INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Anaphora Resolution in Portuguese

An hybrid approach

João Silvestre Marques

Dissertation for obtaining the Master Degree in

Information Systems and Computer Engineering

Jury

President:	Professor Doutor Mário Rui Fonseca dos Santos Gomes
Advisor:	Professor Doutor Nuno João Neves Mamede
Co-advisor:	Professor Doutor Jorge Manuel Evangelista Baptista
Evaluation jury:	Doutora Paula Cristina Quaresma da Fonseca Carvalho

October 2013

Acknowledgements

I would like to thank my supervisor Professor Nuno João Neves Mamede and co-advisor Professor Jorge Manuel Evangelista Baptista for their friendship, guidance and wisdom while always being critic of my work. This outcome would not be possible without their invaluable trust and support.

I also would like to thank Cláudio Diniz, Vera Cabarrão, Rui Talhadas and Alexandre Vicente for their assistence and cooperation.

Last, but not least, I would like to thank my family and friends, particularly my parents and my sister for their continuous support and belief and for always being there for me. They are intrinsically connected with my academic career and life.

Lisbon, October 15^{th} 2013

João Marques

Resumo

Atualmente, devido à imensa quantidade de informação disponível, existe uma necessidade cada vez maior de extrair e processar informação de textos de língua natural. A Resolução de Anáfora é uma das mais relevantes e necessárias tarefas de Processamento de Língua Natural (PLN) para responder a tais necessidades e tem sido objecto de estudo desde há vários anos. A anáfora é um importante mecanismo de coesão textual, na medida em que articula e interliga diferentes partes do texto, garantido a sua unidade semântica.

Este trabalho visa desenvolver um módulo de anáfora pronominal e co-referencial em Português a integrar na cadeia PLN do L^2F, STRING. Este trabalho pretende também melhorar a eficiência do módulo atualmente em uso, cuja avaliação produziu uma medida f de 33.5%. Para tal, anotamos um *corpus* bastante heterogéneo composto por textos de diferentes géneros: textos literários, notícias, artigos de opinião, artigos de revista, entre outros. No total contém 290.000 tokens e a campanha de anotação produziu 9.268 anáforas.

A estratégia adotada assentou na identificação de expressões anafóricas e candidatos através de um sistema de regras; e na seleção do candidato mais provável para antecedente por um modelo construído com base no algoritmo, de aprendizagem automática, de máxima entropia (ME).

A avaliação do sistema, distinguindo as diferentes fases de processamento e os diversos tipos de expressões anafóricas considerados, demonstrou uma melhoria significativa na performance do MRA 2.0, apresentando uma medida f de 82% na identificação de expressões anafóricas, 70% na identificação de candidatos a antecedente e 54% na resolução de anáforas.

Abstract

Nowadays, due to the large amount of information available, there is a growing need of extracting information from natural language texts and processing it. Anaphora Resolution is one of the most relevant Natural Language Processing (NLP) tasks necessary to answer such needs and has been under study for years. Anaphora is an important textual cohesion mechanism as it articulates and connects different parts of the text, ensuring its semantic unit.

This study aims at developing a co-referential, pronominal anaphora resolution module in Portuguese, to incorporate in the L^2F's NLP chain, STRING. This thesis also intends to improve the efficiency of the module currently in use, developed by Nuno Nobre, whose evaluation produced 33.5% f-measure results. To do so, we annotate a quite heterogeneous *corpus*, being composed of texts from different genres: novels, pieces of news, magazine news and newspaper columns, among others. In total, it contains 290,000 tokens and the annotation campaign produced 9,268 anaphoras.

The strategy adopted was based in the identification of anaphors and candidates through a rule system; and in the selection of the most probable candidate for antecedent by a model built based on the (machine learning) algorithm Expectation-Maximization (EM).

The system's evaluation, telling apart the different processing stages and the several anaphor types considered, showed a significant improvement of ARM's 2.0 performance, with a f-measure of 82% on the anaphor identification, 70% on the candidate identification to antecedent and 54% on anaphora resolution.

Palavras-Chave

Keywords

Palavras-Chave

resolução de anáfora

anáfora pronominal

expressão anafórica

anotação de *corpus*

aprendizagem automática

Keywords

anaphora resolution

pronominal anaphora

anaphor

corpus annotation

machine learning

Table of Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Designation in English	Designation in Portuguese
AR	**A**naphora **R**esolution	Resolução de Anáfora
ARM	**A**naphora **R**esolution **M**odule	Módulo de Resolução de Anáfora
EM	**E**xpectation-**M**aximization	Máxima Entropia
IR	**I**nformation **R**etrieval	Recuperação de Informação
L^{2}F	Spoken Language Systems Laboratory	**L**aboratório de Sistemas de **L**íngua **F**alada
MARv	**M**orphosyntactic **A**mbiguity **R**esol**v**er	[módulo de] Desambiguação Morfossintática (estatístico)
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
POS	**P**art **o**f **S**peech	(“parte de discurso”) categoria morfossintática/gramatical
PP	**P**repositional **P**hrase	Sintagma preposicional
RuDriCo	**Ru**le-**Dri**ven **Co**nverter	Conversor baseado em regras
XIP	**X**erox **I**ncremental **P**arser	Analisador Sintático (da Xerox)
XML	E**x**tensible **M**arkup **L**anguage	

List of Terms

Term	Meaning
Corpus	A collection of written or spoken linguistic material in machine-readable form, assembled with explicit criteria and using adequate sampling methoodology for the purpose of studying linguistic structures, frequencies, *etc*., and that is supposed to represent the language (or language variety) from which it was sampled.
F-measure	An evaluation measure that combines *Precision* and *Recall*, *a.k.a.* the *harmonic mean*; though the measures can be accorded different weight, usually, in NLP, they are given the same weight.
Feature	Specification of an attribute and its value.
Instance	A single object of the world from which a model will be learned, or on which a model will be used (*e.g.*, for prediction). In most machine learning work, instances are described by feature vectors.
Metonymy	A figure of speech that designates the substitution of a word for another word (mostly nouns), usually the two having a part-whole relation between them (*e.g.* "suit" for "business executive" or "tracks" for "horse races").
Precision	An evaluation measure that considers the proportion of *correct* answers provided by a system over the set of all 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 (drawn from a golden standard).

Chapter 1

Introduction

In a time when Natural Language Processing (NLP) draws more and more attention, the task of Anaphora Resolution presents itself as critical for many applications such as machine translation, information extraction and question answering [36]. For a machine, it is difficult to select the correct entity (antecedent) to which an anaphor refers to, due to the ambiguous nature of natural languages. To overcome this drawback, a great amount of linguistic knowledge (morphological, lexical, syntactic, semantic and even world knowledge) may be required.

In this work, we present the strategies used throughout time to resolve anaphora, as well as some systems that have been influential in the evolution of the task. We also describe our approach to resolve anaphora in Portuguese texts, based on a number of features, which makes use of the different knowledge already available, combined with a machine learning component.

1.1 Anaphora

1.1.1 Basic notions

An anaphora is a relation between a type of expression whose reference – the anaphor – depends upon another referential element – the antecedent. Consider the following example:

> (1.1) *Luís Figo* é um ex-futebolista português. Em 2001, *ele* foi distinguido como melhor jogador do Mundo.
>
> *Luís Figo* is a former Portuguese football player. In 2001, *he* was distinguished as the world's best player.

As human readers, we immediately see that this sequence state that Luís Figo was distinguished as the world's best player in 2001. However, this deduction actually requires that a link to be established between *Luís Figo* in the first sentence and *he* in the second. Only then, the distinction as the world's best player in 2001 that is mentioned in the second sentence can be attributed to *Luís Figo* in the first. Therefore, the interpretation of the second sentence is dependent of the former, ensuring in this way the *cohesion* between the two sentences.

Besides contributing to the cohesion of the discourse, the two expressions are co-referential since they both refer to the same person in the real world, *Luís Figo*. This does not always stand true:

(1.2) O *homem* que deu o *salário* à sua esposa é mais sábio do que *aquele* que *o* deu à sua amante.

The *man* who gave his *paycheck* to his wife is wiser than the *man* who gave *it* to his mistress.

In the example above, the anaphor *o* and its antecedent *salário* do not correspond to the same referent in real world but to one of a similar description. The same happens between *aquele (homem)* and *homem*. This type of phenomenon is called *identity-of-sense* anaphora as opposite to identity-of-reference.

Anaphora can also be classified according to the antecedent's location: *intrasentential*, if the antecedent is in the same sentence as the anaphor; or *intersentential*, if the anaphoric relation is made across sentence boundaries.

1.1.2 Types of anaphora

In addition to the immense knowledge needed to perform anaphora resolution, the various forms that anaphora can assume make it a very challenging task to teach computers how to solve anaphora [36]. The following types of anaphora can be considered:

Pronominal Anaphora

This type of anaphora is based on the use of personal pronouns (1.3), possessive (1.4), demonstrative (1.5) or reflexive (1.6 and 1.8).

(1.3) O João deu uma prenda à *Maria*. *Ela* gostou muito.

João gave an offer to *Maria*. *She* liked a lot.

(1.4) *Cavaco Silva* não terminou a campanha sem visitar as *suas* raízes.

Cavaco Silva did not finish his campaign without visiting *his* roots.

(1.5) Passos Coelho discursou sobre *a delicada situação do país*. Segundo o Primeiro Ministro, *esta* irá requerer sacrifícios aos portugueses.

Passos Coelho spoke about the *country's delicate situation*. According to the Prime Minister, *this* will require sacrifices from the Portuguese.

In Portuguese, personal pronouns include nominative (tonic) {*eu*, *tu*, *ele*/*ela*, *nós*, *vós*, *eles*/*elas*} and clitic (atonic), accusative forms {*me*, *te*, *o*/*a*, *nos*, *vos*, *os*/*as*}, dative {*me*, *te*, *lhe*, *nos*, *vos*, *lhes*}, oblique {*mim*, *ti*, *si*, *nós*, *vós*}[1] and reflexive pronouns {*me*, *te*, *se*, *nos*, *vos*}. Reflexive pronouns, in Portuguese, appear in the form of the pronouns attached to the verb. Two different syntactic constructions can be seen:

- *Intrinsically reflexive verbal constructions*: verbs that can only be used reflexively, *i.e.*, the pronoun does not correspond to the pronouning of a distributionally free NP or PP. Portuguese intrinsically reflexive verbs include *queixar-se* (complain), *abster-se* (abstain), *suicidar-se* (commit suicide), *etc.* (1.6 and 1.7);

[1] Plus the contraction of preposition *com*, oblique pronouns {*comigo*, *contigo*, *consigo*, *connosco*, *convosco*}.

- *Normal reflexive verbal constructions*: verbs that select a free NP or PP complement which is pronominalized as a reflexive if it refers to the same entity as the verb's subject. In example (1.8), the verb *magoar* (hurt) is employed reflexively whereas in (1.9) it is not. The list of normal reflexive verbs also comprises verbs like *ver* (look), *pentear* (comb), *lavar* (wash), *etc*.

(1.6) *A mãe do Tiago* suicidou-*se*.

Tiago's mother committed suicide.

(1.7) A mãe do Tiago suicidou o Pedro.

Tiago's mother suicided Pedro.

(1.8) *A Maria* magoou-*se*.

Maria hurt *herself*.

(1.9) O Bruno inadvertidamente magoou *o seu irmão mais novo*.

Bruno inadvertently hurt *his baby brother*.

Personal pronouns substitute in discourse an entire entity, forming a syntactic constituent. On the other hand, *possessive* pronouns {*meu*, *teu*, *seu*, *nosso*, *vosso*} correspond to the pronouning of a *de N* (of N). Prepositional phrase function as a determiner of another noun. In Portuguese, they agree in gender and number with the noun they determine. Furthermore, since they are determiners, they contribute to the reference of the head-noun, which can then be zeroed leaving the possessive alone, with the role of anaphor and as a head (on surface) of the constituent: *Ele visitou as suas raízes e eu visitei as minhas [raízes]*.[2] The same determinative function can be seen in the demonstrative pronouns (1.5), as well as certain indefinites. Finally, even articles and numerals can take on themselves this role as anaphors, when discoursive context allows for the zeroing of the phrase head-noun, that they determine (1.10–1.11).

(1.10) O Pedro comprou duas *camisas*: *a* [camisa] azul fica-lhe muito bem.

Pedro bought two *shirts*: the blue *one* suits him very well.

(1.11) O Pedro comprou três *camisas*: *duas* [camisas] eram grandes demais.

Pedro bought three *shirts*: *two* [shirts] were too big.

In some other cases, instead of a reflexive, Portuguese uses oblique pronouns accompanied by focus determiners *próprio* or *mesmo* (self). Like reflexives, the use of these focalizers makes the pronoun refer to the same entity as the verb's subject as in examples (1.12) and (1.13).

(1.12) *A Joana* gosta *dela própria*.

Joana likes *herself*.

[2]Unlike English and other languages there is no autonomous form in Portuguese for this "pronominal" use of the demonstrative (*cp. my*/*mine*).

(1.13) *A Carolina* acredita que *ela mesma* é a rapariga mais inteligente da turma.

Carolina believes *herself* to be the class's most intelligent girl.

Pronominal anaphora usually relates to third person pronouns, both singular and plural, while the first and second person refer to the dialog interlocutors when mentioned in direct speech. We will not cover dialogues as this work will only cover pronominal anaphora in indirect speech.

Definite description

This kind of anaphora that, besides referring to an antecedent, also provides additional information to the reader. Consider the following example:

(1.14) *Cristiano Ronaldo*, após os três golos que apontou ao Deportivo, subiu ao nono lugar do ranking dos 10 maiores goleadores da história do Real Madrid. *Ronaldo* já leva 6 golos no campeonato. *O internacional português*, com 156 golos, porém, ainda está longe do topo que é ocupado por Raúl com 323 golos.

Cristiano Ronaldo, after the three goals scored against Deportivo, climbed to ninth position of the top 10 scorers in Real Madrid's history. *Ronaldo* already has 6 goals in the championship. *The Portuguese international football player*, with 156 goals, however is still far from the top which is occupied by Raúl with 323 goals.

In this piece of text, *Cristiano Ronaldo* is referred as *the Portuguese international football player* which provides the reader with new information about the named entity. This also helps to increase the cohesion of the text. Also, *Cristiano Ronaldo* is referred in the second sentence just as *Ronaldo* which indicates that substring matching can be an effective way to solve definite descriptions.

Noun Anaphora

This anaphora is a specific case of identity-of-sense anaphora in which the antecedent is just the head noun and not the noun phrase.

(1.15) Eu não quero o *casaco* azul. Prefiro *o* preto.[3]

I don't want the blue *jacket*. I prefer the black *one*.

In (1.15), the anaphor *o* points back only to the noun *casaco* instead of the noun phrase *o casaco azul*. The antecedent and the anaphor do not refer to the same *casaco* in real life, thus the relation is not co-referential.

Verb Anaphora

Verb anaphora occurs when a verb anaphor has a verb or verb phrase as its antecedent.

[3]Compose with example (1.10), produced by the same linguistic device (the anaphoric use of an article) but keeping identity of reference.

(1.16) “Portugal deve *renegociar a dívida* tal como *fez* a Alemanha quando precisou", defende o BE.

“Portugal should *renegotiate the debt* as *did* Germany when they needed it", defended the BE.

As we can see above, the verb anaphor *fez* points back to the verb phrase *renegociar a dívida*, which plays the role of antecedent.

Adverb Anaphora

Adverb anaphora can be divided in temporal (1.17) and locative anaphora (1.18). However, it should be noted that most Portuguese adverbs used in this type of anaphora are often used in a non-anaphoric (expletive) way as well (1.19). While in (1.18), the word *lá*/*there* is clearly anaphoric, referring to *Lisbon*, in (1.19) the occurrence of the same word is non-anaphoric.

(1.17) *Antes do 25 de Abril de 1974*, não se podia dizer nada. As coisas eram bem diferentes *então*.

Before April 25th 1974, one could say nothing. Things were quite different *then*.

(1.18) Fernando Pessoa viveu em *Lisboa* e *lá* conheceu Almada Negreiros.

Fernando Pessoa lived in *Lisbon* and *there* he met Almada Negreiros.

(1.19) *Lá* se foi a mesada...

There goes the allowance...

The manner adverb *assim* (thus) can also be considered to have anaphoric value in sentences like (1.20):

(1.20) O Pedro não fala *assim*, *à maneira do norte*.

Pedro does not talk like *that*, *with a Northern accent*.

However, *assim* also functions as a conjunctive adverb, linking different sentences of the same discourse (1.21):

(1.21) O Pedro detesta o Manuel de Oliveira. Assim, decidiu ficar em casa e não veio connosco ver o filme.

Pedro detests Manuel de Oliveira. Thus, he decided to stay home and did not come with us to watch the film.

Zero Anaphora

This form of anaphora is distinguished by the fact that the anaphor is “invisible”. In other words, the anaphor can be viewed as the very zeroing of repeated elements whose presence in the sentence is implicit, making the sentences shorter as well as avoiding repetition of those elements [45].

Zero anaphora can be viewed as the ultimate type of anaphora, which enhances cohesion by reducing the amount of text. It can be applied to nouns, pronouns or even verbs (or verb phrases).

Consider the following example:

(1.22) "*Eles* fizeram um grande trabalho e ∅ são realmente importantes para a equipa."

"*They* have done a great job and ∅ are really important to the team."

In this example (1.22), the subject of *are really important to the team* is a second instance of *They*, which is omitted, since it is co-referent to the subject of the first coordinated clause. This makes the text shorter and avoids the repetition of the anaphor *They*.

Indirect Anaphora

Indirect anaphora requires background knowledge in order to identify the referent. Metonymy and hyperonymy/hiponymy are semantic relations that usually characterize this kind of anaphora.

(1.23) Rindo-se da cara de espanto de Dudley, Harry subiu *a escada*, saltando *três degraus* de cada vez, e correu para o seu quarto.

Laughing at Dudley's face of astonishment, Harry climbed *the stairs*, leaping *three steps* at a time, and ran to his room.

In (1.23), the noun phrase *a escada (stairs)* is regarded as the antecedent of *degraus (steps)*. The reader picks this up since it is known that stairs have steps. However, sometimes the reader has to possess domain knowledge in order to make the necessary inference. In the example:

(1.24) *As Spice Girls* separaram-se em 2001 e *Victoria Beckham* lançou o álbum solo no mesmo ano, seguido por uma sucessão de singles.

The Spice Girls broke up in 2001 and *Victoria Beckham* released the solo album that same year, followed by a succession of singles.

one must know that *Victoria Beckham* was a former member of the group *The Spice Girls* in order to pick up the reference.

1.1.3 A complex problem

Besides the diversity of forms that anaphora may assume, anaphora resolution presents many problems of difficult resolution. Sometimes even human annotators cannot reach without further information an agreement about the correct antecedent of the anaphor.

For instance, in the sentence:

(1.25) João gosta do *seu cabelo* curto mas a namorada prefere-*o* comprido.

João likes his *hair* short but his girlfriend prefers *it* long.

we can consider that in this case the antecedent is only *cabelo* (hair) and that João's girlfriend prefers her hair long – identity-of-sense – or that the anaphor *o* (it) refers to his hair and that João's girlfriend prefers his hair long – identity-of-reference.

Many times the ambiguity fall on semantics, an additional level of knowledge to take into account when resolving anaphora. Like the previous example, in the next sentence:

(1.26) A Maria disse à Carolina que *ela* estava em perigo.

Maria told Carolina that *she* was in danger.

the anaphor *ela* is ambiguous for it can refer both to *Maria* and *Carolina*, while in:

(1.27) A Maria avisou a *Carolina* de que *ela* estava em perigo.

Maria warned *Carolina* that *she* was in danger.

Carolina is by far the most likely antecedent since the semantics of the verb focuses on the person being warned and not on the person who warns. As a general rule, in Portuguese, zero anaphors refer to the *NP* with the same function in the main clause, therefore, the following sentence is unambiguous:

(1.28) *A Marta* disse à Carolina que $\emptyset$ estava em perigo.

Marta told Carolina $\emptyset_{[she]}$ was in danger.

However, the same lexical constraints imposed by the semantics of the verb *avisar* also apply, so that the zeroed subject in (1.29) stays *Carolina* and not *Maria*:

(1.29) A Marta avisou *a Carolina* de que $\emptyset$ estava em perigo.

Marta warned *Carolina* that $\emptyset_{[she]}$ was in danger.

Moreover, an anaphor can sometimes relate to coordinated antecedents like in (1.30) in which the anaphor *They* refers to the two antecedents *Lampard* and *Terry*.

(1.30) "Espero ter o mesmo trajeto de jogadores como o *Lampard* ou o *Terry* aqui no Chelsea. *Eles* fizeram um grande trabalho e são realmente importantes para a equipa."

"I hope to follow the same path of players as *Lampard* or *Terry* here in Chelsea. *They* have done a great job and are really important to the team."

In other cases, some names are regarded as a group and the anaphor and the antecedent do not agree in number or gender.

(1.31) Depararam-se com um verdadeiro *cardume*. Para além disso, *eles* eram fáceis de apanhar.

They found a real *shoal*. Besides *they* were easy to catch.

