In particular, the number of human-activity verbs and the types of metonymical shifts that the system is meant to capture are still scarce. More importantly, rules devised to treat metonymy should use features assigned to the predicative, central elements of the sentences (like the human-activity verbs here used). In this way, a clear separation of the lexicon from the metonymic context is guaranteed and the system would gain flexibility.

Bibliography

[1] ALONSO, OMAR; GERTZ, MICHAEL & BAEZA-YATES, RICARDO. 2007. On the value of temporal information in information retrieval. *SIGIR Forum*, **41**(2), 35–41.

[2] AMARAL, CARLOS; LAURENT, DOMINIQUE; MARTINS, ANDRÉ; MENDES, AFONSO & PINTO, CLÁUDIA. 2004a. Design and implementation of a semantic search engine for Portuguese. *Pages 247–250 of:* LINO, MARIA TERESA; XAVIER, MARIA FRANCISCA; FERREIRA, FÁTIMA; COSTA, RUTE & SILVA, RAQUEL (eds), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, vol. 1.

[3] AMARAL, CARLOS; FIGUEIRA, HELENA; MENDES, AFONSO; MENDES, PEDRO; PINTO, CLÁUDIA & INFORMÁTICA, PRIBERAM. 2004b. A Workbench for Developing Natural Language Processing Tools. *Pages 1–2 of: In Pre-proceedings of the 1st Workshop on International Proofing Tools and Language Technologies (Patras, Greece)*.

[4] AMARAL, CARLOS; FIGUEIRA, HELENA; MENDES, ANDRÉ MARTINS AFONSO; MENDES, PEDRO & PINTO, CLÁUDIA. 2005. Priberam's question answering system for Portuguese. *In: Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005) (Vienna, Austria)*.

[5] AMARAL, CARLOS; FIGUEIRA, HELENA; MENDES, AFONSO; MENDES, PEDRO; PINTO, CLÁUDIA & VEIGA, TIAGO. 2008. Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM. *Chap. 9, pages 171–180 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[6] AÏT-MOKHTAR, SALAH; CHANOD, JEAN-PIERRE & ROUX, CLAUDE. 2001. *XIP Tutorial for Grammar Development*. Xerox Research Centre Europe.

[7] BICK, ECKHARD. 2007. Functional aspects on Portuguese NER. *Chap. 12, pages 145–155 of:* SANTOS, DIANA & CARDOSO, NUNO (eds), *Reconhecimento de entidades mencionadas em português – Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Digitally published.

[8] BORBINHA, JOSÉ LUÍS; PEDROSA, GILBERTO; REIS, DIOGO; LUZIO, JOÃO; MARTINS, BRUNO; GIL, JOÃO & FREIRE, NUNO. 2007. DIGMAP - Discovering our past world with digitised maps. *Pages 563–566 of:* KOVÁCS, LÁSZLÓ; FUHR, NORBERT & MEGHINI, CARLO (eds), *Research and advanced technology for digital libraries*. Berlin, Heidelberg: Springer Verlag, for 11th European Conference, ECDL 2007, Budapest, Hungary, September 2007, Proceedings.

[9] CARDOSO, NUNO. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. *Chap. 11, pages 195–211 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[10] CARVALHO, PAULA; OLIVEIRA, HUGO GONÇALO; SANTOS, DIANA; FREITAS, CLÁUDIA & MOTA, CRISTINA. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. *Chap. 1, pages 11–31 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[11] CHAVES, MARCIRIO SILVEIRA. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. *Chap. 13, pages 231–245 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[12] CHAVES, MARCIRIO SILVEIRA; MARTINS, BRUNO & SILVA, MÁRIO J. 2005. *GKB – Geographic Knowledge Base*. Technical Report 05–12. Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.

[13] COLLINS, MICHAEL & SINGER, YORAM. 1999. Unsupervised models for named entity classification. *Pages 100–110 of:* FUNG, PASCALE & ZHOU, JOE (eds), *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

[14] COSTA, JOÃO. 2009. *O Advérbio em Português Europeu*. Lisboa: Edições Colibri.

[15] CRAVEIRO, OLGA; MACEDO, JOAQUIM & MADEIRA, HENRIQUE. 2008. PorTexTO: sistema de anotação/extracção de expressões temporais. *Chap. 8, pages 159–170 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[16] DINIZ, CLÁUDIO. 2010. *RuDriCo 2 - A converter based on declarative transformation rules*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[17] FERREIRA, LILIANA; TEIXEIRA, ANTÓNIO & DA SILVA CUNHA, JOÃO PAULO. 2008. REMMA - Reconhecimento de Entidades Mencionadas do MedAlert. *Chap. 12, pages 213–229 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[18] HAGÈGE, CAROLINE; BAPTISTA, JORGE & MAMEDE, NUNO. 2008a. Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o Segundo HAREM. *Pages 289–308 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[19] HAGÈGE, CAROLINE; BAPTISTA, JORGE & MAMEDE, NUNO. 2008b. Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre a Xerox e o L2F do INESC-ID Lisboa. *Chap. 15, pages*

261–274 of: MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[20] JOHNSON, STEPHEN CURTIS. 1975. *YACC: Yet Another Compiler Compiler*. Technical Report. AT & T, Bell Laboratories, Murray Hill, New Jersey.

[21] JURAFSKY, DANIEL & MARTIN, JAMES H. 2008. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2 edn. Prentice Hall.

[22] KIYOTA, YOJI; KUROHASHI, SADAO & KIDO, FUYUKO. 2005. Resolution of Modifier-Head Relation Gaps Using Automatically Extracted Metonymic Expressions. *Pages 367–376 of:* SU, KEH-YIH; TSUJII, JUN'ICHI; LEE, JONG-HYEOK & KWONG, OI YEE (eds), *Natural Language Processing – IJCNLP 2004*. Lecture Notes in Computer Science, vol. 3248. Springer, Berlin / Heidelberg.

[23] LAKOFF, GEORGE & JOHNSON, MARK. 1980. *Metaphors We Live By*. University of Chicago Press.

[24] LAUSBERG, HEINRICH. 1982. *Elementos de Retórica Literária*. 3 edn. Lisboa: Fundação Calouste Gulbenkian. Translation, preface and additions by R. M. Rosado Fernandes.