In (1.31), the *shoal*, a semantically collective but a gramatically singular noun, is a set of fish and it is later referred by a plural pronoun *they*.

Due to the complexity that anaphora bears, it is imperative to approach anaphora resolution in a progressive way, in order to achieve reasonable results. In this way, we shall focus only on resolving pronominal anaphora since this is arguably the most important and widespread type.

1.2 Machine Learning

NLP requires a considerable amount of knowledge about morphology, syntax, semantics, pragmatics and general knowledge about the world. However, encoding all this knowledge may not be a feasible task. The public availability of annotated *corpora* produced as part of the MUC-6 [19] and MUC-7 [60] conferences promoted, in the 1990s, a gradual shift of AR focus from heuristic approaches to machine learning approaches. Learning-based co-reference research has remained on the spotlight since then, with results regularly published not only in general NLP conferences, but also in specialized conferences (*e.g.*, the biennial Discourse Anaphora and Anaphor Resolution Colloquium [23]) and workshops. As a clustering task, co-reference has also received a lot of attention in the machine learning community. For all this, machine learning presents itself as an alternative to the traditional knowledge-based systems. However, in order to put in practice the learning algorithms, we need a Portuguese training set, hence we need a Portuguese annotated *corpus*.

Unfortunately, Portuguese anaphorically or co-referentially annotated *corpora* are scarce or not publicly available and annotate one anew is a very time-consuming task, even with very efficient annotation tools. Nonetheless, as a *corpus* annotated with anaphoric links is critical to our work, we decided to annotate one deemed to be large enough to meet our purposes.

1.3 Goals

In this dissertation, we aim at performing the task of *anaphora resolution*, that is, we intend to select the correct entity to which anaphors reffer to, in a Portuguese written text, indirect discourse. We follow a strategy that involves (i) identification of anaphoric expressions (ii) compilation of a list of candidates for the entities referred by each anaphoric expression (iii) elimination of some candidates based on specific restrictions (iv) ordering the remaining candidates according to heuristics.

This research follows the work done by Nuno Nobre on this topic (see Section 2.3), as we try to study the *corpus* and its nature, improve the below-standard results achieved at that time by assessing the different stages of AR and the strategy to apply, developing, and eventually combining, a manual rules-based approach and a machine learning one to produce better results. The Anaphora Resolution Module 2.0 will resort to XIP to get morphological, syntatic and semantic data, vital to all the stages of the process.

1.4 Dissertation Structure

Chapter 2 reviews related work, as it describes different approaches and systems that addressed the problem of anaphora resolution. In the end of the section, we overview the systems studied. Chapter 3 describes the importance of annotation tools for the complex process of *corpus* building, presents some frameworks and compares them. Chapter 4 presents the description of the golden standard *corpus* developed for this study and the process of annotating it. In chapter 5, we propose our methods to resolve the problem of pronominal anaphora resolution. Chapter 6 discusses the role of evaluation as well as the different forms of assessing the efficiency of our system. The chapter then presents the results obtained using this methodology. Finally, Chapter 7 concludes this document, pointing to new directions of study and the further development of the AR module.

Chapter 2

State of the Art

THE research on anaphora resolution dates back to the 1960s, a time where work relied mostly on heuristic rules, not to mention that the texts where humanly pre-processed, thus error-prone, which turns the evaluation and comparison with recent systems very difficult.

In the 1970s and 1980s, research started to incorporate knowledge sources, which translated into better results, with special emphasis to Hobbs' work [24], which is still viewed as one of the most successful algorithms in anaphora resolution [36, p. 72].

In the 1990s and 2000s, as people grew aware of the tremendous complexity of the job at hand, research started to be limited to specific types of anaphora in view of ultimately achieving better results.

Nowadays, anaphora resolution is more and more a subject of research as it plays an increasingly vital role in real-world NLP applications. Proper treatment of anaphoric relations shapes the performance of today's applications such as information extraction, machine translation, text summarization, or dialogue systems. Among many conferences that focus on this task, the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC) has emerged as a regular forum for presentation and discussion of the best research results in this area [23]. Initiated in 1996 at Lancaster University and taken over in 2002 by the University of Lisbon, and then moving out of Europe for the first time in 2009 (Goa, India), the DAARC conference series established itself as a specialized and competitive forum for the presentation of the latest results on anaphora processing, ranging from theoretical linguistic approaches, through psycholinguistic and cognitive work, to *corpora* studies and computational modeling.

In this section, we describe different approaches and cover some systems that we consider most influential on this complex task of resolving anaphora.

2.1 Centering

Centering is a theory about discourse coherence and uses focusing to order candidates and, according to Grosz *et al.* [20], is based on the idea that certain entities mentioned in utterances are more central than the others. Regarding anaphora resolution, this theory proposes that each utterance features a more prominent entity – the *center* – that is more likely to be pronominalised enhancing coherence. Since a discourse is not a mere sequence of utterances, it must have coherence, centering suggests the use of focus registers to keep track of the center of the utterances.

More recently, in 2007, Rosario applied the Centering approach to resolve Portuguese 3rd person nominative, accusative and dative pronouns (reflexive pronouns were not included) in juridic and journalistic text. The author reports an f-measure of 59.3%, which himself considers as “quite reasonable" and state of the art at that time [49]. This approach was evaluated in a *corpus* composed by legal and journalistic texts totaling 2,100 pronouns.

Whilst centering-based strategies have their ground on keeping focus registers [20], on many heuristic-based approaches such as Mitkov’s algorithm (see next section) and Kennedy and Boguraev’s parse-free approach [27], which base their antecedent preference on various factors, preference is given to the subject over direct and indirect objects since evidence suggests that the subject is usually the center [20, 29, 36]. Mitkov’s algorithm goes even further also considering a pre-defined set of verbs (*present*, *outline*, *etc.*) that transfer the focus to the direct object.

Since our approach will be based on a number of different factors, encompassing different theories, the idea of centering will translate into a subject-boosting factor, like it is done in several approaches such as Mitkov’s system (see section 2.2.2).

2.2 Syntax-based Approaches

Syntax-based approaches operate on the rules and principles that control sentence structure, typically represented by syntactic trees. In this section, we cover Hobbs’s naïve approach [24]; Mitkov’s algorithm [36] and Chaves and Rino’s adaptation of Mitkov’s algorithm for anaphora resolution in (Brazilian) Portuguese [6]. There are many others worth mentioning, such as Lappin and Leass’s Resolution of Anaphora Procedure (RAP) [29], that includes a binding algorithm to treat reflexive pronouns; salience weighting strategies, as well as rules to syntactically filter pronoun-NP co-reference; Kennedy and Boguraev’s parse-free approach [27]; Paraboni and Lima’s research on Portuguese possessive pronominal anaphora [44] or Pereira’s work on zero anaphora resolution in (Brazilian) Portuguese [45], but we will focus on the works most directly related to this study.

2.2.1 Hobbs’s approach

Hobbs’s approach [24] is a pronoun resolution method that operates on surface parse trees. The algorithm described below shows how Hobbs’s approach traverses the syntactic tree in a particular order, compiling a list of antecedents while excluding the ones that do not match the anaphor in gender or number.

Hobbs evaluated 300 pronouns from three texts with a variety of structures. The algorithm reached the success rate of 88.3% and, with the inclusion of selectional restrictions, it ascended to 91.7%. However, it is important to notice that the input was pre-processed and corrected by humans to avoid any pre-processing errors. Since most of the times there was only one plausible antecedent, Hobbs also evaluated the cases in which there was more than one candidate and the algorithm worked in 81.8% of those cases.

Hobbs also found out that 90% of the times the antecedent is on the same sentence as the anaphor and in 98% of the cases is on the same or the previous one (that is, long–distance anaphoric relations were very rare).

1. Begin at the NP node immediately dominating the pronoun in the parse tree of the sentence S;
2. Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it *p*;
3. Traverse all branches below node X to the left of path *p* in a left-to-right, breadth-first fashion. Propose as the antecedent any NP node encountered that has an NP or S node between it and X;
4. If the node X is the highest S node in the sentence, traverse the surface parse trees of previous sentences in the text in order or recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as the antecedent. If X is not the highest node in the sentence, proceed to step 5;
5. From node X, go up the tree to the first NP or S node encountered. Call this node X and call the path traversed to reach it *p*;
6. If X is an NP node and if the path *p* to X did not pass through the N-bar node that X immediately dominates, propose X as the antecedent;
7. Traverse all branches below the node X to the left of path *p* in a left-to-right. breadth-first manner. Propose any NP node encountered as the antecedent;
8. If X is the S node, traverse all branches of node X to the right of path *p* in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent;
9. Go to step 4.

Figure 2.1: Hobbs naïve approach.

2.2.2 The Mitkov Algorithm

In 2002, Ruslan Mitkov presented a knowledge-poor, heuristic-based, inexpensive and fast approach for pronominal anaphora resolution to meet NLP practical systems demands, and called it MARS (**M**itkov's **A**lgorithm **R**esolution **S**ystem) [36, p. 145]. The algorithm makes use of *antecedent indicators* to score and rank the candidates according to the likelihood of their being the correct antecedent of the anaphors. The author also defended that his approach is language-independent, presenting results from the extension of the method to other languages besides English, such as Polish or Arabic.

The first step of the algorithm, the pre-processing stage, uses a sentence splitter, a POS tagger and NP grammar rules to extract the noun phrases preceding the anaphor, from the current sentence and from the two previous sentences. The result is a set of NPs, which constitutes a list of candidates.

In the second step, Mitkov filters the list of candidates with number and gender agreement tests, discarding the ones that do not pass them. Collective nouns such as *government*, *team*, *parliament* that can be referred by anaphors in plural that do not match them in number (see Section 1.1.3) are excluded from the test. The result is a narrowed list of candidates.

In the third and final step, the *antecedent indicators* are applied giving the candidates positive or negative scores, increasing or decreasing the likelihood of their being selected. These indicators are listed below:

- *Collocation Match* (CM) – A score of +2 is assigned to those NPs that have an identical collocation pattern, that is, when the NPs and the pronoun follow the same patterns, *v.g.* <NP/pronoun, verb> or <verb, NP/pronoun>.

- *First Noun Phrases* (FNP) – A score of +2 is assigned to the first NP in a sentence. The subject, the theme of an utterance usually appears first and it is thus more likely to establish co-reference with a following anaphor;

- *Immediate Reference* (IR) – A score of +1 is assigned to those NPs appearing in constructions of the form '... (You) V_1 NP ...*conj* (you) V_2 it (conj (you) V_3 it)', where *conj* $\in$ {and/or/before/after/until...};

- *Indefiniteness* (I) – Indefinite NPs are assigned a score of –1;

- *Indicating Verbs* (IV) – A score of +1 is assigned to those NPs immediately following a verb that is an element of a predefined set (*assess*, *check*, *outline*, *present*, *etc.*). According to Mitkov [36, p. 146], empirical evidence suggests that NPs following these verbs usually carry more salience;

- *Lexical Reiteration* (LR) – A score of +2 is assigned to those NPs repeated twice or more in the paragraph in which the anaphor appears and a score of +1 to those repeated once in that paragraph;

- *Prepositional Noun Phrases* (PP) – NPs appearing in prepositional phrases are assigned a score of –1, considering that these are less salient than NPs and hence less prone to be antecedent of an anaphor;

- *Referential Distance* (RD) – NPs in the immediate antecedent clause, but in the same sentence as the pronoun are assigned a score of +2. Those in the previous sentence are assigned a score of +1. The NPs in the sentence

NP candidate	CM	FNP	IR	I	IV	LR	PP	RD	SHR	SI	TP	∑
original cover	0	1	0	0	0	0	0	1	0	0	1	3
original	0	1	0	0	0	1	0	2	1	0	1	6
original glass	0	0	0	0	0	0	-1	2	0	0	1	2

Table 2.1: Scores assignment to potential candidates of example (1.27).

that is two sentences apart from that of the anaphor are assigned a score of 0 whilst the NPs still farther are assigned a score of –1;

- *Section Heading Preference* (SHR) – A score of +1 is assigned to those NPs that also occur in the heading of the section in which the pronoun appears;
- *Sequential Instructions* (SI) – A score of +2 is applied to NPs in the NP_1 position of constructions of the form: 'To V_1 NP_1, V_2 NP_2; (Sentence). To V_3 it, V_4 NP_4';

  (1.) To turn on *the video recorder*, press the red button. To programme *it*, press the 'Programme' key.

- *Term Preference* (TP) – A score of +1 is applied to those NPs identified as representing terms in the genre of the text.

The *antecedent indicators* are summed, producing a number representative of its preference. The candidate with the highest total is selected as the antecedent. If two candidates have an equal score, the candidate with the higher score for IR is proposed. If IR does not hold, the candidate with the higher score for CM is proposed. If CM suggests a tie, the candidate with higher score for IV is selected. If this indicator does not hold again, the most recent candidate is chosen.

The algorithm can be summed up as follows:

1. Extract NPs to the left of the anaphor from the current sentence and from the two previous ones;
2. Discard the candidates that do not agree in gender or number;
3. Apply the antecedent indicators to each candidate and assign them scores; propose the candidate with the highest aggregate score.

Consider the following example provided by the author:

(2.1) **Positioning the original: Standard Sheet Original**
Raise the original cover. Place the original face down on the original glass so that *it* is centrally aligned against the original width scale.

Steps 1 and 2 of the algorithm generated the set of potential candidates as {*original cover*, *original*, *original glass*}. Step 3 assigns the scores to the following candidates as displayed in Table 2.1.

The noun phrase *the original* (score 6) is selected as antecedent for *it*.

Mitkov evaluated MARS on a set of eight technical manuals which contained 247,401 words and 2,263 anaphoric pronouns. MARS operated in fully automatic mode using the FDG-parser [55] as its main pre-processing tool, one of the best available at the time. A module for automatic identification of non-anaphoric occurrences of pronouns was also incorporated in the system. The overall success rate of MARS was 59.35%. Yet, if the anaphor is considered successfully resolved when the whole NP representing its antecedent is selected as such, the system achieves an average success rate of 80.03%. When considering that a pronoun is correctly resolved if only the part of the NP which represented the antecedent was identified, the average success rate ascends to a maximum success rate of 92.27%. However, further tests are needed since the success rate varied greatly in different manuals (between 51.59% and 82.67% in standard mode, when the anaphor is correctly resolved only when the candidate selected matches the antecedent; a candidate including the antecedent is considered incorrect). MARS operated on a manually corrected input.

Mitkov also applied two baseline models: selecting the most recent subject[1] and selecting a randomly chosen NP. The evaluation of these models indicated success rates of 37.78% and 31.82%, respectively. The difference in these results supports the notion that the antecedent indicators are effective.

Nobre [41] also adapted Mitkov's algorithm to resolve our problem – Portuguese pronominal anaphora (see section 2.3). However, even taking into consideration any possible errors occurred in the various pre-processing stages, the score of 33.5% cannot be considered reasonable.

2.2.3 The Mitkov Algorithm for Anaphora Resolution in Portuguese

Chaves and Rino presented RAPM, which stands for ***R**esolução **A**nafórica do **P**ortugês baseada no algoritmo de **M**itkov* (Anaphora Resolution for Portuguese based on Mitkov's algorithm) [6]. It is an adaptation of Mitkov's algorithm that is supposed to better fit the Brazilian Portuguese language. RAPM produces a candidate list from a three-sentence window that is narrowed down by gender and number agreement rules and then each candidate is scored by antecedent indicators, being the highest scored the selected one, just as MARS does. The main difference is in the set of antecedent indicators the authors used. The first five are reminiscent of Mitkov's algorithm (see the previous subsection for more details) while the latter three are novel:

- *First Noun Phrase* (FNP);
- *Indefiniteness* (I);
- *Lexical Reiteration* (LR);
- *Prepositional Noun Phrase* (PP);
- *Referential Distance* (RD);
- *Syntactic Parallelism* (SP) – A +1 score is issued to an NP that has the same syntactic function as the corresponding anaphor;
- *Nearest NP* (NNP) – A positive +1 score is issued to the nearest NP to the anaphor;

[1] Note that a model that would select the most recent NP candidate was not tested.

- *Proper Noun* (PN) – Proper nouns are scored +1.

Chaves and Rino explain the introduction of the new antecedent indicators: the SP factor could be applied since the *corpora* was morphosyntatically and parsed annotated (in Mitkov's resource-poor approach, no parsing was used); after analyzing the *corpora*, the authors noticed that a proper noun candidate tended to be chosen as antecedent (PN); the nearest noun phrase is also frequently the correct antecedent (NNP). The remaining Mitkov's indicators were discarded due to their inadequacy to the *corpus* under focus in this work.

To assess RAPM, three different *corpora* were used: a law, a literary[2] and a newswire one, containing a total of 1,055 3rd personal pronouns. No human pre-processing was done, which entails the selection of incorrect antecedents, chosen due to wrongly morphosyntactically tagged or incorrectly parsed input texts, unlike it was done in MARS. Different combinations of antecedent indicators were tested and the one containing the eight antecedent indicators listed above achieved the best results, with a success rate of 67.01%. This represents a 7.66% boost over standard-mode MARS.

2.3 Anaphora Resolution Module 1.0

In 2010, Nobre developed the Java-based ARM (**A**naphora **R**esolution **M**odule) 1.0 [41], which was integrated on the STRING NLP chain (see Section 5.1). Since we aim to develop ARM 2.0, it is only fitting that ARM 1.0 constitutes a baseline and it is presented in greater detail.

ARM 1.0 also implements Mitkov's knowledge-poor approach, which has been considered language-independent and achieved promising results in Portuguese (see Section 2.2.2). It receives as input the output of XIP in the form of an XML file and it works as a module fully integrated in the processing chain (for more information see section 5.1). This input file provides information such as *chunks* (*e.g.* noun phrases), *dependencies* (*e.g.* subject) and *named entities* (*e.g.* people, institutions).

2.3.1 Dependency Rules

To take advantage of recognizable patterns that support the existence of anaphora, Nobre implemented some dependency rules to locate and extract information from the XIP output.

Dependency rules are composed of three parts:

1. A regular expression pattern;
2. A set of conditions about relations between the nodes of a chunk tree or the nodes themselves, independent of the tree structure;
3. A dependency term.

Altogether, 17 rules were implemented that made possible to locate patterns evidencing the following dependency relations:

- *ACANDIDATE(1,2)*: token 1 is a possible anaphor of token 2;

[2]The literary *corpus* consisted of the whole book "O alienista", by the Brazilian author Machado de Assis, which includes dialogues.

- *ACANDIDATE_POSS(1,2)*: according to gender and number agreement used in implementation (see below), token 1 is the anaphor of token 2;

- *INVALID_ACANDIDATE(1,2)*: according to gender and number agreement used in implementation (see below), token 1 cannot be the anaphor of token 2;

- *IMMEDIATE_REFERENCE(1,2)*: token 1 is in immediate reference with token 2. In example (2.2), a relation, *IMMEDIATE_REFERENCE*, is created, between *a* and *Isabel*.

  (2.2) A Maria viu a *Isabel* and cumprimentou-*a*.
  Maria saw *Isabel* and greeted *her*.

2.3.2 Gender and Number Agreement

In addition to dependency rules, other rules were implemented concerning the correct identification of gender and number of compund nouns, *i.e.*, segments composed by more than one word yet that form a single semantic unit (*e.g. África do Sul/South Africa*). Proper nouns, especially, can be rather ambiguous. For instance, some proper nouns (given names) can be used in a masculine or feminine form (*e.g. João*) but family names do not have gender and can be used in the plural without any formal change but their determiner (*e.g.* "O *Silva*" and "Os *Silva*").

To approach these ambiguities, some rules were introduced on XIP:

1. In noun phrases or prepositional phrases, the gender and number of a noun is the same as the article determining them;

2. The number feature (singular, plural) of a compound noun, is the same as the one of its first noun.

2.3.3 Implementation

Like in Mitkov's approach (see previous section), ARM 1.0 operates on three steps:

1. *Anaphor identification*: 3rd person pronouns including *possessive*, *relative* and *demonstrative pronouns* are identified as possible anaphors. This represents a larger scope comparing to MARS and RAPM since neither of them covered this type of pronouns. The reflexive pronoun *se* was excluded at this time since its correct co-reference resolution is often verb-dependent. For example, it can correspond to an indefinite (non-anaphoric) pronoun (2.3); or it may refer to a post verbal subject NP in pronominal passive-like constructions (2.4).

   (2.3) Precisa-se de ajuda.
   Help is needed.

   (2.4) Vendem-*se casas*.
   Houses for sale.

Solving these syntactic and semantic issues requires much grammatical information that was not available at the time. Thus, in example (2.3), the indefinite interpretation/analysis of the reflexive pronoun *se* results from the fact that the verb *precisar* requires only a PP as its complement, and, since the verb is inflected in the 3rd person-singular, there is no other syntactic slot the pronoun can fill in. In example (2.4), the verb *vender* agrees in number with the NP *casas*, which has the semantic role of OBJECT; as no other NP is present with that number value, *casas* becomes parsed as the verb's subject, and the sentence is analyzed as a pronominal passive construction, where the reflexive pronoun has an anaphoric value.

Only verbs allowing this transformation authorize the passive, pseudo-reflexive (non-anaphoric) interpretation of the pronoun, thus the correct analysis is lexically dependent.

2. *Antecedent candidates identification*: ARM 1.0 identifies nouns and pronouns as antecedent candidates. Coordinated antecedents are taken into account: masculine pronouns can refer to exclusively masculine or mixed (masculine and feminine) coordinated NP antecedents, while feminine pronouns are restricted to exclusively feminine, coordinated NP antecedents. Possessive pronouns skip the gender-number filter (see section 1.1.2);

3. *Selection of the most likely antecedent candidate for each anaphor*: Several features are applied for boosting or penalizing a candidate's chance of being selected as the antecedent of an anaphor. The features used are based on the ones described by Mitkov's and Chaves and Rino's research (see previous section). The candidate with the highest aggregate score is selected as the antecedent. The following features and scores were considered:

    - *First Noun Phrase (FNP)*: a score of +1 is assigned to the first NP in a sentence;
    - *Collocation Match (CM)*: a score of +1 is assigned to those NPs that have an identical collocation pattern to the pronoun;
    - *Syntactic Parallelism (SP)*: an NP in a previous clause with the same syntactic role as the current is awarded a score of +1;
    - *Frequent Candidates (FC*): the three NPs that occur most frequently as competing candidates of all pronouns in the text are awarded a score of +1;
    - *Indefiniteness (IND)*: Indefinite NPs are assigned a score of –2;
    - *Prepositional Noun Phrases (PPN)*: NPs appearing in prepositional phrases are assigned a score of –1;
    - *Proper Noun (PN)*: a proper noun is awarded a score of +2;
    - *Nearest NP (NNP)*: the nearest NP to the anaphor is awarded with a score of +1;
    - *Referential Distance_0 (RD0)*: NPs in the previous clause, but in the same sentence as the pronoun are assigned a score of +2;
    - *Referential Distance_2 (RD2)*: NPs in two sentences distance are assigned a score of –1;
    - *Referential Distance_2+ (RD2+)*: NPs in more than two sentences distance are assigned a score of –3;
    - *Possessive Pronoun Probable Candidate (PPPC)*: a score of +1 is assigned to the candidate "C" if is present on an *ACANDIDATE_POSS(A,C)* relation for anaphor "A";
    - *Possessive Pronoun Invalid Candidate (PPPC)*: a score of –3 is assigned to the candidate "C" if is present on an *INVALID_ACANDIDATE(A,C)* relation for anaphor "A";

2.3.4 Genetic Algorithm

A machine learning approach was used to complement Mitkov's method, in order to train the indicators used and optimize their combination of weights. A genetic algorithm was implemented based on the work of Russel and Norvig [50]. In ARM 1.0, the individual was defined as a set of antecedent indicators and the fitness function was given by f-measure.

During the training phase, several newspaper articles available at Linguateca [51] were used, the model achieving an overall 41.61% f-measure. The values for the indicator features were the ones displayed in the list above.

2.3.5 Evaluation

To evaluate ARM 1.0, 8 texts were used from online forum messages, 1 from a legal *corpora* and 11 from news articles, containing a total of 692 pronouns, in which 334 of these were evaluated by the ARM. The system achieved 30% recall, 38% precision and 33.5% f-measure. The difference between recall and precision suggests that some resolution errors come from pre-processing stages. It may be possible that the more diverse nature of the evaluation *corpus* had an impact in these results.

2.4 Statistical Approaches

Also known as probabilistic, statistical approaches are based from processing statistical data retrieved on large annotated *corpora*, which helps to pick the antecedent from the list of candidates. These approaches emerged on the 1990s and the most influential works include Dagan and Itai's collocation patterns-based approach [9] and Ge, Hale and Charniak's statistical framework for resolution of third person anaphoric pronouns [17].

2.4.1 Collocation patterns-based approach

In 1991, Dagan and Itai presented an approach for resolving third person pronouns based on collocation (co-occurrence) patterns [9].

This approach was innovative since it presented an alternative to selectional restrictions that were under the spotlight at that time. Given that selectional restrictions were based in the assumption that both the anaphor and the antecedent must satisfy the same constraints, Dagan and Itai's system substituted the anaphor with each of the candidates and the candidate with most frequent co-occurence patterns is preferred over the others. To calculate these patterns, the *corpus* had to be processed in order to create the database. The database contained collocational patterns for the following pairs: "subject-verb", "verb-object" and "adjective-noun". The authors called this the 'acquisition' phase. It was followed by the 'disambiguation' step, which used the data collected in the first phase to select the antecedent.

To illustrate their system's behavior, Dagan and Itai used the following example taken from the *Hansard corpus*, used to build the statistical database:

subject-verb	collection	collect	0
subject-verb	money	collect	5
subject-verb	government	collect	198
verb-object	collect	collection	0
verb-object	collect	money	149
verb-object	collect	government	0

Table 2.2: Co-occurrence patterns associated with the verb *collect* based on an excerpt from the *Hansard corpus*

(2.5) They knew full well that companies held tax money aside for collection later on the basis that the government said *it* was going to collect *it*.

In order to resolve the two occurrences of *it* in the above sentence, statistics are gathered for the three candidates that arose: *money*, *collection* and *government*. Table 2.2 lists the patterns achieved when substituting each candidate with the anaphor and the number of times each pattern occurred on the *corpora*.

According to Table 2.2, the candidate picked for the first *it*, which is in subject position, is *government*; and for the second, which is in object position, is *money*.

The experiment was conducted on 59 sentences containing *it* that did not include non-anaphoric occurrences of this pronoun; only intersentential anaphoras were considered and trivial cases in which the anaphor had only one candidate were excluded. The examples where retrieved from a *corpus* of 28 million words. In 21 out of 59 examples, the system did not select a candidate since the threshold of 5 occurrences in the acquisition stage was not met. However, out of the remaining 38 sentences, Dagan and Itai's method selected the correct antecedent 33 times (87%).

While the method might be promising, the experiment was conducted on a very small sample and, thereby, further evaluation was needed to add significance to the results.

This method requires a *corpus* annotated with syntactic or semantic information as the one used by the authors. We have no knowledge of the application of this method to other languages than English, but we consider that it can also be applied to other languages since the statistics are retrieved from a *corpus* and the principle of collocation is language-independent. Another option would be to incorporate this method as a feature or even as a tie-breaker on antecedent factors-based approaches such as MARS.

2.5 Machine Learning Approaches

Machine learning approaches to the problem of anaphora resolution, more specifically to the problem of coreference resolution, have been reasonably successful, and, at first, operated mainly by modeling the problem as a classification task [40]. These strategies are based on weighted features that resort to various kinds of knowledge, as done in manual rules-based approaches, and offer the automation of the acquisition of knowledge from a *corpus*

by learning from a set of examples (patterns). More recently, other models have been explored for overcoming the classification model's weaknesses and achieving better results [39].

Traditional learning-based co-reference resolvers operate by training a model for classifying whether two NPs are co-referring or not ([5], [34], [40], [52]). In spite of their initial success, mention-pair models have two major weaknesses. Firstly, these models only determine how good a candidate is relative to the anaphor, but not how good a candidate is relative to other candidates. In other words, they fail to indicate which candidate is the most probable antecedent. Secondly, they present limitations on their expressiveness since the information extracted from the entities may not be sufficient to make a decision; *e.g.* 'Mr. Clinton' and 'Clinton' would be associated by sub-string matching, 'she' and 'Clinton' can be associated since there is lack of evidence of gender disagreement, and therefore the three NPs would co-refer when 'Mr. Clinton' and 'she' are clearly not co-referent due to gender disagreement [39]. These problems made way to new models such as the cluster-ranking framework, reported by Rahman and Ng in 2010, whose experimental results in co-reference resolution showed its superior performance to competing approaches [47].

As for the first weakness, ranking arguably presents a more natural way of formulating co-reference resolution than classification since it allows all candidate antecedents to be considered simultaneously, and therefore directly captures the competition among them, and the anaphor being resolved by the highest ranking candidate. To address the second issue, the use of cluster-level features, that is, features that are defined over any subset of mentions in a preceding cluster, increase the expressiveness of the model.

In the 1990s, the availability of annotated *corpora*, produced as part of the MUC-6 (*Message Understanding Conferences*) [19] and MUC-7 [60] conferences, gradually shifted the focus of co-reference research from heuristic approaches to machine learning approaches. Unfortunately, large-sized, anaphora-annotated, Portuguese *corpora* are still publicly unavailable.

The annotation of *corpora* is very laborious and time-consuming, specially if we take into account that annotating anaphoric links should be extended to the chain rather than only anaphor-antecedent pairs, since the task may be considered successful only when the anaphoric chain is resolved [36]. Yet, annotated *corpora* are indispensable to training and evaluating language models.

Machine learning methods applied in anaphora resolution typically include ID3, C4.5 algorithm [46] and clustering methods, although rule learners, memory-based learners, statistical learners and support vector machines [25] have been increasingly used [39].

In this section, we present the RESOLVE system [34], Cardie and Wagstaff's clustering algorithm [5], Soon's co-reference resolution of noun phrases approach [52] and Rahman and Ng's cluster ranker model for co-reference resolution [47].

2.5.1 RESOLVE System

McCarthy and Lehnert's RESOLVE [34] system uses the C4.5 decision-tree algorithm to learn how to classify co-referent noun phrases in the domain of business joint ventures. RESOLVE has its ground on a manually annotated text for co-referential noun phrases and on feature vectors pairing anaphors and antecedents. 1,230 feature vectors were created from the entities marked in 50 texts with 322 positive instances (26%) – co-referent pairings

– and 908 (74%) negative instances – non co-referent pairings. The following features and values were used:

- *Name*: Does a reference contain a name? Possible values {yes, no};
- *Joint Venture Child*: Does a reference refer to a joint-venture child, *e.g.* a company formed as a result of a tie-up among two or more entities? Possible values {yes, no, unknown};
- *Alias*: Does one reference contain an alias of the other, *i.e.* does each of the two references contain a name and is one of the names a substring of the other name? Possible values {yes, no};
- *Both joint venture child*: Do both references refer to a joint-venture child? Possible values {yes, no};
- *Common NP*: Do both references share a common NP? Possible values {yes, no};
- *Same sentence*: Do the references come from the same sentence? Possible values {yes, no}.

The MUC-5 [19] English Joint Venture *corpus* was used to evaluate the system, which scored 86.5% f-measure in the unpruned version, while the pruned version reached 85.8%. For both versions, all the pre-processing errors had been manually removed.

2.5.2 Cardie and Wagstaff's clustering algorithm

In 1999, Cardie and Wagstaff reported an unsupervised, domain-independent machine learning approach to resolve co-referential noun phrases based on a clustering strategy [5].

According to Cardie and Wagstaff, each NP is represented as a vector of attribute-value pairs. Given the feature vectors, the clustering algorithm coordinates the application of constraints and preferences to partition the NPs into equivalence classes, one for each real-world entity mentioned in the text.

The eleven features associated to each NP are as follows:

- *Individual Words*: The words contained in the NP are stored as a feature;
- *Head Noun*: The last word in the NP is considered the head noun;
- *Position*: NPs are numbered sequentially, starting at the beginning of the document;
- *Pronoun Type*: Pronouns are marked for case: NOMinative, ACCusative, POSSessive, or AMBiguous (*you* and *it*), all other NPs obtain the value NONE for this feature;
- *Article*: Each NP is marked INDEFinite (contains *a* or *an*), DEFinite (contains *the*), or NONE;
- *Appositive*: If the NP is surrounded by commas, contains an article, and it is immediately preceded by another NP, then it is marked as an appositive; otherwise, it is not;
- *Number*: If the head noun ends in an 's', then the NP is marked PLURAL; otherwise, it is considered SINGular;
- *Proper Noun*: Proper names are identified by looking for two adjacent capitalized words, optionally containing a middle initial;

Feature f	Weight	Incompatibility function
Words	10.0	(# of mismatching words)/(# of words in longer NP)
Head Noun	1.0	1 if the head nouns differ; else 0
Position	5.0	(difference in position)/maximum difference in the document
Pronoun	r	1 if NP_i is a pronoun and NP_j is not; else 0
Article	r	1 if NP_i is indefinite and not appositive; else 0
Words-substring	$-\infty$	1 if NP_i subsumes (entirely includes as a substring) NP_j
Appositive	$-\infty$	1 if NP_j is appositive and NP_i is its immediately predecessor; else 0
Number	∞	1 if they do not match in number; else 0
Proper Name	∞	1 if both are proper nouns, but mismatch on every word; else 0
Semantic Class	∞	1 if they do not match in class; else 0
Gender	∞	1 if they do not match in gender (allows EITHER to match MASC or FEM); else 0
Animacy	∞	1 if they do not match in animacy; else 0

Table 2.3: Features weights in Cardie and Wagstaff's machine learning approach.

- *Semantic Class*: Resorting to WordNet [53], a head noun is characterized as TIME, CITY, ANIMAL, HUMAN or OBJECT. A separate algorithm identifies NUMBERs, MONEY and COMPANYs;
- *Gender*: Also resorting to Wordnet, the gender can be MASCuline, FEMinine, EITHER ou NEUTER;
- *ANIMACY*: NPs classified as HUMAN or ANIMAL are marked ANIM; all other NPs are considered INANIM.

The distance metric is given by Eq. 2.1:

$$dist(NP_i, NP_j) = \sum_{f \in F} w_f \times incompatibility_f(NP_i, NP_j) \tag{2.1}$$

where F corresponds to the NP feature set described above; $incompatibility_f$ is a function that returns a value between 0 and l inclusive and indicates the degree of incompatibility of f for NP_i and NP_j; and w_f denotes the relative importance of compatibility with respect to the feature f (Table 2.3).

Terms with a weight of ∞ represent filters that rule out impossible antecedents: Two NPs can never co-refer when they have incompatible values for NUMBER, PROPER NAME, SEMANTIC CLASS, GENDER and ANIMACY features. Conversely, terms with $-\infty$ force co-reference with compatible values. When computing a sum that involves both ∞ and $-\infty$, the ∞ distance takes precedence, hence co-reference between the two noun phrases is discarded. Each NP is compared to all preceding NPs and, if the distance between two NPs is less than the clustering radius r, then their classes are merged into the same class, *i.e.* considered co-referential.

Regarding the evaluation, the value of r can affect the results: increasing r also increases recall, but decreases precision. In an evaluation on the MUC-6 coreference resolution *corpus*, Cardie and Wagstaff's clustering approach achieves the best f-measure of 53.6% with $r = 4$, which, at the time, was considered average to the task in

the MUC-6 evaluation but held promise as a machine learning system overcoming RESOLVE's 47% on the same dataset [34].

2.5.3 Soon's approach

In 2001, Soon *et al.* presented a learning approach to noun phrases co-reference resolution in unrestricted text. The module tries to establish co-reference between *markables*, textual elements which can be definite noun phrases, demonstrative noun phrases, proper names, appositives and sub-noun phrases. These are identified by a larger co-reference system also featuring sentence segmentation, tokenisation, morphological analysis, part-of-speech tagging, NP identification, named entity recognition and semantic class determination. The machine learning algorithm used to learn a classifier was C5, an updated version of C4.5 [46]. Soon *et al.* devised a twelve-feature vector for training and evaluation. The following features apply to two markables, *i* and *j*, where *i* is the potential antecedent and *j* the anaphor:

- *Distance*: Possible values are {0,1,2...}. If *i* and *j* are on the same sentence, the value is 0; if they are one sentence apart, the value is 1, and so on;
- *i-Pronoun*: Possible values {true, false}. If *i* is a reflexive (*himself*, *herself*), personal (*he*, *him*, *you*) or possessive (*hers*, *her*) pronouns, return true; else return false;
- *j-Pronoun*: The same process as described above, this time for *j*;
- *String match*: Possible values {true, false}. If *i* matches *j* return true; otherwise return false. The comparison is made without articles or demonstrative pronouns;
- *Definite Noun Phrase*: Possible values {true, false}. If *j* starts with the word *the* return true; else return false;
- *Demonstrative Noun Phrase*: Possible values {true, false}. If *j* starts with the word *this*, *that*, *these* or *those* return true; else return false;
- *Number Agreement*: Possible values {true, false}. If *i* and *j* agree in number return true; else return false;
- *Semantic class agreement*: Possible values {true, false, unknown}. *i* and *j* are in agreement if they are on the same semantic class (*e.g. Mr. Lim* and *he* both of the semantic class "male") or if one is parent of the other (*e.g. chairman* with semantic class "person" and *Mr. Lim* with semantic class "male"); *i* and *j* are in disagreement if their semantic classes are not the same and none of them is parent of the other (*e.g. IBM* with semantic class "organization" and *Mr. Lim* with semantic class "male"). If either semantic class is "unknown", the head noun of both markables are compared. If they are the same, return true; else return "unknown". Resorts to WordNet [53];
- *Gender Agreement*: Possible values {true, false, unknown}. If the gender of either markable *i* or *j* is unknown (*e.g. the president*), then the gender agreement feature value is unknown; else if *i* and *j* agree in gender, return true; otherwise return false;
- *Proper name*: Possible values {true, false}. If both *i* and *j* are proper nouns return true; else return false;

- *Alias*: Possible values {true, false}. The value of this feature is positive if both i and j are proper names that refer to the same entity;

- *Appositive*: Possible values {true, false}. If j is an apposition to i, return true; else false.

To train and evaluate their approach, Soon *et al.* used MUC-6 and MUC-7 data. They used 30 annotated training documents from MUC-6 and MUC-7 to train with 12,400 and 19,000 words, respectively. There were altogether 20,910 (48,872) training examples used for MUC-6 (MUC-7), of which only 6.5% (4.4%) are positive examples in MUC-6 (MUC-7). The system achieved a f-measure of 62.6% for MUC-6 (pruning confidence set at 20%) and 60.4% for MUC-7 (pruning confidence set at 40%). It is pertinent to notice that in this system, unlike McCarthy and Lehnert's RESOLVE system and Cardie and Wagstaff's clustering algorithm, the pre-processing is automatically made by the other modules and the input (markables) is not error-free. In fact, Soon *et al.* performed an error analysis concluding that the system identified the markables correctly 85% of the times.

Soon's *et al.* work is a benchmark in machine learning co-reference resolution, as the continuing research in the extension of their work proves [40] [2]. In 2002, Ng and Cardie conducted an experiment in NP co-reference that achieved f-measures of 70.4% and 63.4% in MUC-6 and MUC-7 data, respectively [40]. The improvements arose from extra-linguistic changes to the learning framework and a large-scale expansion of the feature set to include more sophisticated linguistic knowledge. More recently, in 2010, Broscheit *et al.* reported a toolkit-based approach to automatic co-reference resolution on German text, starting from Soon's *et al.* work, to show that machine learning-based co-reference resolution can be robustly performed in a language other than English [2].

2.5.4 Cluster Ranking Model

In Rahman and Ng's joint discourse-new detection and co-reference resolution approach, each training instance $i(c_j, m_k)$ represents a preceding cluster c_j (set of co-refering NPs) and a discourse-old mention (co-referent NP) m_k and consist of cluster-level features. The system used 39 features (describing the candidate m_k, the cluster c_j, and the relationship between them) largely the same as employed by Soon's approach and Ng and Cardie's extension of Soon's work. A training instance is created between each discourse-old mention m_k and each of its preceding clusters c_j and the class value assigned to $i(c_j, m_k)$ is 2 if m_k belongs to c_j, 1 otherwise. New-discourse mentions, that is, NPs that do not co-refer with any of the previous NPs, start a new cluster by creating an additional instance containing the features that solely describe the active mention and has the highest rank value among competing clusters (*i.e.* 2).

Regarding the learning stage, the mentions are processed from left to right. For each active mention m_k, test instances are created pairing it with each of its preceding clusters. To prevent the possibility of m_k being a new-discourse mention, a test instance was added containing features that only describe the active mention (similar to what was done in the training step above). If this additional test instance is assigned the highest rank value by the ranker, then m_k is classified as discourse-new and will not be resolved. Otherwise, it is linked to the cluster that has the highest rank.

For evaluation, 599 documents were selected from the ACE 2005 data set. A training set and a test set were defined following a 80/20 ratio. The baselines used for evaluation comparison included a mention-pair model,

System	Approach Type	Method	Type of anaphora
Hobbs's approach	Syntax-based	Parse-tree analysis	Pronouns
MARS	Syntax-based	Antecedent factors	3rd person personal pronouns
RAPM	Syntax-based	Antecedent factors	3rd person personal pronouns
ARM 1.0	Syntax-based	Antecedent factors	3rd person personal and possessive pronouns
Collocation pattern-based approach	Statistic analysis	Co-occurrence	3rd person pronouns
RESOLVE	Machine learning	C4.5 algorithm	Co-referent noun phrases
Cardie and Wagstaff's approach	Machine learning	Clustering algorithm	Noun phrases
Soon's approach	Machine learning	C5 algorithm	Noun phrases
Rahman and Ng's approach	Machine learning	Cluster ranking	Noun phrases

Table 2.4: Systems' features overview.

an entity-mention model, mention-ranking model and a pipeline cluster ranker. The results show that the joint cluster ranker outperformed the other approaches scoring 76% f-measure in true mentions (manually corrected) and 69.3% when the mentions are extracted automatically and, therefore, have an error associated. The best baseline (the system with the second-best results) is the pipeline cluster ranker which suggests that cluster ranker approaches are the state-of-art in learning-based co-reference resolution.