[25] LOUREIRO, JOÃO. 2007. *Reconhecimento de Entidades Mencionadas (Obra, Valor, Relações de Parentesco e Tempo) e Normalização de Expressões Temporais*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[26] MAMEDE, NUNO. 2007. *XIP: the new text processing workflow*. Technical Report. Laboratório de Sistemas de Língua Falada (L^2F), INESC-ID Lisboa.

[27] MAMEDE, NUNO. 2009. *A cadeia de processamento de Língua Natural do L^2F (em Dezembro de 2009)*. Technical Report. Laboratório de Sistemas de Língua Falada (L^2F), INESC-ID Lisboa.

[28] MAMEDE, NUNO; BAPTISTA, JORGE & HAGÈGE, CAROLINE. 2009. *Nomenclature of chunks and dependencies in Portuguese XIP grammar 2.1*. Technical Report. Laboratório de Sistemas de Língua Falada (L^2F), INESC-ID Lisboa.

[29] MARTINS, BRUNO. 2008. O sistema CaGE no Segundo HAREM. *Chap. 7, pages 149–158 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[30] MARTINS, BRUNO. 2009. *Geographically aware Web text mining*. Ph.D. thesis, Faculdade de Ciências, Universidade de Lisboa.

[31] MARTINS, BRUNO; MANGUINHAS, HUGO & BORBINHA, JOSÉ LUÍS. 2008. Extracting and exploring the geo-temporal semantics of textual resources. *Pages 1–9 of: Proceedings of the 2nd IEEE International Conference on Semantic Computing (ICSC 2008), August 4-7, 2008, Santa Clara, California, USA*. IEEE Computer Society.

[32] MEDEIROS, JOSÉ CARLOS. 1995. *Processamento Morfológico e Correcção Ortográfica do Português*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[33] MOLINIER, CHRISTIAN & LÉVRIER, FRANÇOISE. 2000. *Grammaire des adverbes: description des formes en -ment*. Genève: Droz.

[34] MOTA, CRISTINA. 2008. R3M, uma participação minimalista no Segundo HAREM. *Chap. 10, pages 181–193 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[35] MOTA, CRISTINA. 2009. *How to keep up with language dynamics: A case study on named entity recognition*. Ph.D. thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[36] MOTA, CRISTINA; OLIVEIRA, HUGO GONÇALO; SANTOS, DIANA; CARVALHO, PAULA & FREITAS, CLÁUDIA. 2008. Resumo de resultados do Segundo HAREM. *Pages 379–403 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[37] OLIVEIRA, HUGO GONÇALO; MOTA, CRISTINA; FREITAS, CLÁUDIA; SANTOS, DIANA & CARVALHO, PAULA. 2008. Avaliação à medida do Segundo HAREM. *Chap. 5, pages 97–129 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[38] PALMA, CRISTINA. 2009. *Estudo contrastivo Português-Espanhol de expressões fixas adverbiais*. Master thesis, Universidade do Algarve.

[39] PARDAL, JOANA PAULO. 2007. *Manual do Utilizador do RuDriCo*. Technical Report. Laboratório de Sistemas de Língua Falada (L^2F), INESC-ID Lisboa.

[40] PEIRSMAN, YVES. 2006. Example-based metonymy recognition for proper nouns. *Pages 71–78 of: In Proceedings of the Student Research Workshop of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*.

[41] POIBEAU, THIERRY & KOSSEIM, LEILA. 2000. Proper Name Extraction from Non-Journalistic Texts. *Language and Computers, Computational Linguistics in the Netherlands*, 144–157.

[42] RIBEIRO, RICARDO DANIEL; OLIVEIRA, LUÍS C. & TRANCOSO, ISABEL. 2003. Using Morphosyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. *Pages 143–150 of: PROPOR'2003 - 6th Workshop on Computational Processing of the Portuguese Language*. Heidelberg: Springer-Verlag.

[43] ROMÃO, LUÍS. 2007. *Reconhecimento de Entidades Mencionadas em Língua Portuguesa: Locais, Pessoas, Organizações e Acontecimentos*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[44] SANTOS, DANIEL. 2010. *Identification of relationships between entities*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[45] SANTOS, DIANA; CARVALHO, PAULA; FREITAS, CLÁUDIA & OLIVEIRA, HUGO GONÇALO. 2008. Segundo HAREM: Directivas de anotação. *Pages 277–287 of:* MOTA, CRISTINA & SANTOS, DIANA (eds), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Digitally published.

[46] SCHMID, HELMUT. 1995. *TreeTagger, a language independent part-of-speech tagger*. Technical Report. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[47] SHUTOVA, EKATERINA. 2009. Sense-based interpretation of logical metonymy using a statistical method. *Pages 1–9 of: ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*. Morristown, NJ, USA: Association for Computational Linguistics.

[48] STEFANOWITSCH, ANATOL & GRIES, STEFAN T. (eds). 2006. *Corpus-Based Approaches to Metaphor and Metonymy*. Trends in Linguistics. Studies and Monographs 171. Berlin: De Gruyter Mouton.

[49] XEROX. 2003. *Xerox Incremental Parser – XIP Reference Guide*. Xerox Research Centre Europe.

Appendix A

Classification results

The systems that took part in the HAREM evaluation campaign are shown in the columns of Table A.1 (except for R3M, which did not participate in the classification task). Each line represents a different category and the results attained by the systems regarding the evaluation measures of Precision (P), Recall (R) and F-measure (F). These results correspond to the classification task in a strict ALT evaluation scenario (defined in Section 2.10). The highest score for each measure is shown in bold. For example, in the ABSTRACTION category Priberam had the highest recall and F-measure, but not the highest precision, having been surpassed by REMMA.