2.6 Overview

So far, we have presented different approaches to solving anaphora and outlined different strategies trying to resolve different types of anaphora. Now, we will compare the different systems and discuss the significance of the results achieved. Table 2.4 resumes the main properties of each system. Table 2.5 compares the evaluation subject, type of anaphora, *corpora* genre, number of anaphoras and the results scored by each system. It also indicates whether pre-processing errors were removed before testing.

From the outset, we make clear that one cannot compare straightforwardly the different systems since they try

System	Evaluation Target	Manually Corrected Input	Top Results
Hobbs's approach	100 pronouns from a history book; 100 pronouns from a literary book; 100 pronouns from newspaper	✔	91.7% success rate
MARS	Technical manuals	✔	92.27% success rate
RAPM	Law, literary and newswire corpora	✗	67.01% success rate
ARM 1.0	334 pronouns from 8 forum messages texts, 1 text from a legal *corpora*, 11 texts from news articles	✗	30% recall 38% precision 33.5% f-measure
Collocation pattern-based approach	Hansard corpora	✗	87% success rate
RESOLVE	MUC-5 English joint venture corpora	✔	86.5% f-measure
Cardie and Wagstaff's approach	MUC-6 co-reference resolution corpora	✔	53.6% f-measure
Soon's approach	MUC-6 and MUC-7 corpora	✗	62.6% f-measure
Rahman and Ng's approach	ACE data set	✔	76.0% f-measure

Table 2.5: Systems' evaluation overview.

different methods to resolve different types of anaphora and resort to different textual material to evaluate them. Besides, it is relevant to check if the system's input is pre-processed for it can contain errors that ultimately can tamper with the results scored. Though Mitkov defends that any problem resulting from pre-processing should be removed in order to evaluate exactly the performance of the anaphora resolver [36, p. 177], it can also be argued that a more realistic perspective of the AR task can be obtained if no human preprocessing of errors is performed, as this scenario corresponds better to the real setup of an AR system, which is intended to perform as much automatically as possible.

Given that MARS, Hobbs's, and Dagan and Itai's collocation pattern-based approaches are the best performing systems, according to scores on Table 2.5, even considering that Hobbs's approach rules out any pre-processing errors; it is impressive the success rate that the latter achieved, which consolidates this early research as an important benchmark on the scientific community. Mitkov suggests that in antecedent factors-based approaches such as his, the evaluation of each individual factor should be addressed to assess its real importance, and how the overall performance could improve through changing the weights of the factors [35]. Dagan and Itai's statistic filter can also successfully combine with other strategies, as it was done with Hobbs's algorithm, achieving a 10% boost on the success rate and a 3% increase when coupled with Lappin and Leass's RAP [36, p. 99].

Regarding machine learning systems, RESOLVE stands apart, with 86.5% f-measure. However, these results may not have much significance or, at least, seem highly domain-dependent, when we take into account that scores only an f-measure of 47% on the MUC-6 data set, on which other systems (Cardie and Wagstaff's, and Rahman and Ng's systems) have been evaluated with higher marks. Soon's approach is widely regarded as a reference in anaphora resolution through machine learning methods, dealing with different kinds of noun phrases and offering the idea that learning approaches held promise. Despite the success that Soon's approach enjoyed, the mention-pair model used remains fundamentally weak. Its subpar performance promoted the research upon new models that could address the weaknesses exhibited by that model. In this way, Rahman and Ng's cluster ranker approach scores a f-measure of 76% with no NP extraction errors, but still scores 69% with automatic NP extraction, outperforming the baseline mention-pair model used to comparison by 5.6% and 6.8%, respectively.

When assessing Nobre's work on Portuguese pronominal anaphora based on Mitkov's algorithm and its f-measure score of 33.5%, and comparing it against the systems scores on Table 2.5, it is undeniable that the results are well below the usual standards. Firstly, we will start by trying to find the reasons of such surprising discrepancy.

Chapter 3

Annotation Tools

SINCE the early 1990s, research in anaphora resolution has benefited from the availability of *corpora*, raw or annotated, despite the benefits in that time were limited to collocation patterns extraction, as used in Dagan and Itai's work (see Section 2.4). Nowadays, annotated *corpora* are widely used for training machine learning algorithms (see Section 2.5).

Annotating anaphora is a difficult, time-consuming and labor-intensive task, even when focusing on a single variety of the phenomenona, and bearing in mind that annotators do not always agree in the choice of the correct antecedent for an anaphor. For instance, the MUC co-reference annotating scheme has been target of criticism. Current "co-reference" annotation practice, as exemplified by MUC, has overextended itself, mixing elements of genuine co-reference with elements of anaphora and predication in unclear and, sometimes, contradictory ways. As a result, the annotated *corpus* emerging from MUC is unlikely to be as useful for the computational linguistics research community as one might hope, the more so because generalization to other domains is bound to make problems worse [56].

Certain features are required from an annotation tool. One of the most important is that the tool be free, so it can be widely distributed and used as a standard, common tool. Other factors are also relevant: the annotation level (words, sentences or any segment of text); the possibility to relate the annotated units (such as anaphora or subject relations); the formats with which it is compatible; if it is possible to filter views by the type of annotation or coloring to make it easier to distinguish different units or different relations. All these features help to optimize the annotation process.

Next, we present the CorpusTool [42], Glozz [61], Knowtator [43] and MMAX 2 [37] annotation platforms. Other annotators were considered, namely Domeo [7] and RapTat [18], recommended for annotating biomedical tasks [38], but those four are the ones that were deemed as more adequate to the task of anaphora annotation.

3.1 CorpusTool

CorpusTool[1] is an annotation framework designed by Mick O'Donnell [42] that became available in February, 2007. It is aimed for linguists that do not have experience in programming so that they spend their time annotating text instead of learning how to use the program.

[1] http://www.wagsoft.com/CorpusTool/ (access date: 27/08/2013)

Previewing annotator needs of working with multiple *corpora*, it provides a 'Project Window' to manage different source files as well as adding annotation layers to them. It also provides a graphical tag scheme that, when changed, propagated these changes throughout all the files annotated in that layer. Any segments of text can be annotated and the annotated files are stored on XML files. The tool also provides a text exploring mechanism, as well as statistical information that can be displayed on different views. Recent versions already include structural tagging, this is, relations between the text segments (*e.g.* co-referential links), a limitation in the earlier versions.

The tool is free but not open-source. It is available for Windows and Macintosh and it is rapidly spreading.

3.2 Glozz

Glozz[2] is a free Java-based annotation platform developed by Antoine Widlöcher and Yann Mathet [61]. It is not open-source.

The annotation process on Glozz is based on the concepht of **U**nits-**R**elations-**S**chema (URS) model:

- *Units*: a contiguous span of text starting at one character position and finishing at another one; units can overlap each other or even include others. Glozz also provides an option to consider words as atoms facilitating annotations where words are the 'smaller' segments;
- *Relations*: a link, oriented or not, from one element of the URS model (a unit, a relation or a schema), to another element;
- *Schemas*: a set of as many URS elements as wished. For instance, a given schema can contain some units, but also some relations, and even some other schemas. This enables the construction of recursively deep structures.

Each annotation element can be associated with a feature-value set. The features and possible values to be assigned are set through a simple XML file. Glozz also provides some features that the annotator can find very useful such as different colors for different relations to help the annotator make the distinction between URS elements in a single glance or the possibility to hide some annotation elements to facilitate the analysis of other ones. The annotator can create his own style and customize it according to his liking.

Glozz permits the user to import his .txt file and creates a set of XML files including the annotated *corpus* and the annotation model.

3.3 Knowtator

Knowtator[3] [43] is a free (not open-source) general-purpose text annotation tool that is integrated with the Protégé [28] knowledge representation system, suitable for NLP systems. Builded on the strengths of the widely used Protégé, Knowtator has been developed as a Protégé plug-in that makes use of its knowledge representation capabilities to specify annotation schemas.

Knowtator provides a 'fast annotation' mode that can be very helpful in accelerating the annotation process. Besides, it is easy to define complex annotation schemas and incorporate them. The color assignment to each unit

[2] http://www.glozz.org/ (access date: 27/08/2013)

[3] http://knowtator.sourceforge.net/ (access date: 27/08/2013)

and its customization is another plus. The user can also export the annotation to a simple XML file. However, in our opinion, this tool presents some problems such as the fact that the relations are not displayed by an arrow which would be the most intuitive approach. Instead, relations are represented by way of a different border on the units. The inability to filter the units we want to see is another disadvantage when compared against other tools. In Knowtator, words are the shorter segment of text markables.

3.4 MMAX 2

MMAX 2[4] is a Java-based free annotation tool developed by Christoph Müller and Michael Strube [37]. It is not open-source.

MMAX 2 supports an arbitrary number of levels of annotation, each of which resides in a separate file, as well as relations between them. In this tool, the schema is defined via an XML file. A MMAX 2's feature that many annotators may find useful is the flexible and customizable display where the annotator define the colors of an annotation element and its style (plain, bold, italic). Plus, the user can hide some elements to better assess others. MMAX also provides a project wizard but the shorter segment a user is allowed to annotated is a word. The annotation files are saved in XML format.

3.5 Overview

Bearing in mind that text annotation is a time-consuming and laborious task, the use of an annotation tool is critical in helping a human annotator. There are a number of features that an annotation framework should provide to facilitate the annotator's job. In other words, it should allow a number of options starting with a friendly graphical interface to allow the human annotator an efficient interaction with the annotated text. Moreover, many view an annotation tool and its input and output formats indispensable so to adapt the tool to their own NLP application. The possibility of coloring units as to better distinguish them or to hide some units or relation types can also be very helpful when dealing with diverse types of annotations (*e.g.* see only direct anaphors in a text also annotated with pronominal and locative anaphora). Table 3.1 sums up the properties of the annotation tools studied.

Our search for an annotation tool focused on annotating anaphoric chains but not only. On the long run, this framework should be able to deal with several types of elements and relations (*e.g.* proper nouns, subjects, zero anaphora, 3rd person pronouns) which increases the importance of the 'coloring units' and 'hiding units' features. Besides, considering the anaphora annotating task in Portuguese, it is important to annotate parts of word since some anaphors in the Portuguese language appear in the form of pronouns attached to the verb (1.9 and 1.10) which are not always tokenized as individual words.

According to Table 3.1, Glozz and MMAX 2 stand apart as the more complete frameworks. We choose Glozz, which we considered to have a friendlier interface. This is especially important since the annotation framework now chosen will be used in years to come, to annotate different things by different people.

[4] http://mmax2.net/ (access date: 27/08/2013)

System	Free	Open source	Annotation Level	Relations	Annotation Format	Units Colors	Hiding options
CorpusTool	✔	✗	Any piece of text	✔	XML	✗	✗
Glozz	✔	✗	Any piece of text	✔	XML	✔	✔
Knowtator	✔	✗	Words & sentences	✔	XML	✔	✗
MMAX2	✔	✗	Words	✔	XML	✔	✔

Table 3.1: Annotation tools' features overview.

Chapter 4

Corpus

This chapter introduces and describes the golden standard *corpus* we built and the way we approached the annotation process, from the annotation model chosen and the annotators and their skills to the process of annotation itself and the way it has evolved according to the assessment that took place along the way.

4.1 Golden standard *corpus*

To develop a machine learning approach to anaphora resolution, we needed to build a *corpus* annotated with anaphoric relations to supply the training instances to the system and to serve as a golden standard for the system's evaluation. In this chapter, we describe in detail the *corpus* created for these purposes.

The dataset that will be used to train and evaluate our system is a fragment of the European Portuguese LE-PAROLE *corpus* [14]. The *corpus* is quite heterogeneous, being composed of texts from different genres: novels, pieces of news, magazine news and newspaper columns, among others. In total, it contains 290,000 tokens.

The *corpus* was automatically POS-annotated by Palavroso and STRING, and manually corrected. The initial PAROLE tagset was adapted to the STRING functionalities and linguistic specifications [32] and consists now of 12 categories of POS labels (noun, verb, adjective, pronoun, article, adverb, preposition, conjunction, numeral, interjection, punctuation and symbol) and 11 fields (*scilicet*, category (CAT), subcategory (SCT), mood (MOD), tense (TEN), person (PER), number (NUM), gender (GEN), degree (DEG), case (CAS), syntactic features (SYN), and semantic features (SEM)). No category uses all the fields.

The annotation campaign identified 9,268 anaphoric relations (94.3%) and 560 cataphoras (5.7%). The breakdown of the anaphoras by anaphor type is shown in Table 4.1:

The type of anaphor was identified based on the NLP chain STRING output (see section 5.1). This comprises an error margin that is associated with the annotation errors, as some anaphors were identified with a unexpected POS type (such as prepositions *a*, for instance). There were 7,001 anaphoras (75.5%) that had the antecedent in the same sentence, while 2,267 (24.5%) did not. From these, 1,028 anaphoras had the antecedent in the previous sentence, 364 with a sentence between, and 223 with two sentences between. This hints that the majority of anaphoras do not surpass the three-sentence distance between anaphor and antecedent. The annotated anaphora in which the anaphor is farther from the antecedent reports a distance of 146 sentences between them.

Type of anaphor		Number of anaphoras	Percentage
Pronouns	Relative	3,663	39.52%
	Personal	3,470	37.44%
	Possessive	689	7.43%
	Indefinite	607	6.55%
	Demonstrative	188	2.03%
	Total	8,617	92.97%
Articles		338	3.65%
Numerals		74	0.80%
TOTAL		9,268	100%

Table 4.1: *Corpus* anaphoras composition.

A rapid analysis of Table 4.1 confirms that pronouns are the most representative category of anaphors, particularly the personal and relative pronouns.

4.2 Annotation Process

Attending to the time-consuming nature of the process of annotating *corpora*, we considered that this should not be a one-person task. Thus, it was necessary to define an annotation model to guarantee the consistency of the whole process. In other words, it was necessary to make sure that each and every annotator performs this task in the same way.

4.2.1 Annotation Model

In using Glozz platform [61], we relied on URS model (see Section 3.2). Regarding anaphora, we proposed the following units, relations and color scheme to be used in the annotation process[1]:

- *Units*: head of noun phrases (red), head of prepositional phrases (red), pronominal anaphors (yellow) and verb phrases (green);
- *Relations*: anaphora, *i.e.*, an oriented arc from the anaphor to its antecedent (blue);

The annotator operates on a *corpus* already annotated with the units mentioned above, thought to be the most useful for the task, and s/he just had to annotate the anaphora relations as described above.

To improve the consistency of the process, specific guidelines were devised in order to clearly state the general principles governing the annotation campaign (and to be renewed/reviewed if necessary). Though these general principles are provided in full detail in Appendix A, one can already state some basic annotation schemata. Thus, we define that zero anaphora should not be annotated at this time. In the case of coordinated antecedents, an anaphoric relation should link the anaphor to each of the antecedents that compose the coordinated antecedent.

[1] In the Glozz platform different color codes can be defined for both the units and the relations. This was also done in this campaign, and the color codes chosen in such a way that it would help the annotators identify the markables.

Furthermore, when two (or more) antecedents refer to the same entity, the closest one should be preferred over the others.

The guidelines also present some particularly problematic situations and the solutions adopted for each case.

4.2.2 Annotators and their qualifications

The annotation process was carried out by 5 annotators with credentials and expertise in Portuguese Linguistics and NLP:

Anotator #1 has a PhD in Linguistics and has large experience in NLP;

Anotator #2 is graduated in Information Systems and Computer Engineering and is the author of this MSc thesis;

Anotator #3 has a PhD degree in Electrotechnic and Computers Engineering and has large experience in NLP;

Anotator #4 is graduated in Linguistics and is currently finishing his MA in the same area;

Anotator #5 is graduated in Linguistics and is currently finishing his MA in the same area; s/he also has a lot of experience in *corpora* annotation.

4.2.3 Inter-annotator Agreement Calculation

In order to calculate the inter-annotator agreement, we partitioned the *corpus* into $5+1$ parts. Each annotator took the task of annotating one of the parts, but before that, all annotators worked on the same part to infer the *Fleiss' kappa coefficient* (k) [16]. k is a statistical measure for assessing the reliability of agreement between a fixed number of raters/annotators when assigning categorial ratings to a number of items or when classifying items. The Kappa coefficient is defined by the following equation (4.2):

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4.1}$$

where $Pr(a)$ is the relative observed agreement among annotators, and $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly attributing each category rate. k varies between a maximum of 1 (total and complete agreement between annotators) and 0, when there is no agreement other than what would be expected by chance (as defined by $Pr(e)$).

However, taking into account the specificity of anaphora annotation, particularly the fact that there is no fixed number of categories since the number of candidates vary in each case, it is not possible to calculate $Pr(e)$ using the observed data. Therefore, the general formula of k was adapted as follows: let N be the total number of anaphors, let n be the number of annotators, and let c be the number of candidates for each anaphor. The anaphors are indexed by $i = 1,...N$ and the candidates are indexed by $j = 1,...c+1$, where $c+1$ represents the case where an anaphor has not been annotated. Let n_{ij} represent the number of raters who assigned the $i-th$ anaphor to the $j-th$ candidate. The k calculation will thus take the form of equation 4.2

$$k = Pr(a) = \frac{1}{N}\sum_{i=1}^{N} P_{\mathrm{i}} = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{c+1} n_{ij}^2 - Nn\right) \tag{4.2}$$

where P_i is the extent to which raters agree for the ith subject (*i.e.*, compute how many rater-rater pairs are in agreement, relative to the number of all possible rater-rater pairs).

After the development of the first version of annotation directives, a copy was handed to each annotator as each one was asked to annotate a set of 12 magazine articles (14,856 tokens, 566 anaphoras, 48 cataphoras). k reached 49.8% which was considered an unreliable value. In view of this result, the annotations were compared and corrected in a sequence of group meetings. The main points of disagreement were (i) the presence of apposition (some annotators indicated the apposition and others the main NP as the antecedent), (ii) the selection of different antecedents from the same anaphoric chain. All these were settled by consensus, eventually refining the initial formulation of the directives to clarify any less precise indication.

The performance results of each rater was as follows:

Annotator	1st experiment accuracy	2nd experiment accuracy	Improvement
#1	75.0 %	81.9%	+6.9%
#2	76.7 %	85.1%	+8.4%
#3	72.2 %	82.4%	+10.2%
#4	54.4 %	56.9%	+2.5%
#5	77.9 %	88.3%	+10.4%

Table 4.2: Annotators' performance on the first and second experiments.

After updating the annotation directives renewal/update, a new set of 32 news articles (15,590 tokens, 522 anaphoras, 47 cataphoras) was selected and the annotators were asked to annotate according to the new , improved version of the guidelines. Then the agreement was calculated again to verify whether it was acceptable this time. The five annotators now achieved a k of 70.8%, which represented a major improvement from the first annotation experiment. According to statistics displayed in Table 4.2 and retrieved after a similar process of correction as the one occurred in the first experiment, Annotator #4 was considerably below the standards achieved by the other raters. Without Annotator #4, k rises to 78.7%, which can be considered as a reliable value. Therefore, it was decided that Annotator #4 was excluded from the annotation campaign and his/her part was delivered to Annotator #5. Most of the remaining *corpus* annotation stayed in charge of Annotator #2, the author of this thesis, and Annotator #5, the consistently most accurate annotator.

Still, after the second experiment, some extra guidelines were added to the annotation directives in order to improve even further the quality of the *corpus* annotation. This time, the attributive constructions as anaphors and intervals as cataphors were discarded (for the fully detailed annotation directives, see Appendix A).

Chapter 5

Architecture

In this chapter, we describe the solution we proposed to solve the problem of pronominal anaphora in Portuguese. The system will operate on the output of L^2F NLP chain – STRING [31] (see Section 5.1). Basically, we developed an hybrid approach: a rule-based approach to retrieve the anaphors and respective potential antecedent candidates; a model based on a machine learning approach, which required a *corpus* annotated with anaphoric relations (see previous chapter), to select the most probable candidate for antecedent. We experimented, tweaking and combining different features to assess the system and determine the version that produces the best results.

5.1 STRING

STRING (**ST**atistic and **R**ule-based **N**atural lan**G**uage processing chain) [31] is a NLP chain developed by the Spoken Language Systems Laboratory (L^2F of the INESC–ID). In its current architecture, it has 7 modules as shown in Figure 5.1. The different applications are written in different programming languages (C++, Java and Perl).

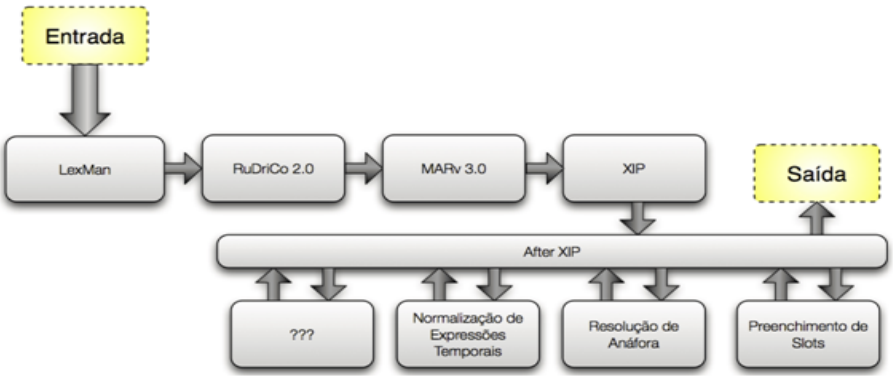

Figure 5.1: Natural Language Processing chain STRING

The first module is called LexMan (**Lex**ical **M**orphological **An**alyzer) [13] [58] and its function is to split the input into sentences and tokens, as well as to identify words, compound words, punctuation marks, numbers, among other textual units. It is also responsible for tagging these tokens with their potential POS and their

grammatical values (category, subcategory, number, gender, time, mood, *etc.*).

RuDriCo (**Ru**le **Dri**ven **Co**nverter) [12] is mainly responsible for disambiguating the word forms tokenised and tagged by the previous module. It uses disambiguation rules to select the correct POS tags for a given word, considering its context. It also performs the reconstitution of the lexical units in contracted words (*e.g.* 'nas': 'em' + 'as') and, finally, it is also in charge of grouping certain compound lexical units into a single token.

The next module is MARv (**M**orphossyntactic **A**mbiguity **R**esol**v**er) [48], a probabilistic disambiguation tool resorting to a 250k word POS-annotated *corpus*. It selects a single POS tag for each word using the Viterbi algorithm [59] to select the most adequate tag for each word. MARv 3.0 makes use of trigrams to provide contextual information and unigrams which codify lexical information.

XIP (**X**erox **I**ncremental **P**arsing) [1] splits sentences into syntactic segments – chunks – and resolves, sentence by sentence, their dependencies (subject, direct complement, *etc.*). XIP also applies local grammars, to identify certain productive multiword units, and morphosyntactic disambiguation rules, besides adding lexical/syntactic/semantic information.

"After XIP" module comprehends a few modules that operate on top of XIP, that is, they make use of the information provided by XIP and the previous modules but their output does not serve as an input or dependency for other modules to operate on.

Anaphora Resolution is the module this work will focus on. Its purpose is to identify anaphoric expressions and to point out, from the anaphoric chain, the closest antecedent to which that expression refers.

For instance, in example (5.1)

> (5.1) *O Eng. Pedro Matos* casou com a *Ana Silva* em 22 de Abril de 2009. *Ele* doutorou-se no Instituto Superior Técnico em 2008. *A Ana* licenciou-se no ISCTE em 2004 e trabalhou na TAP entre Janeiro de 2004 e Dezembro de 2005. *Ela* trabalha na Microsoft desde Janeiro de 2006.
> *The Engineer Pedro Matos* married Ana Silva on April 22, 2009. *He* received his doctorate at the Instituto Superior Técnico in 2008. *Ana* graduated from the ISCTE in 2004 and worked in TAP between January 2004 and December 2005. *She* has worked at Microsoft since January 2006.

the output is

Anaphor - Ele	Antecedent - O Eng. Pedro Matos
Anaphor - Ela	Antecedent - A Ana

Finally, we have the modules for Time Normalization [21] [22] [30] [33] and Slot Filling [4]. Time Normalization includes the tasks of normalizing temporal references whereas Slot Filling concerns the identification of entities, information retrieval about them and the relations between them and grouping all that information by entities. Future modules are being developed to complement these, namely, for event identification and ordering [3] and semantic role labelling [54].

Next we present the Anaphora Resolution Module that is the focus of this dissertation.

5.2 Anaphora Resolution Module 2.0

We experimented an hybrid approach to anaphora resolution. As discussed in Section 2.5, new models have been proposed recently such as the cluster ranking model [8] [11] [47] [57], with promising results. However, it is important to notice that most learning methods deal with co-reference while we focus only on anaphora resolution.

We tried to make the most out of every piece of linguistic knowledge that STRING provides us. After all, this is the promise that machine learning methods offer: automating the acquisition of a large amount of knowledge from *corpora* (by learning a set of examples), an impediment to the development of robust knowledge-based systems [36, p.87].

Firstly, we will try to apply to our task the ranker model, an approximation of the cluster ranker model reported by Rahman and Ng [47], as their results are advanced as state of the art (see ection 2.5). Since we are not treating co-reference and we already have the NPs provided by XIP and the anaphors by the 'Anaphor extractor' from the rule-based system, we do not need to extract mentions (NPs). Moreover, resolving an anaphor is different than resolving co-reference, since an anaphor will always refer to an antecedent and, hence, will always belong to a cluster containing that antecedent. Therefore, there are no 'new-discourse mentions' to process and the cluster "part" of Rahman and Ng's approach does not apply. The ranker will remain the same, though, as the goal is to indicate which candidate is the most probable. As for the feature set to use, it will be adapted according to experimental results and it will include features describing the candidate, the anaphor and the relationship between them. This seems a good starting point as it is also done in the learning methods studied.

The annotated *corpus* (see section 4.1) will serve as a golden standard for the system's evaluation.

We can divide the problem of anaphora resolution in three stages:

- Identification of anaphors;
- Compilation of the list of candidates;
- Choice of the most probable candidate.

Regarding the first two stages, we devised a rule-based system whose result, *i.e.*, the anaphors and the list of candidates, will serve as the input to a learning model that will order the candidates by the probability of their being the anaphor's antecedent. The most probable candidate will then be selected as the antecedent of the anaphor.

5.2.1 Anaphor Identification

Our program uses the information (particularly syntactic information) provided by STRING in order to identify a token as an anaphor. Through the parsing of the text, the nodes that are retrieved as anaphors are:

- Articles that constitute a single node, that is, articles that are not incorporated in NPs or PPs (Fig. 5.2);

  (5.2) Duas universidades: *a* de Lisboa e *a* do Porto.
  Two universities: the *one* from Lisbon and the *one* from Porto.

- Nodes named "REL" in STRING are also retrieved, as they represent relative pronouns;
- Pronouns incorporated on a NP or PP that do not violate any of the following rules:

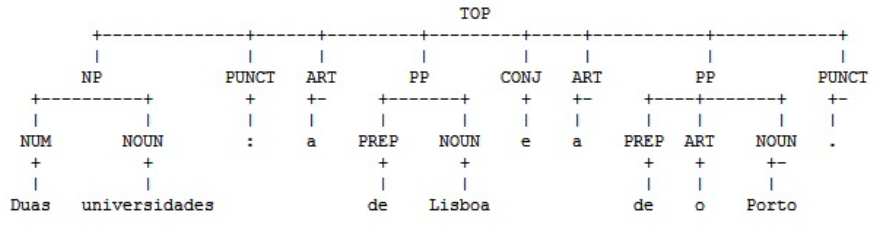

Figure 5.2: Parse tree produced by STRING that shows two articles as nodes not incorporated in a NP or PP.

- Pronouns cannot be 1st or 2nd person. 1st or 2nd person pronouns refer to the participants in a dialog, and are not addressed in this dissertation;
- Pronouns cannot be a predicative subject, that is, the pronouns that are preceded by the Portuguese verb *ser* (in English, the "to be" verb). This rule is detected through the dependency "PREDSUBJ" that features the verb and the pronoun as parameters.

  (5.3) O jogo decisivo é este.

  The decisive match is this one.

  In the example (5.3), demonstrative pronoun *este* is preceded by *é* which in Portuguese is the conjunction of verb *ser* in the 3rd person, present form. The dependency "PREDSUBJ" links the verb and pronoun triggering the rule excluding then the pronoun as an anaphor;
- Only if the pronoun is not demonstrative, indefinite or possessive, can it be present in a coordination. This rule intends to rule out pronouns that are determining a noun but do not have present that (pronoun–noun determining) dependency as it occurs with the other pronoun whom it shares the coordination. In example (1.34), this rule excludes *Estas* (These) to be identified as an anaphor;

  (5.4) Estas e outras coisas são perigosas.

  These and other things are dangerous.
- Indefinite pronoun *se* is ruled out as it is not anaphoric (5.5). Also, *se* clitic pronouns attached to a verb with "PASS-PRON" feature, corresponding to the pronominal passive-like construction, are discarded as they are being used in a expletive way (5.6).

  (5.5) Dizia-se que era uma decisão irrevogável.

  It was said to be an irrevocable decision.

  (5.6) Elas aborreciam-se de morte a ver telenovelas.

  They hated to death watching soap operas.

- Cardinal Numerals, irrespective of their form as words or algarisms, that agree with all of the following rules:

  - Cardinal numerals that are the head of the node (5.7); otherwise, if they are determining a noun, they are excluded (5.8);

    (5.7) O *Pedro* e o *Rui* são os melhores amigos. Os *dois* vão juntos para a escola.

    Pedro and *Rui* are best friends. The *two* go together to school.

(5.8) Os dois irmãos são muito parecidos.

The two brothers are very much alike.

– Cardinal numerals that have the feature “TIME” (5.9), the feature “CURR" (currency) (5.10) or contain an “-" (5.11) are not identified as anaphors.

(5.9) O 25 de Abril foi há muito tempo.

April 25th was a long time ago.

(5.10) Nos dias de hoje, uma cerveja custa 1 euro.

Nowadays, a beer costs 1 euro.

(5.11) O resultado fixou-se nos 3-0.

The result was set at 3-0.

– Cardinal numerals present in a coordination are ruled out, since we do not consider the intervals (even if we did, they would be a cataphor) (5.12).

(5.12) Irei à praia entre as duas e quatro horas.

I will go to the beach between two and four o’clock.

- Ordinal numbers that are not determining any other node are identified anaphors. STRING already makes the distinction between the use of the word as an adverb (Fig. 5.3) and as a numeral (Fig. 5.4).

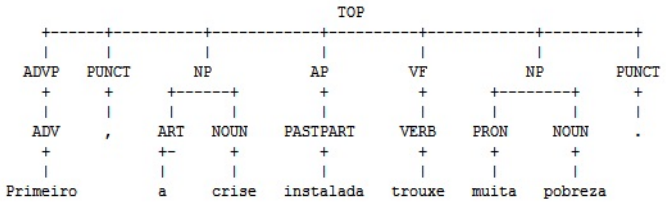

Figure 5.3: STRING correctly identifies *primeiro* as an adverb.

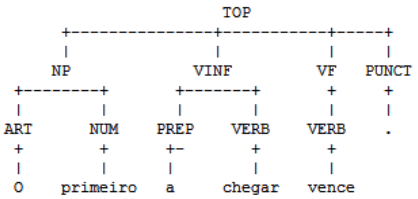

Figure 5.4: STRING correctly identifies *primeiro* as a numeral under a NP.

Also, we compiled a list of indefinite and demonstrative pronouns that are traditionally considered not to be used in a non-anaphoric manner and, therefore, are automatically excluded as anaphors. This list contains the following words:

- *isto*, *isso*, *aquilo* (this, that, that);
- *o porquê*, *o como* (because, how);
- *toda a gente*, *alguém*, *algo*, *ninguém*, *nenhum* (everybody, someone, something, nobody, none);

- *tudo*, *nada* (everything, nothing);
- *algures*, *nenhures* (somewhere, nowhere);
- *mesmo*, *o tal*, *um certo*, *próprio* (the same, such, certain, self);

5.2.2 Compilation of candidate list

Like the anaphor identification stage, the candidate identification is also made throughout the parsing of the text. Nouns that are heads of NPs and PPs are identified as potential candidates. When STRING identifies that two or more nouns are present in a coordination, they also constitute a coordinate candidate (5.13).

(5.13) O *João* e o *Pedro* foram a casa da Rita.

João and *Pedro* went to Rita's home.

Besides, if a pronoun is (left-side) closer to a relative pronoun anaphor than any other candidate, it is also identified as a candidate for that anaphor to prevent cases such as (5.14).

(5.14) Foi *aquilo que* nos levou a agir assim.

It was *that what* made us act like that.

At last, the span of text from which the candidate list of an anaphor is to be retrieved is limited to a two sentence window – only the candidates that are on the same sentence at the left of the anaphor, or in the previous two sentences, are selected. Exception is made to the relative pronoun anaphors, whose candidates must be selected from the same sentence and at the anaphor's left side[1].

5.2.3 Selection of the best candidate

The ordering of the candidate list (and the choice of the most probable one) is based on the model generated through the application of a machine learning method applied to the *corpus* we annotated. To do this, we used the WEKA software[2] [62].

Our system identifies the anaphors and candidates for each anaphor, and creates an instance for each pair anaphor-candidate with several features displayed in Table 5.1 in page 44. As we implemented a supervised learning (based on the annotation), each instance contains the target feature (T) *is_antecedent* that could be either *true* if the candidate is the antecedent for the anaphor, or *false* otherwise. The remaining feature values are retrieved using STRING. The features are grouped in three types: anaphor-related features (A), candidate-related features (C) and features related to the relationship between anaphor and candidates (R). In particular, *anaphor_gender*, *anaphor_number*, *candidate_gender* and *candidate_number* can present an "IND" value which means that the gender/number of anaphor/candidate is indefinite.

We also had to define the machine learning method adequate to our task. Since we want to be able to order the candidate list in order to pick the best candidate, we could not use a classifier due to the possibility that there

[1]The process of annotation presented strong evidences that the relative pronoun anaphor's antecedent is almost every time in the same sentence as the anaphor, and often immediately at its left side.

[2]**WEKA** (**W**aikato **E**nvironment for **K**nowledge **A**nalysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.

http://www.cs.waikato.ac.nz/ml/weka/ (access date: 04/08/2013)

could be not even one single valid candidate or that more than one candidate could be classified as the anaphor's antecedent. Therefore, we chose the Expectation-Maximization algorithm (EM) [10]. EM is a soft-clustering convergence method, which means, in our case, that it provides the probabilities of an instance (pair anaphor-candidate) to belong to a cluster. Running EM in two clusters (one represents that the candidate is the antecedent for that anaphor, the other represents that it is not), we are able to get the probabilities of each candidate to be the antecedent and therefore we are able to choose the best one.

The *corpus* feature extraction produced 97,167 instances[3]. Each anaphor has in average 10.37 candidates, increasing to 14.09 candidates on non-relative anaphors and decreasing to 4.56 candidates on relative anaphors.

[3]Our system runs on a Intel(R) Celeron(R) G550 2.6GHz PC. It took exactly 11 minutes to generate the training examples from the training documents for our whole *corpus*. The training time for the EM algorithm to generate the model from all the training examples was 73 seconds.

Type	Feature	Description	Possible values
R	distance	number of sentences between anaphor and candidate	{numeric}
R	same_sentence	verifies if the anaphor and candidate are in the same sentence	{true, false}
R	gender_agreement	verifies if anaphor and candidate agree in gender	{true, false}
R	number_agreement	verifies if anaphor and candidate agree in number	{true, false}
R	share_relation_with_verb	verifies if the anaphor and candidate have a relation with the same verb (*e.g.* subject, direct complement)	{true, false}
A	anaphor_gender	gender of the anaphor	{MASC(uline), FEM(inine), IND(efinite)}
A	anaphor_number	number of the anaphor	{SG (singular), PL(ural), IND(efinite)}
A	anaphor_type	type of the anaphor	{PRON(oun), ART(icle), NUM(eral)}
A	anaphor_pronoun_type	type of the pronoun	{PERS(onal), POSS(essive), DEM(onstrative), IND(efinite), NULL (anaphor is not a pronoun)}
A	is_anaphor_clitic	verifies if the anaphor is a clitic	{true, false}
A	is_anaphor_subject	verifies if the anaphor is a subject	{true, false}
A	is_anaphor_direct_ complement	verifies if the anaphor is a direct complement	{true, false}
A	is_anaphor_indirect_ complement	verifies if the anaphor is an indirect complement	{true, false}
C	candidate_gender	gender of the candidate	{MASC(uline), FEM(inine), IND(efinite)}
C	candidate_number	number of the candidate	{SG (singular), PL(ural), IND(efinite)}
C	is_candidate_a location	verifies if the candidate has a location feature	{true, false}
C	is_candidate_an organization	verifies if the candidate has an organization feature	{true, false}
C	is_candidate_conjoint	verifies if the candidate comprehends more than one entity	{true, false}
C	is_candidate_ demonstrative	verifies if the candidate is preceded by a demonstrative pronoun	{true, false}
C	is_candidate_human	verifies if the candidate has a human feature	{true, false}
C	is_candidate_a proper_noun	verifies if the candidate is a proper noun	{true, false}
C	is_candidate_indefinite	verifies if the candidate is indefinite	{true, false}
C	is_candidate_a location	verifies if the candidate has a location feature	{true, false}
C	is_candidate_NE	verifies if the candidate is a named entity	{true, false}
C	is_candidate_subject	verifies if the candidate is a subject	{true, false}
C	is_candidate_direct complement	verifies if the candidate is a direct complement	{true, false}
C	is_candidate_indirect complement	verifies if the candidate is an indirect complement	{true, false}
C	is_candidate_NP_or_PP	verifies if the candidate is a NP or PP	{true, false}
C	order_of_candidate	order of the candidate; 1 if is the closest candidate (regarding the anaphor), 2 if is the second closest, and so on	{numeric}
C	number_of_candidates	number of candidates for the same anaphor	{numeric}
T	is_antecedent	verifies if the candidate is the antecedent for the anaphor	{true, false}

Table 5.1: Features used in ARM2.0.

Chapter 6

Evaluation

THIS chapter presents the evaluation of the system's performance in the anaphora resolution task. As in any NLP task, evaluation is of critical importance to anaphora resolution and we realized that the attention paid to evaluation has been insufficient. If the discrepancy between systems makes the low attention to evaluation surprising, the fact that the results obtained in our human annotation by qualified annotators (see section 4.2) are inferior to some of the systems studied (see section 2.6) makes evaluation even much more relevant. Section 6.1 presents a detailed view of all the evaluation metrics that have been used and the different insights that they could provide, section 6.2 presents the results that have been obtained and finally, section 6.3 discusses and analyzes the system's performance.

6.1 Metrics

To evaluate the AR system (and to improve it), we performed the evaluation in three stages:

1. *Anaphor Identification*: to perform a more complete evaluation in the rule-based anaphor identification, this stage includes anaphor-by-type identification:
    - Personal pronouns;
        - Personal pronouns, except *se*[1];
        - *se* pronouns;
    - Relative pronouns;
        - *que* pronouns;
        - *onde* pronouns;
    - Possessive pronouns;
    - Demonstrative pronouns;
    - Indefinite pronouns;

[1] As it has been already explained, *se* is a particularly difficult case of Portuguese anaphora, since it can be used in a reflexive (and thus anaphoric) way (*e.g.* A *Marta* lavou-*se*. / *Marta* washed *herself*.) or not (*e.g.* Acredita-*se* que o Presidente morreu. / It is believed by everyone, in general, that the President died.).

- Cardinal numerals;
- Ordinal numerals;
- Articles;
- Cataphors;
- Total anaphors;

We made the distinction between *se* personal pronoun and the other peronal pronouns due to the reflexive, indefinite gender and number nature of the *se* pronoun, besides the difficulty that identifying expletive *se* encompasses. We also distinguish *que* and *onde* pronouns, as *que* represents the vast majority of relative pronouns and *onde* is a special case where the antecedent is usually a toponym (place names). Also, the annotation process stressed out sentences were *onde* was used in an expletive way (1.41).

(1.41) O Pedro colocou os óculos onde o Jorge colocou a mochila.
Pedro put the glasses where Jorge put his backpack.

2. *Candidate identification*: once again, to better assess the results and the main areas needing of improvement, the evaluation of candidate rule-based identification encompasses several informations:

- *Total anaphoras*;
  - *Antecedent identified*;
  - *Antecedent not found*;
    - *Antecedent out of reach*;
    - *Antecedent not among the candidates generated*;
  - *Total of non-conjoint antecedents*;
    - *Non-conjoint antecedents identified*;
    - *Non-conjoint antecedents not identified*;
  - *Total of conjoint antecedents*;
    - *Conjoint antecedents identified*;
    - *Conjoint antecedents not identified*;
  - *Total of non-relative antecedents*;
    - *Non-relative antecedents identified*;
    - *Non-relative antecedents not identified*;
  - *Total of relative antecedents*;
    - *Relative antecedents identified*;
    - *Relative antecedents not identified*;
- *Total cataphoras*;
  - *Antecedent identified*;
  - *Antecedent not found*;

  In the candidate identification section, it is of the utmost importance to assess when the antecedent is found among the candidates and when it is not (and in this case, if it is because the antecedent is out of reach or if the program fails to identify it). Also, we feel that it would be important to make the distinction between conjoint antecedents and not-conjoint antecedents since our program

does not identify conjoint antecedents when they are not coordinated. The separate assessment of relative antecedents results from the fact that relative anaphor candidates are always intrasentential, unlike the rest of the anaphors' candidates, which are limited by a two sentence window.