		System						
		CaGE	PorTexTO	Priberam	REMBRANDT	REMMA	SEI-Geo	XIP
Abstr.	**P**			0.1099	0.1956	**0.2231**		
	R			**0.5132**	0.1433	0.0392		
	F			**0.1810**	0.1655	0.0667		
Amount	**P**			0.1055	**0.4127**	0.3589		0.3100
	R			0.7099	**0.7176**	0.5202		0.6558
	F			0.1836	**0.5241**	0.4247		0.4209
Event	**P**			0.0668	0.5630	0.4044		**0.7250**
	R			**0.4327**	0.2026	0.1473		0.1897
	F			0.1158	0.2980	0.2159		**0.3007**
Location	**P**	0.5319		0.3471	0.5484	0.5700	**0.6800**	0.6770
	R	0.5527		**0.6816**	0.6607	0.5089	0.5138	0.5047
	F	0.5421		0.4599	**0.5993**	0.5377	0.5853	0.5783
Org.	**P**	0.3425		0.2277	0.5350	**0.5829**		0.5136
	R	0.1851		**0.5010**	0.3231	0.2397		0.2625
	F	0.2403		0.3131	**0.4029**	0.3397		0.3475
Person	**P**	0.4136		0.4712	**0.7683**	0.6666		0.7122
	R	0.2851		**0.7157**	0.5368	0.3677		0.5332
	F	0.3376		0.5682	**0.6320**	0.4740		0.6099
Thing	**P**			0.0762	0.0451	**0.2227**		
	R			**0.3899**	0.0566	0.0318		
	F			**0.1275**	0.0502	0.0557		
Time	**P**	0.0823	0.6694	0.0832	0.5904	0.4744		**0.6812**
	R	0.0294	0.5419	0.1826	0.4030	0.2538		**0.7314**
	F	0.0434	0.5990	0.1143	0.4790	0.3307		**0.7054**
Work	**P**			0.0798	**0.5251**	0.5146		0.4670
	R			**0.3323**	0.2171	0.1212		0.0847
	F			0.1287	**0.3072**	0.1962		0.1434

Table A.1: State of the Art: Comparison (results from the classification task, strict ALT. P: Precision; R: Recall; F: F-measure).

Appendix B

POS categories

The grammar assumes the existence of the following POS categories:

Name	Category	Examples
Adj	adjective	espirituoso
Adv	adverb	ontem, amanhã
Art	article	o, a
Conj	conjunction	e, ou
Foreign	foreign word	court
Interj	interjection	ui
Noun	common or proper noun	nariz, Manuel
Num	numeral	vinte, 1290, primeiro
Pastpart	past participle	amado, lavado, sido
Prep	preposition	em, para, com
Pron	pronoun	ele, meu, este, algo
Punct	punctuation	`;`, `:`, `.`
Rel	relative pronoun	qual, que
Symbol	special symbol	$, %, #
Verb	verb	comi, andaram

Table B.1: XIP: list of POS categories (from Mamede *et al.* [28]).

Appendix C

Classification Directives

This document presents the classification directives developed and adopted for this thesis. Despite having been inspired by the HAREM evaluation campaign's directives, they differ in several essential aspects (see Section 3.3.3).

NER is the NLP task that focuses on locating and classifying entities in a given text. Ambiguity pervades natural language, which makes this a rather challenging task. In order to provide a clear and reproducible classification of NEs, this document presents a set of delimitation and classification directives developed and adopted for this thesis.

This document is organized as follows: Sections C.1, C.2 and C.3 address the three main categories: `AMOUNT`, `HUMAN` and `LOCATION`, respectively. Section C.4 addresses the metonymy problem and the methods used to solve it.

C.1 The AMOUNT category

The `AMOUNT` (Portuguese: VALOR) category is meant to capture several distinct types of entities appearing in texts with numeric quantifiers. The rationale behind this category in NER is to provide a simple IE procedure for many IR/IE tasks. Even if other types of non-numeric quantifiers may be present, the use of number words is mandatory for this type of NE.

C.1.1 Delimitation

The NE consists of the entire Noun Phrase (NP) or Prepositional Phrase (PP), including the head noun, e.g. "20 quilos" (20 kilos), even if this noun is not a measure unit, e.g. "20 laranjas" (20 oranges).

If the quantified expression is in a PP, the NE must also include the preposition introducing that PP, e.g. "O Pedro precisou de 20 laranjas" (Pedro needed of 20 oranges), "O barco estava a 50 milhas da costa" (The boat was at 50 miles from the coast).

C.1.2 AMOUNT types

At this stage, this general category comprises the following types: QUANTITY (Portuguese: QUANTIDADE), CURRENCY (Portuguese: MOEDA), CLASSIFICATION (Portuguese: CLASSIFICAÇÃO) and SPORTS RESULTS (Portuguese: RESULTADOS DESPORTIVOS). Whereas the first type deals with both absolute and relative quantities (QUANTITY), the CURRENCY type deals with expressions that designate money, and the CLASSIFICATION type is meant to capture normal ordinals, e.g. "Ele ficou em 1º lugar" (He took first place). The SPORTS RESULTS type is meant to encompass results from sporting events.

C.1.2.1 QUANTITY type

The QUANTITY type encompasses percentages, e.g. "10%", "10 por cento" (10 percent) and other fractional values, e.g. "1/2" (half), "três quartos" (three quarters); isolated numbers; determinative phrases involving measure units with a numeric quantifier (e.g. 200 g).

Since we are dealing with amounts, the NE must contain at least one number word; this may be a number formed by one or several digits ("3", "342"), including the fractional ("3,42") and thousand separators ("1.000") or it may be a number written in full ("three", "three hundred and forty-two").

If the measure unit is part of a determinative phrase, usually linked to the head noun by "de" (of), the entire NP or PP is to be captured, e.g. "500 g de manteiga" (500 g of butter), "dois decilitros de leite" (two deciliters of milk).

Expressions involving intervals of some sort must be considered as a whole, i.e. as a single NE (and NOT as two NEs, one for each part of the interval); e.g. "entre 10 e 20 laranjas" (between 10 and 20 oranges), "de 10 a 20 laranjas" (from 10 to 20 oranges).

The NE must include other eventual quantifiers on the number determinant, such as adverbs, e.g. "aproximadamente/cerca de/por volta de 20 laranjas" (approximately/about/around 20 oranges).

Correct annotations

- A taxa de desemprego é <EM CATEG="VALOR" TIPO="QUANTIDADE">**10%**</EM>.
- A taxa de desemprego é <EM CATEG="VALOR" TIPO="QUANTIDADE">**10 por cento**</EM>.
- Encontrei <EM CATEG="VALOR" TIPO="QUANTIDADE">**4 cães**</EM> na rua.
- O edifício tem <EM CATEG="VALOR" TIPO="QUANTIDADE">**150 metros**</EM> de altura.
- Javier Sotomayor saltou <EM CATEG="VALOR" TIPO="QUANTIDADE">**2,45 metros**</EM>, tendo batido o record do mundo do salto em altura.
- O camião pesa <EM CATEG="VALOR" TIPO="QUANTIDADE">**500 kg**</EM>.
- A frequência medida foi <EM CATEG="VALOR" TIPO="QUANTIDADE">**1,4 kHz**</EM>.
- A taxa de desemprego está <EM CATEG="VALOR" TIPO="QUANTIDADE">**entre 9 e 10%**</EM>.
- O barco estava <EM CATEG="VALOR" TIPO="QUANTIDADE">**a 50 milhas**</EM> da costa.