3. *Anaphora Evaluation*: includes the division-by-type of anaphor as done in the anaphor identification evaluation and also covers the previous two evaluation stages providing information of when one of the previous stages fails to identify the anaphor or the antecedent, respectively, or when the machine learning system chooses the incorrect candidate:

   - *Total anaphoras*;
     - *Antecedent correctly identified*;
       - *Antecedent correctly identified + antecedent between the candidates generated*;
         - *Anaphoras correctly resolved when the antecedent is the single candidate*;

The aforementioned factors are materialized in *precision* (Eq. 6.1), *recall* (Eq. 6.2) and *f-measure* (Eq. 6.3) which are the most common measures in evaluating NLP systems [26]. If we consider n the number of anaphoras in the text, t the number of anaphors that were resolved (irrespective of their correct/incorrect solution), s the number of anaphors which have been successfully resolved (true positives), k the number of one-antecedent candidate anaphors[2] and m the gender-number agreement solvable anaphors; then, we have the following measures:

$$Precision = \frac{s}{t} \tag{6.1}$$

$$Recall = \frac{s}{n} \tag{6.2}$$

$$f-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6.3}$$

Still, these measures cannot capture the *level of difficulty* involved in the anaphora resolved. For instance, if most of the anaphors present in the *corpus* used for evaluation have only one candidate, it is thus expected that the system will achieve higher evaluation marks in the previous measures. In this way, measures such as *critical success rate* (Eq. 6.4) evaluate the performance of the algorithm only in 'tougher' anaphoras.

$$Critical\ success\ rate = \frac{s-k-m}{n-k-m} \tag{6.4}$$

Also, it is important to evaluate the algorithm against baseline models to see how effective an approach is when compared with basic models. Baseline models in anaphora resolution usually settle on gender and number

[2] According to Mitkov [36, p. 177], the antecedents can be 'true' (manually corrected) or 'automatic' if we are evaluating the algorithm or the system, respectively. In our case we will evaluate the system, so the antecedents are automatically extracted, as we do not correct the cases were the antecedent is not among the candidates.

agreement and the choice of the nearest NP or the nearest subject. The nearest-NP-choice will serve as the baseline for comparison. Comparing with similar approaches is also helpful to place the approach against the field's state-of-the-art systems.

At last, it is extremely useful to break down the evaluation process by looking to different components. We believe this is the case of our approach, and that performing evaluation on the features individually or on different combinations may provide insights about their relevance for the task at hand. Such information was used to refine the relevance of each feature and to determine which one should play the tiebreak role.

6.2 Results

It is important to analyze AR along each stage, since each step's efficiency constitutes a ceiling to the performance of the next phase. In other words, if the anaphor is not successfully found or if the antecedent is not present in the candidate list, the anaphora will not be resolved however good the model may be. This section decomposes the AR process in all its stages, and evaluates each one of them separately, and, finally, shows the results for AR task as a whole. It also places ARM 2.0 against baseline models and analyzes ARM 2.0 in the field's state-of-the-art.

6.2.1 Anaphor Identification

Through the application of the manual rules described in section 5.2.1, ARM 2.0 is able to identify pronouns, numerals and articles as potential anaphors. Table 6.1 analyzes its success presenting the results decomposed by type of anaphor:

Type of anaphor		Found	Correct	Reference	Recall	Precision	F-measure
Personal pronouns	*se*	2,611	1,612	1,632	98.77%	61.74%	75.98%
	All except *se*	1,970	1,782	1,923	92.67%	90.46%	91.55%
	All	4,581	3,395	3,554	95.55%	74.11%	83.48%
Relative pronouns	*que*	3,721	3,038	3,593	84.55%	81.64%	83.07%
	onde	244	223	246	90.65%	91.39%	91.02%
	All	4,169	3,444	4,101	83.98%	82.61%	83.29%
Possessive pronouns		994	953	1,071	88.98%	95.88%	92.30%
Demonstrative pronouns		340	162	191	84.82%	47.65%	61.02%
Indefinite pronouns		404	122	222	54.95%	30.20%	38.98%
Cardinal Numerals		355	26	50	52.00%	7.32%	12.83%
Ordinal Numerals		41	11	25	44.00%	26.83%	33.33%
Articles		519	120	126	95.24%	23.12%	37.21%
TOTAL		11,403	8,233	9,339	88.16%	72.2%	79.39%

Table 6.1: Results for the evaluation of anaphor identification.

NOTE: "Found" means the number of anaphors identified by the program. "Correct" means the number of anaphors correctly identified by the program while "Reference" indicates the number of anaphoras annotated in the *corpus*.

Apart from numerals, articles, and demonstrative and indefinite pronouns, the figures in table 6.1 show that the system is fairly efficient in correctly identifying most types of anaphor (f-measure from 75.98% to 92.30%). It is also clear that the precision is typically lower than the recall, a fact that can be explained by the decision of not annotating co-referent anaphoras (the manual rules that were developed cannot discern co-referent anaphoras from identity-of-sense anaphoras). This is especially relevant in the case of the articles, since they are the usual anaphor in noun anaphora, a specific case of identity-of-sense anaphora (section 1.1.1). Cataphora events also help to lower the recall, as potential anaphors were identified that were, in reality, cataphors: from the 560 cataphoras present in the *corpus*, 264 cataphors were incorrectly considered as potential anaphors, which represents 2.32% of all the anaphors identified by ARM 2.0. Special cases of annotation or XIP errors are also among the reasons that prevented a higher precision and recall.

The *se* anaphor identification precision of 61.74% is also explained by the difficulty in discerning anaphoric and expletive *se*, even taking into account special dependencies provided by XIP (see examples 5.5 and 5.6). Numerals and indefinite pronouns lower the overall results, which is explained by the special difficulties that these types of anaphor raise. For instance, a considerable number of indefinite pronouns can be used both in an anaphoric or in an expletive way. Example (6.1) illustrates the impersonal expletive manner that impersonal pronouns *alguns* and *outros* are often used:

(6.1) *Alguns* acreditam que isto mudará, *outros* já não têm esperança.

Some believe that this will change, *others* do not even have hope.

Numerals identification is a particularly troublesome problem since they are often expletive as they can refer to time in ways that XIP currently cannot fully track it, *e.g.* João nasceu naquele ano (1995). / João was born in that year (1995)., or can be associated with other symbols for different meanings such as temperature, time, currency, *etc.*, *e.g.* 45” / 45 minutes; 25$; O resultado foi de 3-0. / The result was 3-0.; O handicap é 28. / The handicap is 28[3].

The anaphor identification attained a f-measure of 79.39%, which combines a solid 88.16% recall with a lower 72.2% precision, due to the aforementioned factors.

6.2.2 Candidate Identification

As it happens with the anaphors' identification, the compilation of a list of candidates is also made through manually crafted rules. Compiling a list of candidates poses the question of determining the search space from where the antecedent candidates will be retrieved, or, in other words, how far should we consider candidates be from their anaphor. The farther we go, the more probable it is for the antecedent to be present in the list; but it also means that many more candidates are considered, posing a greater and more complex challenge to the model, which would have to choose the correct candidate from a wider list. Therefore, we established a 2-sentence limit for all anaphors (the candidates considered have to be on the left-side of the anaphor in the same sentence or in the two previous sentences), unless the anaphor is a relative pronoun, in which case only the candidates on the left-side of the anaphor that belong to the same sentence will be considered. In the annotation process, it clearly stood out the

[3]XIP can track some symbols through dependencies like currency marks, but numerals can be associated with many other symbols.

Type of anaphor		Number of anaphors	Maximum	Average
Personal pronouns	"se"	1,628	1,072	20.11
	All except "se"	1,920	1,672	40.08
	All	3,548	1,672	30.92
Relative pronouns	"que"	3,591	160	3.21
	"onde"	246	25	3.62
	All	4,099	160	3.38
Possessive pronouns		1,071	995	25.15
Demonstrative pronouns		190	181	19.08
Indefinite pronouns		222	546	33.21
Cardinal Numerals		50	227	31.50
Ordinal Numerals		25	377	45.40
Articles		109	116	12.40
TOTAL		9,372	1,672	17.71

Table 6.2: Average and maximum distance in number of words between anaphor and antecedent of the anaphoras annotated in the *corpus*.

intrasentential nature of relative pronoun anaphora. In fact, in most cases, the antecedent of a relative pronoun is the closer candidate as it can be seen in table 6.2.

Table 6.3 sums up the results of candidate identification where the "Correct" column represents the times that an anaphor is correctly identified *and* the antecedent is present in the candidates list. In other words, it presents the ceiling for the machine learning model which is represented in the "Recall" column. Column "Drop-off" presents the ceiling drop-off from the anaphor identification (recall comparison between candidate identification and anaphor identification). Our program identified the antecedent in the candidates list 84.69% of the times, while in 9.18% of the cases it did not identify the antecedent as a candidate, that was due to the antecedent being out-of-range.

As expected, the large majority of the antecedents of relative pronouns are on the same sentence as the anaphor. Numerals and indefinite pronouns present a low ceiling following the already low results on its identification. Comparing the average distance between and anaphor and candidate in table 6.2, and the drop-off of recall from the anaphor identification stage and the candidate identification stage in table 6.3, we can see a relation between distance and candidate identification recall. Personal pronouns, which represent 37.44% of the *corpus* anaphors, report a significant -21.08% drop-off explained by the distance that in average separates anaphor and antecedent. On the other hand, relative pronoun anaphors are usually very close to the antecedent resulting in only 4.24% recall from anaphor identification phase to candidate identification phase. Overall, the candidates evaluation coupled with the anaphor evaluation reached a f-measure rate of 68.44%. The ceiling for the model is 76%, which means that for all the potential anaphoras identified, in 76% of those, the model is in conditions to correctly resolve the anaphora.

Table 6.4 provides interesting insights in the effect that applying gender and number filters may have in identi-

Type of anaphor		Found	Correct	Reference	Recall	Drop-off	Precision	F-measure
Personal pronouns	"se"	2,611	1,347	1,632	82.54%	-16.23%	51.59%	63.49%
	All except "se"	1,299	1,782	1,923	67.55%	-25.12%	65.94%	66.74%
	All	4,581	3,395	2,646	74.47%	-21.08%	57.76%	65.06%
Relative pronouns	"que"	3,721	2,882	3,593	80.21%	-4.34%	77.45%	78.81%
	"onde"	244	215	246	87.40%	-3.25%	88.11%	87.75%
	All	4,169	3,270	4,101	79.74%	-4.24%	78.44%	79.08%
Possessive pronouns		994	852	1,071	79.55%	-9.43%	85.71%	82.52%
Demonstrative pronouns		340	124	191	64.92%	-19.90%	36.47%	46.70%
Indefinite pronouns		404	80	222	36.04%	-18.91%	19.80%	25.56%
Cardinal Numerals		355	12	50	24.00%	-28.00%	3.38%	5.93%
Ordinal Numerals		41	11	25	44.00%	0%	26.83%	33.33%
Articles		519	120	103	81.75%	-13.49%	19.85%	31.94%
TOTAL		11,403	8,233	9,339	76.00%	-12.16%	62.25%	68.44%

Table 6.3: Results for the evaluation of anaphor identification and presence of antecedent in candidates list.

NOTE: "Found" means the number of anaphors identified by the program. "Correct" means the number of anaphors correctly identified by the program while "Reference" indicates the number of anaphoras annotated in the *corpus*.

fying the antecedent as a possible candidate.

Keep in mind that *se*, *que* and *onde* pronouns are not marked for number nor gender. Possessive pronouns also skip this filter since in Portuguese they agree with the noun they determine instead of their antecedent. The indefinite pronouns, the articles and the ordinal numbers also suffer a significant fall, explained by the fact that this type of anaphora often includes subset relationships making way for a number of cases where there is gender/number disagreement between the anaphor and its antecedent. Sentence (6.2) illustrates this situation with an ordinal numeral anaphor:

(6.2) Os *jogadores* foram apresentados aos adeptos. O *primeiro* foi Messi.

The *players* were presented to the fans. The *first* was Messi.

6.2.3 Selection of the best candidate

As described in chapter 5, we built an Expectation-Maximization (EM) model to select the best candidate from a candidate list. To perform the evaluation, we applied the EM model with and without gender and number filters. The model was also compared against a baseline against a baseline, which consists in just picking the candidate closer to the left of the anaphor. This baseline was also applied with and without filters for gender and number. Table 6.5 provides the detailed results for each of the four systems tested.

The EM model here developed outperforms the baseline consistently, except in the case of the relative pronouns anaphors. This can be easily explained by the small distance that most of the times separates the relative pronoun and its antecedent (table 6.2). In this type of anaphora, the baseline is a little better (approximately 2-3%). It is also clear that the application of gender and number filters slightly improves the results of the models here used (between 1.5 and 3%). This should be related with the fact that the EM model has a shorter number of candidates

Type of anaphor		Without filters	Variation with the application of gender & number filters
Personal pronouns	"se"	83.56%	0%
	All except "se"	72.90%	-3.93%
	All	77.94%	-2.06%
Relative pronouns	"que"	94.87%	0%
	"onde"	96.41%	0%
	All	94.95%	1.13%
Possessive pronouns		89.40%	0%
Demonstrative pronouns		76.54%	-2.47%
Indefinite pronouns		65.57%	-17.21%
Cardinal Numerals		46.15%	-3.84%
Ordinal Numerals		100.00%	-63.64%
Articles		85.83%	-35.00%
TOTAL		86.21%	-2.02%

Table 6.4: Effect of the application of gender and number filters.

NOTE: The total should be read: "For all the anaphors well identified, in 86.21% of the times the antecedent is present in the candidates' list. Introducing gender and number filters, this value drops -2.02%".

Type of anaphor		EM model			EM model w/ G&N filters			Baseline			Baseline w/ G&N filters		
		R	P	F	R	P	F	R	P	F	R	P	F
Personal pronouns	"se"	63.91%	39.95%	49.17%	**66.97%**	**41.86%**	**51.52%**	42.22%	26.39%	32.48%	42.59%	26.62%	32.76%
	All except "se"	41.5%	40.51%	41.00%	**44.31%**	**43.25%**	**43.77%**	11.39%	11.12%	11.25%	21.48%	20.96%	21.22%
	All	51.82%	40.19%	45.27%	**54.74%**	**42.46%**	**47.82%**	25.56%	19.82%	22.33%	31.18%	24.19%	27.24%
Relative pronouns	"que"	60.65%	58.56%	59.59%	60.40%	58.32%	59.34%	63.60%	61.41%	62.49%	**64.24%**	**62.03%**	**63.12%**
	"onde"	60.57%	61.07%	60.82%	63.01%	63.52%	63.26%	63.41%	63.93%	63.67%	**63.41%**	**63.93%**	**63.67%**
	All	60.03%	59.05%	59.54%	59.81%	58.84%	59.32%	61.86%	60.85%	61.35%	**62.57%**	**61.55%**	**62.06%**
Possessive pronouns		32.21%	34.71%	33.41%	**34.27%**	**36.92%**	**35.55%**	18.3%	19.72%	18.98%	18.49%	19.92%	19.18%
Demonstrative pronouns		49.21%	27.65%	35.41%	**54.45%**	**30.59%**	**39.17%**	29.32%	16.47%	21.09%	40.31%	22.65%	29.00%
Indefinite pronouns		15.77%	8.66%	11.18%	**18.02%**	**9.90%**	**12.78%**	5.41%	2.97%	3.83%	8.11%	4.46%	5.76%
Cardinal Numerals		20.0%	2.82%	4.94%	**22.00%**	**3.10%**	**5.43%**	4.00%	0.56%	0.98%	14.00%	1.97%	3.45%
Ordinal Numerals		16.00%	9.76%	12.12%	**16.00%**	**9.76%**	**12.12%**	12.00%	7.32%	9.09%	4.00%	2.44%	3.03%
Articles		46.83%	11.37%	18.30%	**53.17%**	**12.91%**	**20.78%**	17.46%	4.24%	6.82%	30.95%	7.51%	12.09%
TOTAL		51.93%	42.53%	46.76%	**53.44%**	**43.77%**	**48.12%**	40.00%	32.76%	36.02%	42.98%	35.20%	38.70%

Table 6.5: Precision, recall and f-measure results of EM model with and without gender and number filters against closer-candidate baseline with and without filters.

to choose from and also because the baseline model can avoid choosing incorrectly the closer candidate, if it does not agree in gender and/or in number with the anaphor. Below, we present some bar charts illustrative of the ceiling that each stage faces from the previous stage. Note that these charts only present recall numbers and do not consider the anaphors incorrectly identified.

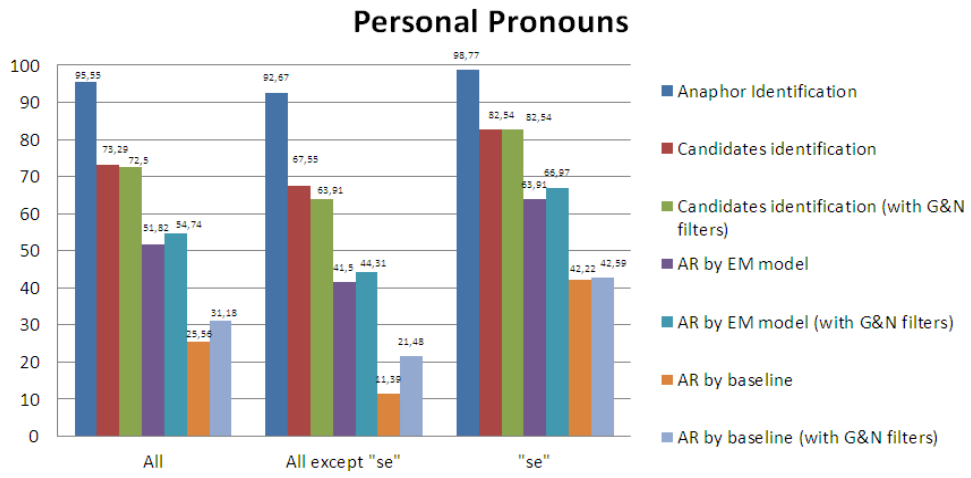

Figure 6.1: Performance of the different stages of AR of personal pronouns anaphors.

Figure 6.1 shows that EM model with gender and number filters outperforms the others systems, attaining a f-measure of 47.82% and outperforming the baseline with filters by a 20.08% margin. The improvement of results after application of gender and number filters (especially on the baseline model, that just picks the closer one as opposite to the EM model that relies on 30 features to make the choice) in other personal pronouns than *se* is explained by the shorter list of candidates, resulting from discarding the candidates that do not agree in gender and number with the anaphor. As for the *se* pronouns, they are not marked for gender nor number and, thus, they should not change. However, there is a little improvement with the application of gender and number features. Pronouns *se* AR produces better results than the other personal pronouns, a fact explained by the percentage of anaphoras in which the model is in condition to produce the right answer (anaphor well identified and antecedent present in the candidates list) being larger for the *se* pronouns.

About relative pronouns, often not marked for gender and number, figure 6.2 shows that the baseline (that always select the closer candidate) produces slightly better results than the machine learning model that was developed. These results confirm that relative pronouns often have their antecedent immediatelys at their left-side.

Figure 6.3 presents the results for the remaining types of pronouns addressed in this work: possessive, demonstative and indefinite. The results show that EM model with filters is the better approach for all these pronouns. As previously discussed in section 1.1.2, possessive anaphors in Portuguese do not agree in gender and number with the antecedent but with the noun they determine (which underwent zeroing). The pronominal use of demonstrative and indefinite is also due to the same zeroing. Some indefinites are gender/number marked. All *demonstrative* are number and gender marked. These grammatical features are reflected in the use of gender and number filters. However, even the results achieved by the best model in indefinite pronouns AR are very low (18.02%), due to the low ceiling in which they operate. In view of the low representativeness of this type of pronoun in the *corpus* (6.55% – chapter 4), we decided to leave the indefinite pronouns out of the scope of ARM2.0.

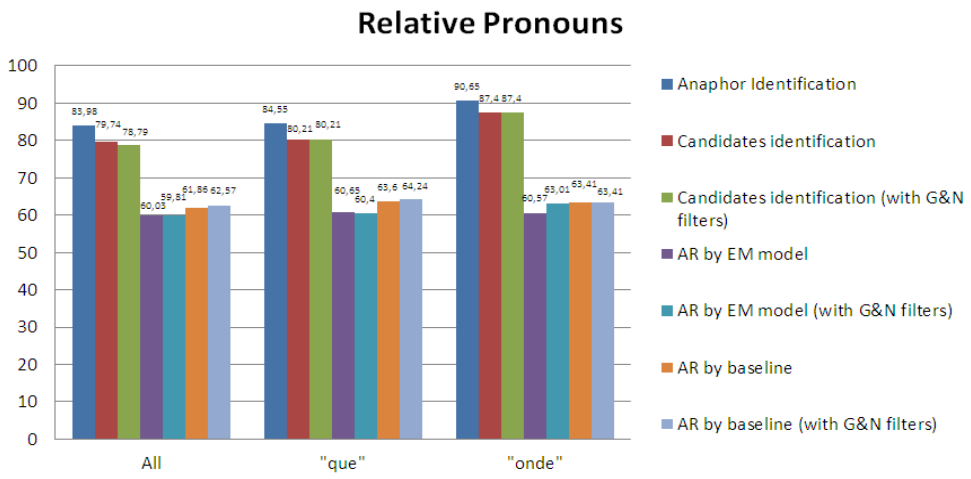

Figure 6.2: Performance of the different stages of AR of relative pronouns anaphors.