- O bife pesa <EM CATEG="VALOR" TIPO="QUANTIDADE">**cerca de 200 gramas**</EM>.
- O aumento salarial foi <EM CATEG="VALOR" TIPO="QUANTIDADE">**de menos de 10%**</EM>.

Incorrect annotations

For purposes of convenience to the reader, and in order to represent a delimitation error (not a classification one), we will write the misplaced words in brown. If the words are placed outside the NE delimitation tag, it means that they should have been included in it. Otherwise, it means that they should have been left out of it.

- A taxa de desemprego é <EM CATEG="VALOR" TIPO="QUANTIDADE">**10**</EM> **por cento**.
- Encontrei <EM CATEG="VALOR" TIPO="QUANTIDADE">**4**</EM> **cães** na rua.
- O edifício tem <EM CATEG="VALOR" TIPO="QUANTIDADE">**150**</EM> **metros** de altura.
- A frequência medida foi <EM CATEG="VALOR" TIPO="QUANTIDADE">**1,4**</EM> **kHz**.
- A taxa de desemprego está **entre** <EM CATEG="VALOR" TIPO="QUANTIDADE">**9**</EM> **e** <EM CATEG="VALOR" TIPO="QUANTIDADE">**10%**</EM>.
- O barco estava **a** <EM CATEG="VALOR" TIPO="QUANTIDADE">**50 milhas**</EM> da costa.
- O bife pesa **cerca de** <EM CATEG="VALOR" TIPO="QUANTIDADE">**200 gramas**</EM>.

C.1.2.2 CURRENCY type

This type is responsible for encompassing all monetary expressions; these include abbreviations such as USD, EUR or GBP, which may be combined with numbers (digits or not), as well as fully written expressions such as "10 dólares americanos" (10 american dollars) or "dez dólares do Canadá" (ten dollars from Canada). The currency unit itself must be included in the NE, as well as any prepositions or quantifiers related to other ways of describing the quantity (as it is also done to QUANTITY).

Generic references to currency ("the euro", "the american dollar", etc.) are not to be classified.

As in the QUANTITY type above, intervals such as "entre 10 e 20 euros" (between 10 and 20 euros) or "de 10 a 20 milhões de dólares" (from 10 to 20 million dollars), are also to be captured as a whole NE.

Correct annotations

- O casaco custa <EM CATEG="VALOR" TIPO="MOEDA">**200 euros**</EM>.
- O iPhone 4 vai custar <EM CATEG="VALOR" TIPO="MOEDA">**299 USD**</EM>.
- Gastei <EM CATEG="VALOR" TIPO="MOEDA">**mil pesos chilenos**</EM> no jantar.
- <EM CATEG="VALOR" TIPO="MOEDA">**1 dinar tunisino**</EM> está par a par <EM CATEG="VALOR" TIPO="MOEDA">**com 1 dinar da Argélia**</EM>.
- Isso custa <EM CATEG="VALOR" TIPO="MOEDA">**entre 5 e 10 euros**</EM>.

- As moedas de ouro valiam <EM CATEG="VALOR" TIPO="MOEDA">**mil**</EM>, <EM CATEG="VALOR" TIPO="MOEDA">**dois mil**</EM> e <EM CATEG="VALOR" TIPO="MOEDA">**4 mil euros**</EM>.

- As moedas de ouro valiam <EM CATEG="VALOR" TIPO="MOEDA">**1**</EM>, <EM CATEG="VALOR" TIPO="MOEDA">**2**</EM> e <EM CATEG="VALOR" TIPO="MOEDA">**4 mil euros**</EM>.

Regarding this last example, it is important to notice that despite being isolated numbers, "1" and "2" are not actually isolated quantities. So, instead of classifying them as AMOUNT QUANTITY, which would typically be their classification, in this case they ought to be classified as AMOUNT CURRENCY because they represent "1000" and "2000" euros, respectively. This is easier to detect in the example before the last, because instead of "1" and "2", the sentence indicates "mil" (thousand) and "dois mil" (two thousand), i.e. the "thousand" factor is not omitted.

Incorrect annotations

- O casaco custa <EM CATEG="VALOR" TIPO="MOEDA">**200**</EM> **euros**.

- O iPhone 4 vai custar <EM CATEG="VALOR" TIPO="MOEDA">**299**</EM> **USD**.

- <EM CATEG="VALOR" TIPO="MOEDA">**O peso chileno**</EM> valorizou-se.

- O <EM CATEG="VALOR" TIPO="MOEDA">**dinar tunisino**</EM> está par a par com o <EM CATEG="VALOR" TIPO="MOEDA">**dinar da Argélia**</EM>.

- As moedas de ouro valiam <EM CATEG="VALOR" TIPO="QUANTIDADE">**mil**</EM>, <EM CATEG="VALOR" TIPO="QUANTIDADE">**dois mil**</EM> e <EM CATEG="VALOR" TIPO="MOEDA">**4 mil euros**</EM>.

- As moedas de ouro valiam <EM CATEG="VALOR" TIPO="QUANTIDADE">**1**</EM>, <EM CATEG="VALOR" TIPO="QUANTIDADE">**2**</EM> e <EM CATEG="VALOR" TIPO="MOEDA">**4 mil euros**</EM>.

C.1.2.3 CLASSIFICATION type

The CLASSIFICATION type is meant to capture normal ordinal numbers, such as "1º" (1^{st}) or "primeiro" (first) when used to rank individual athletes in sporting competitions. As such, expressions such as "4.ª classe" (4^{th} grade) or "o 44.º Presidente dos EUA" (the 44^{th} President of the USA) must not be marked as AMOUNT CLASSIFICATION. Typical expressions to be marked are those related to a competition of some sort, for example, "Ele chegou em primeiro lugar" (He arrived in first place), "Ele ocupa a terceira posição" (He ranks third) or "Ele assegurou a segunda posição" (He secured the second position), just to name a few.

Correct annotations

- <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**1.º lugar**</EM> para a McLaren-Mercedes.
- Reprovei na 4.ª classe.
- <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**O primeiro lugar**</EM> coube a Usain Bolt.
- Lewis Hamilton assegurou <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**a segunda posição**</EM> na grelha de partida do GP do Mónaco.
- Na categoria de Infantis, ficaram <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**em 1.º lugar ex-equo**</EM> Carlos Miguel Salvado (Bandolim) e Patrícia Alexandra Marques (Acordeão);

Incorrect annotations

- O <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**primeiro lugar**</EM> coube a Usain Bolt.
- Reprovei na <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**4.ª**</EM> classe.
- O <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**44.º**</EM> Presidente dos EUA.

C.1.2.4 SPORTS RESULTS type

This type encompasses quantities that are related to results of sporting events. For example, football scores: 2–0 or "três bolas a zero" (three-nil); tennis and squash sets: 6–3, 3–2; among others.

The prepositions involved in the fully written form of these expressions should be included in the NE as well as adjectives such as "igual", e.g. "o jogo ficou 2 igual".

Correct annotations

- Portugal ganhou <EM CATEG="VALOR" TIPO="RESULTADOS_DESPORTIVOS">**7-0**</EM> à Coreia do Norte.
- Ao intervalo, o Benfica já estava a ganhar <EM CATEG="VALOR" TIPO="RESULTADOS_DESPORTIVOS">**por três bolas a zero**</EM>.
- O jogo acabou <EM CATEG="VALOR" TIPO="RESULTADOS_DESPORTIVOS">**2 igual**</EM>.

Incorrect annotations

- Portugal ganhou <EM CATEG="VALOR" TIPO="CLASSIFICACAO">**7-0**</EM> à Coreia do Norte.

C.2 The HUMAN category

The HUMAN (Portuguese: HUMANO) category is meant to capture several distinct types of entities referring to people, such as the names of an individual person, groups of people, jobs people have, among others; a NE belongs to this category if it can be the subject of verbs such as "pensar" (to think), "dizer" (to say), "considerar" (to consider), "afirmar" (to affirm), etc. Under this category, one distinguishes

between HUMAN INDIVIDUAL and HUMAN COLLECTIVE: whereas the former designates individual entities, the latter covers cases of collective organizations. This distinction is further explained in detail in Sections C.2.1 and C.2.2.

C.2.1 INDIVIDUAL type

The HUMAN INDIVIDUAL type comprises of the subtypes listed below.

1. PERSON (Portuguese: PESSOA);
2. POSITION (Portuguese: CARGO).

C.2.1.1 Delimitation

The full string of a person's name must be entirely classified as a single NE, e.g. "António Faria Fagundes" or "José António dos Santos Taveira". Whenever the name is preceded by a formal or honorific designation or a kinship degree, that designation must also be included in the NE, e.g. "Dr. Manuel Martins" (id.), "Tio João" (Uncle John). The same happens with roman numerals appearing in the names of popes, kings (and, more rarely, in the name of some individuals): Bento XVI, D. João II.

C.2.1.2 PERSON subtype

This subtype is meant to capture people's individual names, usually expressed by proper names. The NE includes their formal designation, such as "Sr." (Mr.), "Sra." (Mrs.), "Dr." (id.); kinship degrees, like "tio" (uncle), "avô" (grandfather). Proper names include nicknames, such as "Tarzan Taborda" (Albano Taborda Curto Esteves, the famous Portuguese wrestler); diminutives, like "Zé" for "José" (Joe for Joseph); initials ("JFK"); mythological ("Júpiter") and religious names, such as "São Pedro" (Saint Peter).

In the case of job titles or equivalent (see also Section C.2.1.3), when accompanying the proper name but *not* separated by comma, as in "o Presidente Cavaco Silva" (the President Cavaco Silva), the job/position must be included in the NE, whether it is capitalized or not. If the proper name is in apposition to the title, and separated by comma, only the name is included in the NE. The same rationale applies to kinship degrees. In other cases, when the proper name is complemented by a toponym, such as "a Rainha Isabel II de Inglaterra" (Queen Elizabeth II of England), both the job/position and the toponym must be included in the NE.

Vocatives such as "Sua Alteza" (Your Highness) and their abbreviations (S.A.R. / Y.R.H., which means "Sua Alteza Real" / "Your Royal Highness") are not to be included in the NE, even if they appear next to the name or title of the NE: "Sua Excelência o Presidente da República Cavaco Silva" (His Excellency the President of the Republic Cavaco Silva).

Correct annotations

- O meu nome é <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Diogo Oliveira**</EM>.

- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Presidente Cavaco Silva**</EM> esteve ontem presente em Espanha.
- O meu <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**tio João**</EM> era farmacêutico.
- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**padre Melícias** </EM> chegou atrasado à Missa.
- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**acordeonista Miguel Sá**</EM> deu ontem uma entrevista.
- Sua Alteza Real a <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA"> **Rainha Isabel II de Inglaterra**</EM> foi de férias para as Caraíbas.
- Os jogadores <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Eduardo** </EM> e <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Fábio Coentrão**</EM> foram os melhores em campo.

Incorrect annotations

- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Sr. José António** </EM> **dos** <EM CATEG="PESSOA" TIPO="INDIVIDUAL">**Santos Tavares**</EM> é do Benfica.
- O **Presidente** <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Cavaco Silva**</EM> esteve ontem presente em Espanha.
- O **acordeonista** <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Miguel Sá**</EM> deu uma entrevista.
- <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Sua Alteza Real a Rainha Isabel II de Inglaterra**</EM> foi de férias para as Caraíbas.
- <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Vossas Excelências** </EM> desculpem-me.
- Os jogadores <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">**Eduardo e Fábio Coentrão**</EM> foram os melhores em campo.

C.2.1.3 POSITION subtype

The POSITION subtype should be used in reference to a position that is occupied by one person but which may be occupied by other individuals in the future. That is, in a given context, POSITION can concretely represent a person, but by referring to his or her position. For example, "Papa" (Pope) or "Rainha da Suécia" (Queen of Sweden). The name of institutions, such as "secretário-geral da ONU" (UN Secretary-General) or, in the case of national-wide positions, the name of the corresponding country

("Presidente da França", France's President), must be included in the NE. The NEs of this type, however, never include person's proper names, which are then classified as HUMAN INDIVIDUAL PERSON. In the case of country names, the strings with the corresponding gentilic adjectives are to be considered as NEs as well: "presidente francês" (French president), "rainha inglesa" (English queen). Finally, the NE is retrieved irrespective of the position noun being spelled in upper or lower case initials: "Presidente da França" (France's President) and "presidente da França" (France's president).