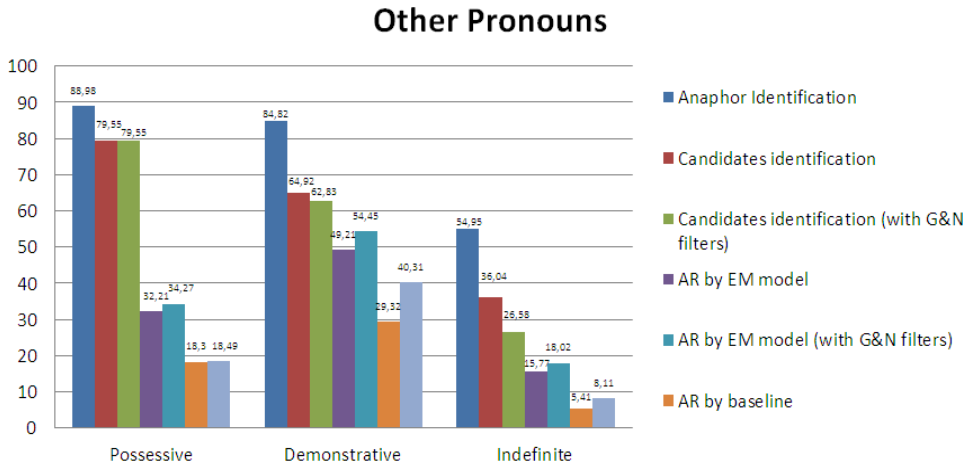

Figure 6.3: Performance of the different stages of AR of possessive, demonstrative and indefinite pronouns.

According to figure 6.4, EM with filters once again outperforms the other systems, and articles AR presents an interesting recall of 53.17%. As it happens with indefinite pronouns, the low recall and precision, plus the low number of numerals in the *corpus* (0.80%) lead us to exclude numerals from the scope of our system at this time. The problem with indefinite and demonstratives pronouns, articles and numerals is that the system marks as anaphors many instances that are not in fact anaphoric, and this excessive number of false-positives, while producing a good recall, hurts the precision significantly (table 6.1). This means that an effort must still be developed to improve the anaphor identification module.

6.2.4 Model efficiency

As discussed before, the anaphor and candidates identification efficiency mark a ceiling for the model's performance. In this section, we approach the model efficiency considering the ceiling set by the previous stages. Figure 6.5 displays the efficiency of the developed EM model with and without gender and number filters.

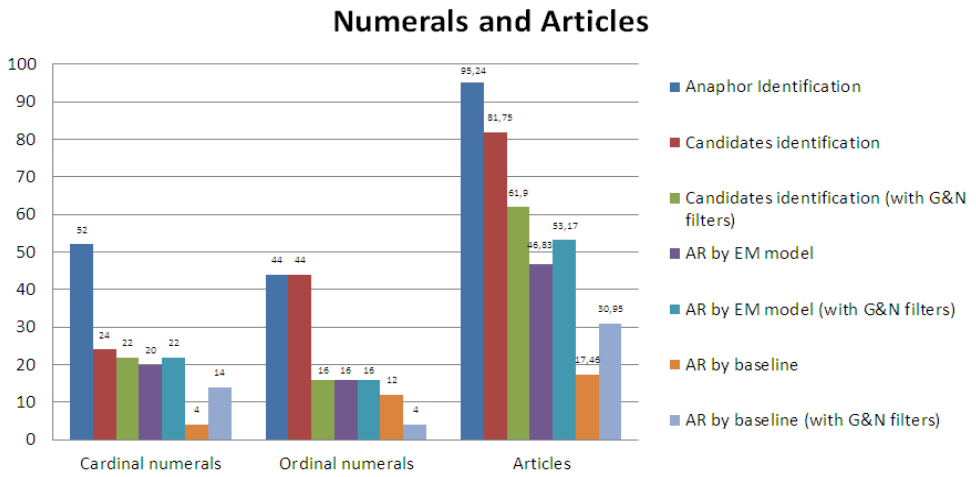

Figure 6.4: Performance of the different stages of AR of numerals and articles.

The results show that the introduction of gender and number filters improve the efficiency of the model, which is natural since that causes a decreasing of the number of candidates for the model to choose from. The model achieves good results with the relative pronouns, which can be explained by the lower number of candidates selected, due to the fact that the candidates have to be on the same sentence. The introduction of gender and number filters present no significant discrepancy since relative pronouns are usually not marked for gender and number. Possessive pronouns are the type of anaphor in which the model has more difficulties in selecting the correct antecedent. A usually considerable number of candidates (25.15 in average), coupled with the uselessness of gender and number filters, makes this type of anaphor a tough challenge for the model as its efficiency drops below 50%. Regarding indefinite pronouns, articles and numerals, it is not uncommon for the anaphor and the antecedent to disagree in number, thus providing a significant boost in the model efficiency with the application of the filters, since these non-agreement situations are counted as cases where the model has no conditions to resolve the anaphora successfully.

6.2.5 Variation in *corpus*

Most of the AR systems (see chapter 2) reported results for solving 3rd person pronouns using *corpora* that do not include dialogues (technical reports on MARS [36], MUC-5 [34], MUC-6 [5] and MUC-7 [47]). Our system also does not cover 1st and 2nd pronouns either, but our *corpus* is of a very diverse nature, presenting a large variety of textual genres, and including literary texts (novels), which are particularly rich in dialogues. Empirical evidence from these latter texts showed that dialogues often display an increased distance between the anaphors and their antecedents (table 6.6).

In this section, we present the results obtained with our system on the *golden standard corpus* but without the texts taken from the two novels, which are rich in dialogues.

Table 6.7 shows that there are little fluctuations in the anaphor identification in the *corpus* with and without novels. In general, the results of anaphor identification without novels are slightly worse (approximately -2%). However, when we compare the candidate identification in the *corpus* with and without novels (table 6.8), we conclude that the ceiling is clearly higher: +4.34% recall in *corpus* without novels (excluding indefinite pronouns

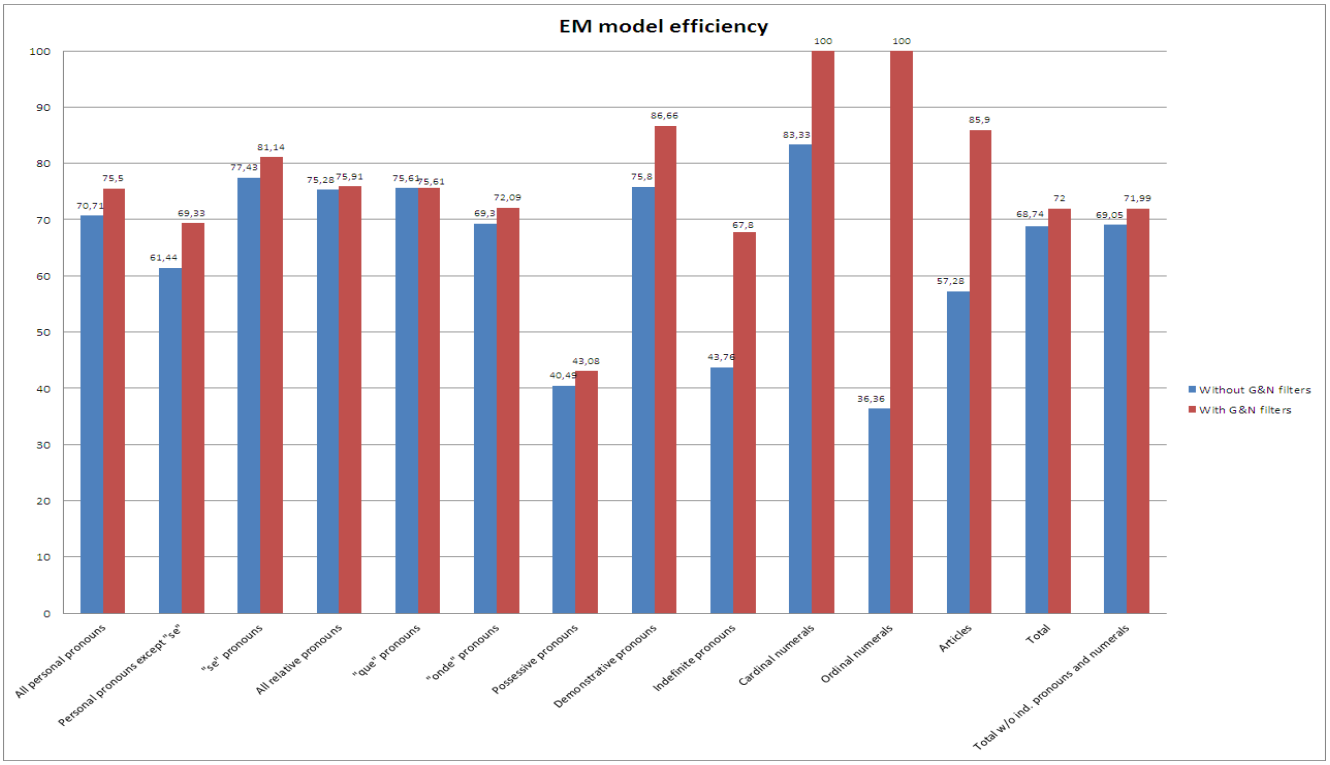

Figure 6.5: EM model efficiency with and without number and gender filters.

Type of anaphor		Entire *Corpus*			*Corpus* without novels			Avg. variation
		No. of anaphoras	Max.	Avg.	No. of anaphoras	Max.	Avg.	
Personal pronouns	*se*	1,628	1,072	20.11	1,039	368	11,57	-8.54
	All except *se*	1,920	1,672	40.08	704	299	25.01	-15.07
	All	3,548	1,672	30.92	1,743	368	17.00	-13.92
Relative pronouns	*que*	3,591	160	3.21	2,800	160	3.42	+0.21
	onde	246	25	3.62	190	25	3.76	+0.14
	All	4,099	160	3.38	3,226	160	3.61	+0.23
Possessive pronouns		1,071	995	25.15	915	995	21.59	-3.56
Demonstrative pronouns		190	181	19.08	139	159	13.71	-5.37
Indefinite pronouns		222	546	33.21	124	270	14.81	-18.40
Cardinal Numerals		50	227	31.50	33	90	18.39	-13.11
Ordinal Numerals		25	377	45.40	22	377	48.14	+2.74
Articles		109	116	12.40	84	44	10.25	-2.15
TOTAL		9,372	1,672	17.71	6,322	995	10.69	-7.02

Table 6.6: Comparison of distance between anaphors and antecedent in the entire *corpus* and in the *corpus* without novels.

NOTE: Total average variation (last column, bottom line) should be read “In average, anaphor and antecedent are 7.02 words closer in the *corpus* without novels than in the entire *corpus*.

Type of anaphor		Entire *corpus*			*Corpus* without novels		
		Recall	Precision	F-measure	Recall	Precision	F-measure
Personal pronouns	*se*	98.77%	61.74%	75.98%	-0.11%	-4.81%	-3.78%
	All except *se*	92.67%	90.46%	91.55%	-4.02%	-6.23%	-5.17%
	All	95.55%	74.11%	83.48%	-1.11%	-9.34%	-6.64%
Relative pronouns	*que*	84.55%	81.64%	83.07%	+3.21%	+3.15%	+3.18%
	onde	90.65%	91.39%	91.02%	-2.23%	+2.99%	+0.28%
	All	83.98%	82.61%	83.29%	+0.16%	+3.07%	+1.61%
Possessive pronouns		88.98%	95.88%	92.30%	-0.02%	-0.34%	-0.17%
Demonstrative pronouns		84.82%	47.65%	61.02%	+0.79%	+1.93%	+1.77%
Indefinite pronouns		54.95%	30.2%	38.98%	-12.21%	-11.64%	-11.23%
Cardinal Numerals		52.00%	7.32%	12.83%	-12.61%	+2.69%	-4.54%
Ordinal Numerals		44.00%	26.83%	33.33%	+6.00%	+0.67%	+2.15%
Articles		95.24%	23.12%	37.21%	-12.29%	-5.27%	-7.83%
TOTAL		88.16%	72.2%	79.39%	-1.62%	-2.32%	-2.07%
TOTAL w/o indefinite pronouns & numerals		89.29%	76.15%	82.20%	-1.47%	-1.73%	-1.63%

Table 6.7: Comparison of results of anaphors identification evaluation in *corpus* with and without novels.

and numerals). In other words, there are 4.34% more cases in which the model is in conditions to operate successfully. Personal and demonstrative pronouns represent the greater improvement, especially the personal pronouns, which present a 11.13% higher ceiling. Recall that personal pronouns represent 37.77% of the *corpus*. Note also that in the *corpus* without novels, the program successfully includes the antecedent in the candidate list in 89.90% of times, with the antecedent being out of range in only in 2.90% of the cases, and the program failing to detect the antecedent in the remaining 7.2% of the cases. This is natural given the shorter distance between anaphors and their antecedents in the *corpus* without novels (table 6.6). This stands out against the complete *corpus* where 9.18% of the times the antecedent was out of range (in the entire *corpus* there are, in average, further 7.02 words between the anaphor and its antecedent) which, in part, explains the higher ceiling in the *corpus* without novels.

Naturally, with a higher ceiling, EM model achieves better results in personal and demonstrative pronouns where the difference in the ceiling was greater. For the same reason, we verify that relative pronouns AR in the *corpus* without novels is slightly worse. However, the best improvement is related to the possessive pronouns, which registered an almost 30% boost. If the variations in the other type of anaphors are explained by a higher/lower ceiling, that is not the case for the possessive pronouns. This surprising result imposed a better look into this matter. Applying the model trained in the *corpus* without novels and applying it to the entire *corpus*, we achieved a recall of 60.78%, a precision of 65.49% and an f-measure of 63.05%, which hints that the presence of dialogues, in some way, deteriorates the generated model, regarding possessive pronouns.

Overall, the results from the *corpus* without novels are a solid 4.79% better in terms of f-measure and 5.10% better if we exclude indefinite pronouns and numerals.

Type of anaphor		Entire *corpus*			*Corpus* without novels		
		Recall	Precision	F-measure	Recall	Precision	F-measure
Personal pronouns	"se"	82.54%	51.59%	63.49%	+6.70%	-0.09%	+1.82%
	All except *se*	67.55%	65.94%	66.74%	+9.75%	+7.51%	+8.59%
	All	73.29%	56.84%	64.03%	+11.13%	+1.05%	+4.65%
Relative pronouns	*que*	80.21%	77.45%	78.81%	+2.61%	+2.56%	+2.58%
	onde	87.40%	88.11%	87.75%	-2.14%	+2.90%	+0.29%
	All	79.74%	78.44%	79.08%	-0.22%	+2.54%	+1.16%
Possessive pronouns		79.55%	85.71%	82.52%	+1.11%	+0.91%	+1.01%
Demonstrative pronouns		64.92%	36.47%	46.70%	+9.90%	+6.86%	+8.18%
Indefinite pronouns		36.04%	19.8%	25.56%	+0.25%	-1.36%	-2.00%
Cardinal Numerals		24.00%	3.38%	5.93%	+0.24%	-0.43%	-0.83%
Ordinal Numerals		44.00%	26.83%	33.33%	+6.00%	+0.67%	+2.15%
Articles		81.75%	19.85%	31.94%	-6.75%	-3.71%	-5.38%
TOTAL		75.55%	61.88%	68.04%	+4.08%	+2.42%	+3.11%
TOTAL w/o indefinite pronouns & numerals		76.90%	65.58%	70.79%	+4.34%	+5.12%	+4.81%

Table 6.8: Comparison of results of candidates identification evaluation in *corpus* with and without novels.

Type of anaphor		Best results – complete *corpus*			Best results – *corpus* without books			Variation		
		R	P	F	R	P	F	R	P	F
Personal pronouns	*se*	66.97%	41.86%	51.52%	74.06%	42.74%	54.2%	+7.09%	+0.88%	+2.68%
	All except *se*	44.31%	43.25%	43.77%	52.06%	49.46%	50.73%	+7.75%	+6.21%	+7.48%
	All	54.74%	42.46%	47.82%	65.18%	44.7%	53.03%	+10.44%	+2.24%	+4.79%
Relative pronouns	*que*	64.24%	62.03%	63.12%	63.68%	61.52%	62.58%	-1.56%	+0.51%	-0.54%
	onde	64.63%	65.16%	64.89%	63.16%	67.42%	65.22%	-1.47%	+2.26%	+0.33%
	All	62.57%	61.55%	62.06%	60.01%	61.10%	60.55%	-2.56%	-0.45%	-1.51%
Possessive pronouns		34.27%	36.92%	35.55%	61.97%	66.55%	64.18%	**+27.7%**	**+29.63%**	**+28.63%**
Demonstrative pronouns		54.45%	30.59%	39.17%	61.15%	35.42%	44.86%	+6.70%	+4.83%	+5.71%
Indefinite pronouns		18.02%	9.90%	12.78%	27.42%	13.18%	17.8%	+9.40%	+3.28%	+5.02%
Cardinal Numerals		22.00%	3.10%	5.43%	21.21%	2.49%	4.46%	-0.79%	-0.61%	-0.97%
Ordinal Numerals		16.00%	9.76%	12.12%	31.82%	17.5%	22.58%	+15.82%	+7.74%	+10.46%
Articles		53.17%	12.91%	20.78%	57.95%	12.47%	20.52%	+4.78%	-0.44%	-0.26%
TOTAL		53.44%	43.77%	48.12%	59.22%	47.82%	52.91%	+5.78%	+4.05%	+4.79%
TOTAL w/o indefinite pronouns & numerals		54.59%	46.55%	50.25%	60.33%	51.13%	55.35%	+5.74%	+4.58%	+5.10%

Table 6.9: Precision, recall and f-measure variation on models when the novels are removed from the *corpus*.

Type of anaphor		Anaphor identification			Candidates identification			Anaphora resolution		
		R	P	F	R	P	F	R	P	F
Personal pronouns	*se*	98.77%	61.74%	75.98%	82.54%	51.59%	63.49%	66.97%	41.86%	51.52%
	All exc. *se*	92.67%	90.46%	91.55%	63.91%	62.39%	63.14%	44.31%	43.25%	43.77%
	All	95.55%	74.11%	83.48%	72.5%	56.24%	63.34%	54.74%	42.46%	47.82%
Relative pronouns	*que*	84.55%	81.64%	83.07%	80.21%	77.45%	78.81%	64.24%	62.03%	63.12%
	onde	90.65%	91.39%	91.02%	87.40%	88.11%	87.75%	63.41%	63.93%	63.67%
	All	83.98%	82.61%	83.29%	78.79%	77.5%	78.14%	62.57%	61.55%	62.06%
Possessive pronouns		88.98%	95.88%	92.3%	79.55%	85.71%	82.52%	60.78%	65.49%	63.05%
Demonstrative pronouns		84.82%	47.65%	61.02%	62.83%	35.29%	45.20%	54.45%	30.59%	39.17%
Articles		95.24%	23.12%	37.21%	61.09%	15.03%	24.19%	53.17%	12.91%	20.78%
TOTAL		89.29%	76.15%	82.20%	75.83%	64.67%	69.81%	58.98%	50.30%	54.30%

Table 6.10: Precision, recall and f-measure of all AR stages of the final ARM 2.0. model in the entire *corpus*.

6.2.6 Building the best model

The results presented above led us to take some decisions in order to build the final and best model:

- Gender and number filters are to be used: The filters' application proved to provide a consistent overall improvement;
- Exclusion of indefinite pronouns and numerals from AR module: The results were not good enough neither on recall nor on precision. In the very first stage of anaphor identification the numbers were already too low, thus conditioning the subsequent phases and, ultimately, the overall AR module results;
- Relative pronouns should resolved by closer-candidate criteria: Closer-candidate baseline consistently outperformed EM model for relative pronouns, even if only by a slim margin;
- Use the model trained in the *corpus* without novels, since the results thus achieved are better, particularly regarding possessive pronouns, which registered a very significant boost. This decision is connected with the abundant presence of dialogues in the novels in the *corpus*. Naturally, it may have to be adapted to the textual genre of the text to be processed.

Once the final ARM 2.0. structure has been defined, we now take a look at the final evaluation results. Table 6.10 displays the results achieved from stage to stage. The exclusion of indefinite pronouns and numerals, associated with the boost on possessive pronouns resolution, propel ARM 2.0. to overall results above 50% on precision and recall (recall results bordering the 60% mark).

On the other hand, figure 6.6 allows us to compare the ceilings that are being carried from anaphor identification to candidates identification and, in turn, to anaphora resolution. Relative and *se* pronouns stand out as the best results due to the usually greater proximity between anaphor and antecedent in this cases. Possessive pronouns also achieve a very interesting 60.78% recall. Nonetheless, it is clear that the ceiling from the previous stage of candidate identification greatly influences AR results. On another hand, other personal pronouns than *se* have the lower results of 44.31% recall, explained by the lower ceiling inherited, as well as the greater number of candidates associated whitin the two previous sentences search space.