Correct annotations

- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="CARGO">**secretário-geral da ONU**</EM> foi visitar o Haiti na sequência do terramoto de Janeiro de 2010.
- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="CARGO">**Papa**</EM> é a figura máxima da Igreja Católica.
- Intervenção de Encerramento de Sua Excelência o <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="CARGO">**Primeiro Ministro**</EM>.
- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="CARGO">**Presidente da República**</EM> é, de uma forma geral, o <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="CARGO">**Chefe de Estado**</EM>.

Incorrect annotations

- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="CARGO">**secretário-geral**</EM> **da ONU** foi visitar o Haiti na sequência do terramoto de Janeiro de 2010.
- O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="CARGO">**Papa Bento XVI**</EM> veio recentemente a Portugal e visitou Lisboa, Porto e Fátima.

C.2.2 COLLECTIVE type

The COLLECTIVE type is meant to capture NEs that designate organizations related to the administration and governance of a territory, such as ministries or municipalities, but also those that designate institutions, companies and groups.

We divide the COLLECTIVE type into three subtypes: ADMINISTRATION (Portuguese: ADMINISTRAÇÃO), INSTITUTION (Portuguese: INSTITUIÇÃO) and GROUP (Portuguese: GRUPO). Their differences are explained in Sections C.2.2.2, C.2.2.3 and C.2.2.4.

C.2.2.1 Delimitation

A COLLECTIVE NE must include all words belonging to the name of a specific organization, such as "Ministério da Cultura" (Ministry of Culture), "Sport Lisboa e Benfica" (id.), "Departamento dos Alunos do IST" (IST's Student Department) or "Departamento de Marketing da General Motors" (General Motors' Marketing Department). In the latter, notice that although the organization is "Marketing Department", "General Motors" is included in the NE, since we want to extract as much knowledge as possible.

C.2.2.2 ADMINISTRATION subtype

The ADMINISTRATION subtype is mainly concerned with capturing entities that designate administrative organizations, such as ministries, municipalities, counties or state departments. Furthermore, it is also concerned with organizations that govern at an international or supranational level (such as the United Nations or the European Union).

Correct annotations

- O <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="ADMINISTRAÇÃO">**Ministério da Saúde**</EM> contratou 30 médicos uruguaios para trabalharem no 112.
- Em Portugal, a <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="ADMINISTRAÇÃO">**Presidência do Conselho de Ministros**</EM> é o departamento governativo ...
- O <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="ADMINISTRAÇÃO">**Parlamento iraniano**</EM> votou hoje uma lei que obriga o <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="ADMINISTRAÇÃO">**Governo**</EM> a "acelerar" o programa nuclear ...

Incorrect annotations

- O <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="ADMINISTRAÇÃO">**Parlamento** </EM> **iraniano** votou hoje uma lei que obriga o <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="ADMINISTRAÇÃO">**Governo**</EM> a "acelerar" o programa nuclear ...

C.2.2.3 INSTITUTION subtype

The INSTITUTION subtype captures entities that designate institutions in general, not included in HUMAN COLLECTIVE ADMINISTRATION, whether they are institutions (in the strict sense), companies (for- or non-profit) or other kinds of organizations, such as societies, clubs, etc. The subtype also encompasses associations and other organizations that promote a cooperative spirit; universities, schools, communities and political parties are also included.

Correct annotations

- A <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**RTP**</EM> despediu vários trabalhadores.
- O <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Benfica**</EM> contratou novos jogadores.
- A <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Google**</EM> é muito famosa pelo seu motor de busca.
- O acidente está a ser investigado pela <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Polícia Judiciária**</EM>.
- A <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Igreja Católica** </EM> sempre se viu como uma união ou comunhão na diversidade.

- O <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Instituto Superior Técnico**</EM> faz parte da <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Universidade Técnica de Lisboa**</EM>.
- A <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Companhia Nacional de Bailado**</EM> foi criada em 1977.

C.2.2.4 GROUP subtype

The GROUP subtype is used to capture every other COLLECTIVE entities, as long as they have a proper designation, that is, a non-descriptive proper name (e.g. musical groups, such as "U2", "Spice Girls", "Metallica", "Pearl Jam", etc).

Correct annotations

- Ontem fui assistir ao concerto dos <EM CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="GRUPO">**U2**</EM>.

C.3 The LOCATION category

LOCATION (Portuguese: LOCAL) is a very broad category, so it was divided into several types and subtypes. The main purpose behind the inclusion of this category in the NER task is to capture specific place entities in texts. These include not only geographical and natural locations (hydronyms: oceans, seas, lakes, rivers; oronyms: mountains, mountain chains; deserts, forests, woods, beaches, islands, archipelagos, etc), but also human created or human defined locations, like countries, counties, parishes, and other administrative circumscriptions; they also include street, neighbourhood, road, and highway names. Besides these entities, virtual locations (e.g. Internet websites) are also included.

C.3.1 Delimitation

For the delimitation of a LOCATION NE, all words that are necessary to provide a full and unambiguous identification must be included; this typically means matching the longest string as possible including most classifiers appearing in front of proper names. This criterion is different from the one present in the directives adopted in HAREM evaluation campaigns (Santos *et al.* [45]). The rationale behind it is to enrich the extracted information, which may be useful for other NLP applications.

For example, the NE "rio Tejo" (Tejo river) and "cidade de Viseu" (city of Viseu) must, in all cases, include the word "rio" (river) and "cidade" (city) whether or not it is capitalized. Similarly, the names of streets, avenues, places, etc. must always include the words "rua", "avenida" and "praça", respectively (or their abbreviations, "R.", "Av.", "Pça"), capitalized or not.

C.3.2 LOCATION types

The LOCATION category consists of three main types: CREATED (Portuguese: CRIADO), PHYSICAL (Portuguese: FÍSICO) and VIRTUAL (Portuguese: id). Each of these main types is divided into several

subtypes, which are further described in Sections C.3.2.1 through C.3.2.3.

C.3.2.1 CREATED type

The CREATED type is meant to capture NEs that were created or delimited by humans. It is divided into the following subtypes:

1. **COUNTRY** (Portuguese: PAÍS): it includes countries, principalities and unions of countries, such as the European Union. It also includes conventional designations of certain countries such as "País do sol nascente (Japão)" (Land of the rising sun (Japan)) or "Império do meio (China)" (Middle Kingdom);

2. **DIVISION** (Portuguese: DIVISÃO): it includes population aggregates such as cities, towns and villages, as well as other administrative divisions like States in Brazil, municipalities, districts, provinces in Portugal, administrative regions (Algarve) or tax districts. It also includes conventional designations of certain cities, such as "Cidade Maravilhosa (Rio de Janeiro)" (Wonderful City), "Cidade das Luzes (Paris)" (The City of Light), etc;

3. **REGION** (Portuguese: REGIÃO): cultural or traditional location, with no administrative value, such as "o Médio Oriente" (The Middle East), "o Terceiro Mundo" (The Third World), "o Nordeste brasileiro" (the Brazilian northeast) or "a Raia" (the border region between Portugal and Spain), etc;

4. **CONSTRUCTION** (Portuguese: CONSTRUÇÃO): it includes all kinds of construction, from buildings, clusters of buildings or specific areas of a building, to bridges, dams, ports, etc;

5. **STREET** (Portuguese: RUA): it includes all kinds of roads, streets, avenues, alleys, squares, small squares, etc; it also includes general designations for parts of a town, such as "Baixa" (Downtown), with no administrative value.

Correct annotations

- Quem vive no <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAÍS">**Mónaco**</EM>, o principado mais badalado do planeta, não sabe o que é imposto de renda.

- Qualquer cidadão da <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAÍS">**União Europeia**</EM> pode agora escrever ao Parlamento Europeu.

- Segundo dados do INE de 2006, o <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**concelho de Sintra**</EM>, apesar de ter menos residentes do que o <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**concelho de Lisboa**</EM>, é o que mais crianças tem.

- A fiscalização aconteceu em <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**Mato Grosso do Sul**</EM>.

- O governador Jon Corzine promulgou segunda-feira a lei que abole a pena de morte no <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**estado de New Jersey**</EM>.

- Os Estados Unidos não pretendem construir novas bases militares em <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="REGIÃO">**África**</EM>, apesar da criação do novo comando militar africano (AFRICOM).

- O <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="CONSTRUÇÃO">**Aeroporto da Madeira** </EM> e o <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="CONSTRUÇÃO">**Aeroporto de Porto Santo**</EM> são ponto de partida e de chegada de várias companhias aéreas internacionais.

- Virar à direita no cruzamento da <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="RUA">**Av. Lusíada**</EM> com a <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="RUA">**Avenida dos Combatentes**</EM>.

Incorrect annotations

- Segundo dados do INE de 2006, o **concelho de** <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**Sintra**</EM>, apesar de ter menos residentes do que o **concelho de** <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**Lisboa**</EM>, é o que mais crianças tem.

- O governador Jon Corzine promulgou segunda-feira a lei que abole a pena de morte no **estado de** <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**New Jersey**</EM>.

- O **Aeroporto da** <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="ILHA">**Madeira**</EM> e o **Aeroporto de** <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="ILHA">**Porto Santo** </EM> são ponto de partida e de chegada de várias companhias aéreas internacionais.

C.3.2.2 PHYSICAL type

The PHYSICAL type is meant to capture NEs that were named (not created) by humans. It is divided into the following subtypes:

1. **WATERCOURSE** (Portuguese: AGUACURSO): it includes rivers, streams, creeks, tributaries, waterfalls, etc;

2. **WATERMASS** (Portuguese: AGUAMASSA): it includes lakes, seas, oceans, gulfs, straits, canals, ponds, reservoirs, etc;

3. **RELIEF** (Portuguese: RELEVO): it includes mountains, ridges, hills, plains, plateaus, valleys, etc;

4. **PLANET** (Portuguese: PLANETA): it includes all celestial bodies;

5. **ISLAND** (Portuguese: ILHA): it includes islands and archipelagos;

6. **NATURALREGION** (Portuguese: REGIÃONATURAL): it designates a geographical/natural region, such as the Balkans, the Sahara Desert, the Amazonas region, etc.[1]

Correct annotations

- Primeiro visitei o <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="AGUACURSO">**Tamisa**</EM> e de seguida as <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="AGUACURSO">**Cataratas do Niagara**</EM>.
- O <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="AGUAMASSA">**Estreito de Gibraltar**</EM> é um estreito que separa o <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="AGUAMASSA">**Golfo de Cádis**</EM> do <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="AGUAMASSA">**Mar de Alborão**</EM>.
- A <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="RELEVO">**Serra da Estrela**</EM> é a maior elevação de Portugal Continental, e a segunda maior em território da República Portuguesa (apenas o <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="RELEVO">**Pico**</EM>, nos Açores, a supera).
- Quer a <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="PLANETA">**Terra**</EM>, quer <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="PLANETA">**Marte**</EM>, ficam na <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="PLANETA">**Via Láctea**</EM>.
- O <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="ILHA">**arquipélago dos Açores**</EM> é composto por nove ilhas, divididas em três grupos (ocidental, central e oriental), sendo que a maior das quais é a <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="ILHA">**ilha de São Miguel**</EM>.
- Uma viagem por 13 países ligando o calor das areias do <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="REGIÃO">**Deserto do Sahara**</EM>, à neve do frio da <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="REGIÃO">**Escandinávia**</EM>.

Incorrect annotations

- O <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="ILHA">**arquipélago dos Açores**</EM> é composto por nove ilhas, divididas em três grupos (ocidental, central e oriental), sendo que a maior das quais é a **ilha de** <EM CATEG="LOCAL" TIPO="FÍSICO" SUBTIPO="ILHA">**São Miguel**</EM>.

C.3.2.3 VIRTUAL type

The VIRTUAL type is meant to capture virtual NEs that are deemed to belong to the LOCATION category for being employed as a place for information. However, pieces of literature are not included: published books of all major forms (novels, poems, short stories, novellas) covering all genres (epic, lyric, drama,

[1] Ambiguity may arise between this type and LOCATION CREATED REGION (Section C.3.2.1), LOCATION CREATED COUNTRY (id.), or even HUMAN COLLECTIVE ADMINISTRATION (Section C.2.2.2). In such cases, preference is given to the geographical type.

romance, satire, tragedy, comedy, biography, etc). All these are treated in a different category, called OBRA (WORK), which has not been considered for this thesis. VIRTUAL is divided into the following subtypes:

1. **SITE** (Portuguese: SÍTIO): it includes all virtual locations: Web, WAP, FTP, etc;
2. **DOCUMENTS** (Portuguese: DOCUMENTOS): it includes other LOCATION VIRTUAL entities, such as regulations, laws, standards, decrees, directives, "planos director" (a regulatory document that specifies planning and land management in a given Portuguese municipality), etc.

Correct annotations

- Costumo aceder a <EM CATEG="LOCAL" TIPO="VIRTUAL" SUBTIPO="SÍTIO"> **www.google.pt**</EM> para fazer pesquisas.
- Para mais informações sobre este concurso, consultar <EM CATEG="LOCAL" TIPO="VIRTUAL" SUBTIPO="DOCUMENTOS">**Regulamento**</EM> (.pdf 57 KB / .doc 34 KB).
- <EM CATEG="LOCAL" TIPO="VIRTUAL" SUBTIPO="DOCUMENTOS">**Decreto-Lei nº 3/2008** </EM>, de 7 de Janeiro, pelo Secretário de Estado da Educação, Valter Lemos.

Incorrect annotations

- Foi Camões quem, <EM CATEG="LOCAL" TIPO="VIRTUAL" SUBTIPO="DOCUMENTOS">**nos Lusíadas**</EM>, comparou ...

C.4 Metonymy

This section presents the set of directives to be used when dealing with cases of metonymy; it is out of its scope to present a detailed explanation of what metonymy is and how the system captures the metonymical relations (see Section 3.3.7). Below is a brief explanation that serves as a contextualization.

C.4.1 Context

Metonymy occurs when words are used in a different context than that of their basic distributional class. For example, "Portugal" is a LOCATION CREATED COUNTRY NE, but when used in the sentence "Portugal acha que a crise veio para ficar" (Portugal thinks the crisis is here to stay), it is being used in a different context; in these cases, there is a shift from one distributional class to another and, as such, the output must differ accordingly.

Not all basic distributional classes can be matched to every other classes, and there are only some obvious possible paths. In this thesis, we will only deal with three shifts:

- LOCATION to HUMAN (Section C.4.2);
- HUMAN COLLECTIVE to HUMAN INDIVIDUAL (Section C.4.3);
- HUMAN COLLECTIVE to LOCATION (Section C.4.4).

The following sections present the criteria that is used in order to mark a category as metonymical, and they also present some examples of the intended output. Regarding the latter, the main difference is that besides keeping the basic distributional class, one extra field is added: MET-CAT (metonymical category). This field indicates which (metonymical) category is being referred to by the entity; this indication should provide a sufficient level of detail in order to understand the shift. Therefore, for example, if an organization is being referred to as a location, it is sufficient to indicate MET-CAT="LOCAL" (LOCATION), because the shift was from HUMAN to LOCATION. However, if an organization is being referred to as a person, it is not enough to indicate MET-CAT="HUMANO" (HUMAN) because both entities are human (the former is COLLECTIVE, the latter is INDIVIDUAL). Thus, in such cases, it is more appropriate to indicate MET-CAT="HUMANO INDIVIDUAL".

C.4.2 LOCATION to HUMAN shift

This shift occurs whenever a LOCATION NE is referred to as a HUMAN NE. LOCATION NEs answer questions that involve the interrogative adverb "where", such as "onde vives?" (where do you live?). In order to detect this shift, it is useful to determine whether a LOCATION NE is being used in a sentence in an unusual way (in terms of syntax), e.g. if it is being used as a subject.

It is important to notice that this shift assumes that the metonymical category (in this case, HUMAN) is always of the COLLECTIVE type. At this stage, we are not worried about making the distinction between COLLECTIVE and INDIVIDUAL since we feel that it is sufficient to provide a level of detail at the TYPE level.

C.4.2.1 Examples

- Cristiano Ronaldo conquistou a admiração da <EM MET-CAT="HUMANO COLECTIVO" CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAÍS">**Inglaterra**</EM>.
- <EM MET-CAT="HUMANO COLECTIVO" CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="DIVISÃO">**Lisboa**</EM> ficou horrorizada com essa notícia.
- <EM MET-CAT="HUMANO COLECTIVO" CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAIS">**Montenegro**</EM> obteve a independência da <EM MET-CAT="HUMANO COLECTIVO" CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAIS">**Sérvia**</EM> em 2006.
- <EM MET-CAT="HUMANO COLECTIVO" CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAIS">**Portugal**</EM> ganhou autonomia em relação à <EM MET-CAT="HUMANO COLECTIVO" CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAIS">**Espanha**</EM>.

C.4.3 HUMAN COLLECTIVE to HUMAN INDIVIDUAL shift

This shift occurs whenever a HUMAN COLLECTIVE NE is referred to as a HUMAN INDIVIDUAL NE, which typically happens when the NE is subject of a "human-activity verb". These verbs are associated with human behaviour and typically have a human subject, e.g. "dizer" (to say), "acreditar" (to believe), "adiar" (to postpone), "delegar" (to delegate), etc.

C.4.3.1 Examples

- O <EM MET-CAT="HUMANO INDIVIDUAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**GNR**</EM> constatou que eu ia em excesso de velocidade, e portanto ia ser multado.

- O <EM MET-CAT="HUMANO INDIVIDUAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="ADMINISTRACAO">**Governo de Portugal**</EM> deliberou, e decidiu que as coisas iam mudar.

C.4.4 HUMAN COLLECTIVE to LOCATION shift

This shift occurs whenever a HUMAN COLLECTIVE NE is referred to as a LOCATION NE. This happens when the human NE is used in a locative syntactic position, such as in the locative argument position of locative verbs or of locative prepositions, e.g. "O Pedro apareceu na SIC" (Peter appeared on SIC).

C.4.4.1 Examples

- No ano passado estive na <EM MET-CAT="LOCAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Google**</EM>.

- O anúncio a que me referia estava no <EM MET-CAT="LOCAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Diário de Notícias**</EM> de ontem, mas também passou na <EM MET-CAT="LOCAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Rádio Comercial**</EM>.

- Fiz uma pesquisa avançada no <EM MET-CAT="LOCAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Google**</EM> e obtive melhores resultados do que no <EM MET-CAT="LOCAL" CATEG="HUMANO" TIPO="COLECTIVO" SUBTIPO="INSTITUIÇÃO">**Yahoo**</EM>.