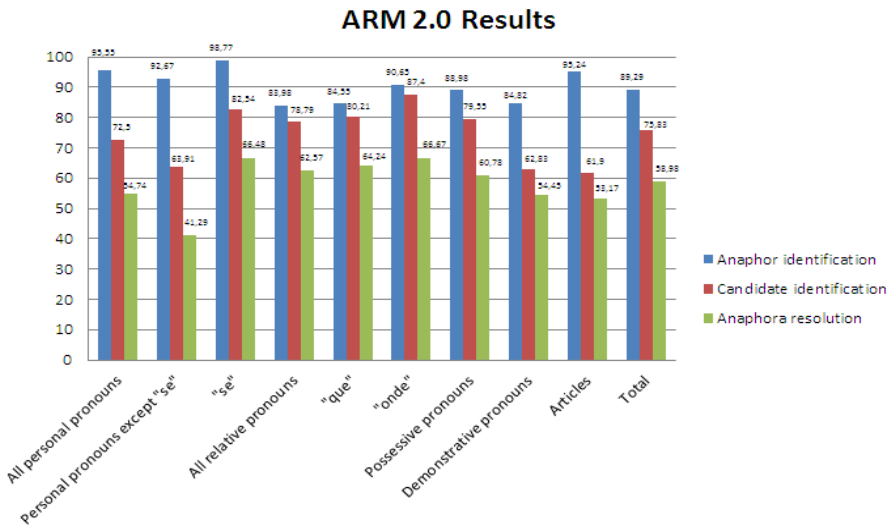

Figure 6.6: Global performance of the different stages of ARM2.0 for each type of anaphor.

Finally, we take a look at the model efficiency, that is, how does the system performs when it only considers the cases in which the anaphor is correctly identified *and* the antecedent is among the candidates. In other words, how does the model perform when it has all the conditions necessary to resolve the anaphora.

In the same token, we also assess *critical success rate* (Eq. 6.4), that is, the efficiency of the model when it discards all the anaphoras that can be resolved in a trivial way, namely, when there is only a single candidate antecedent for the anaphor or all other candidates but one are excluded on the basis of gender and number agreement.

Figure 6.7 compares the model efficiency and the critical success rate, and their breakdown by anaphor type. This figure shows the efficiency of the model in resolving "tougher" anaphoras and the impact of "trivial" anaphoras in the evaluation. Relative pronouns are the type of anaphor that take most advantage of these cases (682 cases, 26.58%), since the one sentence window applied in these type of anaphors promote single candidate anaphoras, hence registering the major drop-off when discarding the gender-number agreement and single candidate solvable anaphoras. A little portion of personal pronouns, excluding *se*, are also resolved under these terms (76 cases, 9.57%), which is natural if we consider that only the accusative and nominative 3rd person are marked for gender and number. On the other hand, *se* pronouns rarely are resolved on the basis of a single candidate or gender and number agreement. This can be explained by the fact that this type of anaphor compiles a list of candidates, whose range reports a two sentence window, minimizing the single candidate scenario. Considering that *se* pronouns are also not marked for gender and number, it is natural the little impact of critical success rate in this type of anaphor (10 cases, 0.38%). The remaining types of anaphors are only very rarely resolved under these conditions, the possessive pronouns are not even submitted to gender and number filters (each of the remaining types of anaphora registered under 10 gender-number or single-candidate solvable anaphoras).

The model efficiency is relatively good ranging between 64.61% in personal pronouns (excluding *se*) and 86.66% in demonstrative pronouns resolution. We consider an overall efficiency of 77.78% a very solid value. Even when considering only "tougher" anaphoras, ARM 2.0 AR model attains a 72.68%, which continues to be a reliable rating.

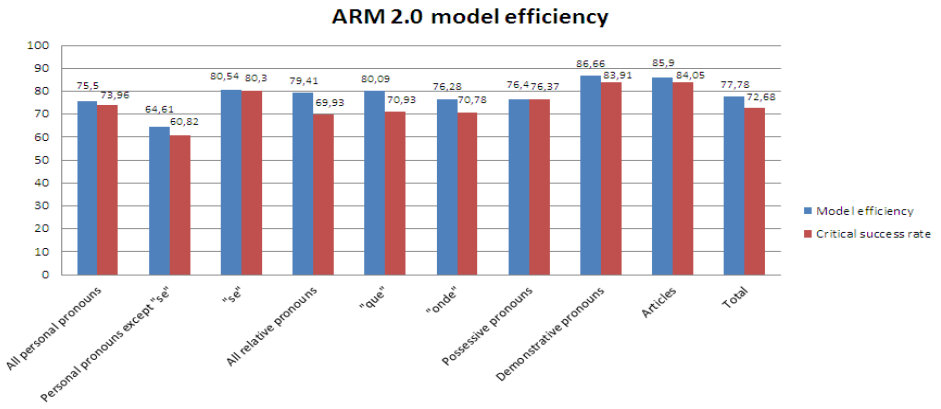

Figure 6.7: Performance of the ARM2.0 AR model for each type of anaphor.

6.3 Discussion

So far, we have described ARM 2.0 and evaluated it in detail. Now, we will compare and evaluate our system and the different systems studied and discuss the significance of the results achieved to determine where ARM 2.0 stands against AR state-of-art.

Table 6.11 resumes the main properties of each system studied, while table 6.12 compares the evaluation subject and the results scored for each system.

Before assessing the results of other systems, recall that after the second round of human annotation, the best annotator reached a 88.3% of accuracy mark (section 4.2). Thus, we deem that is safe to assume that that 88.3% mark is our ceiling.

As it was said in chapter 2, we cannot compare straightforwardly all these systems, since they work on different types of anaphora, the *corpus* being used are different, not to mention the impact that manually-corrected input have in results.

Thus, looking solely to the results, it stands out that our system ranks only above ARM 1.0 [41] and Cardie and Wagstaff's approach [5]. However, it is important to notice that top-performing systems like MARS [36], Hobbs's naïve approach [24] and collocation pattern-based approach [9] operate in very small or technical *corpora*, with a limited number of anaphoras. This is substantially different from our multiple-genre, 9,268 anaphoras *corpus*. Another property that we cannot overlook is the type of anaphora treated. Most systems focused only on 3rd personal pronouns, while our system has a more extensive scope as it includes relative, possessive and demonstrative pronouns, besides articles.

Of all the systems studied, ARM 1.0 is, naturally, the most similar. Unfortunately, we could not evaluate ARM 1.0 with the *golden standard corpus*. Nonetheless, ARM 1.0 was aimed only at 3rd personal and possessive anaphora[4]. In these type of pronouns, ARM 2.0 attained a 48.5% f-measure, which represents a 15% improvement towards ARM 1.0. It is also interesting to compare RAPM [6] results since they are from the other system studied that dealt with Portuguese (even though this was the Brazilian variety) system studied and its *corpus* included over 1,050 anaphoras. They only focus on 3rd personal pronouns and achieved a good 67.01% success rate. Still, the

[4]However, ARM 1.0. did not considered *se* pronouns.

System	Approach Type	Method	Type of anaphora
Hobbs's [24]	Syntax-based	Parse-tree analysis	Pronouns
MARS [36]	Syntax-based	Antecedent factors	3rd person personal pronouns
RAPM [6]	Syntax-based	Antecedent factors	3rd person personal pronouns
ARM 1.0 [41]	Syntax-based	Antecedent factors	Demonstrative, relative, possessive and 3rd person personal pronouns
ARM 2.0	Syntax-based and Machine learning	EM algorithm	3rd person personal and possessive pronouns
Collocation pattern-based approach [9]	Statistical analysis	Co-occurrence	3rd person pronouns
RESOLVE [34]	Machine learning	C4.5 algorithm	Co-referent noun phrases
Cardie and Wagstaff [5]	Machine learning	Clustering algorithm	Noun phrases
Soon [52]	Machine learning	C5 algorithm	Noun phrases
Rahman and Ng [47]	Machine learning	Cluster ranking	Noun phrases

Table 6.11: Systems' features overview.

corpus is different from the one we used here and that may strongly impact the results, as we have seen, when we compared the results with and without literary texts (novels). In fact, it stands out that the baseline with the RAPM *corpus*, consisting in choosing the closer candidate as an antecedent, reached only a 55.49% success rate. This clearly contrasts with the 40.60% achieved with our *corpus* when using the equivalent approach (37.86% with gender-number filters).

Considering machine learning approaches, RESOLVE [34] seems to be highly domain-independent when we take into account that it scores only an f-measure of 47% on the MUC-6 data set. Rahman and Ng's system [47] stands apart with 76.0%, even with pre-processed errors manually removed, which does not happens with Soon's approach [52] that presents a +8.3% f-measure improvement towards ARM 2.0.

In face of the results reported by most of the aforementioned systems, it could be posited that ARM 2.0 still has a significant room for improvement. However, it is relevant to notice that these system's achievement of such high rate has only been possible because hard man-work and human expertise was provided to feed the system with correct input data. This contrasts very deeply with our own strategy, which aims at getting raw texts and resolving its anaphors in an entirely automatically way, something that is much closer to a real scenario of a NLP system in use.

Nonetheless, ARM 2.0 represents a step forward as it improved ARM 1.0 not only in performance but in resolving a more extensive scope of anaphoras and evaluating them in an extensive and unprecedentedly large Portuguese annotated *corpus*.

System	Evaluation Target	Manually Corrected Input	Top Results
Hobbs [24]	100 pronouns from a history book; 100 pronouns from a literary book; 100 pronouns from newspaper	✔	91.7% success rate
MARS [36]	2,263 pronouns from technical manuals	✔	92.27% success rate
RAPM [6]	1,050 pronouns from law, literary and newswire *corpora*	✗	67.01% success rate
ARM 1.0 [41]	334 pronouns from 8 forum messages texts, 1 legal text, 11 texts from news articles	✗	30% recall 38% precision 33.5% f-measure
ARM 2.0	*Golden standard corpus*[1](9,268 anaphoras)	✗	58.98% recall 50.30% precision 54.30% f-measure
Collocation pattern-based approach [9]	Hansard corpora (59 examples)	✗	87% success rate
RESOLVE [34]	MUC-5 English joint venture corpora	✔	86.5% f-measure
Cardie and Wagstaff [5]	MUC-6 co-reference resolution corpora	✔	53.6% f-measure
Soon [52]	MUC-6 and MUC-7 corpora	✗	62.6% f-measure
Rahman and Ng [47]	ACE data set	✔	76.0% f-measure

Table 6.12: Systems' evaluation overview.

[1] see section 4.

Chapter 7

Conclusions

This final chapter presents a brief summary of the main aspects of this study, along with some final remarks. We conclude with future work that we consider could be a good start for improving the current system.

7.1 Synopsis

Anaphora resolution is arguably one of the most intriguing and difficult tasks in Natural Language Processing. This dissertation aimed to improve the understanding of AR challenges in Portuguese and to improve the performance of a NLP system developed at L^2F/INESC-ID Lisboa, by implementing the AR module, responsible for identifying anaphors, produce a candidate list and choose the most probable one for antecedent.

Chapter 2 carried out a comparison between 8 of the most influential systems in the definition of the task, and it also described ARM 1.0, the existing system that we tried to improve. The chapter presented those multiple approaches, from rule-based and statistical approaches to machine learning-based systems; and discussed the advantages and disadvantages that each one featured. Each system was described in detail and the results and their significance were compared.

In Chapter 3, we explored an array of annotation platforms in order to choose the one that could help us maximize the performance of such an important task as annotation. After comparing the characteristics of each one, we chose Glozz, since it stood apart as the more complete, friendlier-interface platform, persuaded that it could better help annotators in a time-consuming and laborious task such as annotating a *corpus*.

Chapter 4 outlines the critical importance of an annotated *corpus* in an automatic NLP task and it describes the *corpus* features, the annotators qualification, the annotation process itself and the way in which it was evaluated. The need of annotation directives to ensure the consistency of the whole process was highlighted. The results showed that AR is a difficult task even for humans, let alone for computers. A ceiling of 88.3% was established as the AR goal for Portuguese, genre-unbounded texts.

In Chapter 5, we present and describe in detail the architecture of ARM 2.0 with all the rules, features and dependencies that the three stages of AR (anaphor identification, candidate list compilation, selection of the best candidate) demanded. We proposed an hybrid approach, in which the anaphor and candidates were retrieved based on rules; the candidate selection is underpinned by an Expectation-Maximization model. The XIP parser was instrumental in this process, since it was based on its output that we were able to compute a set of 30 features for

the machine learning models.

Finally, Chapter 6 presented the evaluation of ARM 2.0. First, we defined the evaluation measures and the types of anaphors that would be covered by the AR task and how the evaluation would be conducted and organized. A second section presented the evaluation results in detail, at each stage of the AR process, enhancing the fact that each step imposed a ceiling on the next one.

The evaluation itself was conducted on each phase of the process, with the EM model developed being compared against a closer-candidate baseline, with and without G.N. filters. The removal of two novels from the *corpus*, was also tested, as an attempt to remove the dialogues' impact in the system's performance. This showed clearly the effect of dialogues' structure in the results, especially in the case of possessive pronouns, where the model trained boosted the accuracy for this type of anaphors in almost 30%. Finally, we discussed the model efficiency alone, not taking into account anaphoras impossible to be resolved, due to the inefficiency of previous stages. The model, when given the necessary conditions to resolve anaphoras, resolves them successfully 77.8% of the times, which can be considered a solid rate.

7.2 Future Work

In the following items, we present different topics that should be focus of attention in a future work in AR so ARM of the L^2F/XIP system STRING can still be further improved:

- The type of anaphor is a very interesting way of structuring the anaphora resolution task. Considering human anaphora resolution, it can be argued that anaphora resolution strategies should be adjusted according to the anaphora type. Therefore, a *corpus* annotated with a wide range of anaphoric relations such as co-reference, metonymy, subset relations, superset relations, inalienable possession (body parts), family relations, zero anaphora and identity of sense, could help to better assess the type of anaphora and, ultimately, to better resolve it;

- Anaphora is a discursive device that enhances the cohesion of the text, making a sentence interpretation dependent from other sentences. At the moment, our system resolves anaphoras simultaneously, which does not allow to take advantage from information acquired in a sequential process. In example (7.1), after resolving the first anaphor *ela* to its antecedent *Sara*, knowing that *Ela* and *dela* are not co-referential (since *dela* is not reflexive nor followed by a focus determiner), *Sara* is excluded as a candidate for the second anaphor *dela*, leaving *Carolina* as the only potential candidate for antecedent;

  (7.1) A *Sara* gosta da *Carolina*. *Ela* sempre gostou *dela*.
  Sara likes *Carolina*. *She* always liked *her*.

- In ARM 2.0, anaphor identification was ensured by manually crafted rules. Although the results achieved can be considered quite reasonable, it would be very interesting to apply machine learning methods to the very first stage of anaphora resolution, particularly in the case of indefinite pronouns and numerals, which hindered, in a significant way, the results of the anaphor identification stage;

- As presented in the evaluation (section 6.2.5), the analysis took into consideration the impact of literary texts (two novels) that integrated the *corpus*, in the overall performance of the AR system. In particular,

the presence of dialogues was deemed to influence the AR task, so experiments were carried out, removing these two novels entirely, as an attempt to remove the dialogues from the scope of the task. Further studies are required to confirm this hypothesis, and we suggest that a specific model be made to spot dialogues in text and, eventually, help the AR system to adapt its strategy to the nature of the text;

- Other machine learning algorithms, besides EM, should be tried and compared with the current approach (such as COBWEB [15] or Naïve Bayes classifier [63]). It is also desirable to use different knowledge sources, besides grammatical and textual features, namely at semantic and pragmatic level, to improve the AR task;

- Dagan and Itai's collocation pattern-based approach showed promise despite the small examples the authors tested. It would be very interesting to retrieve statistical data like the aforementioned collocation patterns and evaluate the impact of this information in the anaphora resolution task.

The results are deemed as satisfactory, as they met the goals of choosing an annotation framework, building an annotated Portuguese *corpus* and developing an hybrid approach that extended the scope and improved the performance of the previous AR module. The gap between ARM 2.0 results and the ones reported by some of the systems studied, even taking into account their different scope, the different *corpora* they used, and the fact that their input was previously corrected, shows that there is still room for improvement. The development and analysis of an unprecedently large Portuguese annotated *corpus* provides the conditions to continue to improve the Anaphora Resolution Module.

Bibliography

[1] Ait-Mokhtar, S., Chanod, J.-P., and Roux, C. (2002). Robustness beyond Shallowness: Incremental Dependency Parsing. *Natural Language Processing*, 8(2/3):121–144.

[2] Broscheit, S., Ponzetto, S. P., Versley, Y., and Poesio, M. (2010). Extending BART to provide a Coreference Resolution System for German. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC'10, pages 164–147, Valletta, Malta. European Language Resources Association (ELRA).

[3] Cabrita, V. (2012). Events, Anaphora and Computer Assisted Language Learning. Master's thesis, Instituto Superior Técnico, Lisboa.

[4] Carapinha, F. (2013). Extracção Automática de Conteúdos Documentais. Master's thesis, Instituto Superior Técnico, Lisboa.

[5] Cardie, C. and Wagstaff, K. (1999). Noun Phrase Coreference as Clustering. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, EMNLP/VLC '99, pages 82–89, College Park, Maryland, USA.

[6] Chaves, A. R. and Rino, L. H. (2008). The Mitkov Algorithm for Anaphora Resolution in Portuguese. In *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language*, PROPOR '08, pages 51–60, Aveiro, Portugal. Springer-Verlag.

[7] Ciccarese, P., Ocana, M., and Clark, T. (2012). Open Semantic Annotation of Scientific Publications using Domeo. *Journal of Biomedical Semantics*, 3(Suppl 1):1–14.

[8] Culotta, A., Wick, M., Hall, R., and Mccallum, A. (2007). First-order Probabilistic Models for Coreference Resolution. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, HLT-NAACL '07, pages 81–88, New York, USA.

[9] Dagan, I. and Itai, A. (1991). A Statistical Filter for Resolving Pronoun References. In Feldman, Y. A. and Bruckstein, A., editors, *Artificial Intelligence and Computer Vision*, pages 125–135. Elsevier Science Publishers B.V.

[10] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

[11] Denis, P. and Baldridge, J. (2007). A Ranking Approach to Pronoun Resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1588–1593, Hyderabad, India. Morgan Kaufmann Publishers Inc.

[12] Diniz, C. (2010). Um Conversor baseado em Regras de Transformação Declarativas. Master's thesis, Instituto Superior Técnico, Lisboa.

[13] Diniz, C. and Mamede, N. (2011). *LexMan – Lexical Morphological Analyzer*. Manual, INESC-ID, Lisboa.

[14] do Nascimento, M., Veloso, R., Marrafa, P., Pereira, L., Ribeiro, R., and Wittmann, L. (1998). LE-PAROLE: do Corpus à Modelização da Informação Lexical num Sistema-multifunção. *Actas do XIII Encontro Nacional da Associação Portuguesa de Linguística*, 2:115–134.

[15] Fisher, D. (1987). Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning*, 2(2):139–172.

[16] Fleiss, J. (1971). Measuring Nominal Scale Agreement among many Raters. *Psychological Bulletin*, 76(5):378–382.

[17] Ge, N. and Hale, J. (1998). A Statistical Approach to Anaphora Resolution. In Charniak, E., editor, *Proceedings of the 6th Workshop on Very Large Corpora*, COLING '98, pages 161–170, Montreal, Québec, Canada. Association for Computational Linguistics.

[18] Gobbel, G., Reeves, R., Speroff, T., Brown, S., and Matheny, M. (2011). Automated Annotation of Electronic Health Records using Computer-adaptive Learning Tools. volume 1, Washington D.C., USA.

[19] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: a Brief History. In *Proceedings of the 16th Conference on Computational Linguistics*, COLING '96, pages 466–471, Copenhagen, Denmark. Association for Computational Linguistics.

[20] Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225.

[21] Hagège, C., Baptista, J., and Mamede, N. (2008). Identificação, Classificação e Normalização de Expressões Temporais em Português: a Experiência do Segundo HAREM e o Futuro. In Mota, C. and Santos, D., editors, *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: o Segundo HAREM*, chapter 2, pages 33–54. Linguateca. http://www.inesc-id.pt/ficheiros/publicacoes/5758.pdf/.

[22] Hagège, C., Baptista, J., and Mamede, N. (2009). Portuguese Temporal Expressions Recognition: from TE Characterization to an Effective TER Module Implementation. In *7th Brazilian Symposium in Information and Human Language Technology*, STIL '09, pages 1–5, São Carlos, Brazil. Sociedade Brasileira de Computação.

[23] Hendrickx, I., Devi, S., Branco, A., and Mitkov, R., editors (2011). *8th Discourse Anaphora and Anaphor Resolution Colloquium on Anaphora Processing and Applications*, Revised Selected Papers DAARC '11, Faro, Portugal. Springer-Verlag.

[24] Hobbs, J. R. (1978). Resolving Pronoun References. *Lingua*, 44:311–338.

[25] Joachims, T. (1999). Advances in Kernel Methods. chapter Making Large-Scale Support Vector Machine Learning Practical, pages 169–184. MIT Press.

[26] Jurafsky, D. and Martin, J. (2009). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in Artificial Intelligence. Pearson Prentice Hall.

[27] Kennedy, C. and Boguraev, B. (1996). Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In *Proceedings of the 16th International Conference on Computational Linguistics*, COLING '96, pages 113–118, Copenhagen, Denmark. John Wiley and Sons, Ltd.

[28] Knublauch, H., Fergerson, R., Noy, N., and Musen, M. (2004). The Protégé OWL Plugin: an Open Development Environment for Semantic Web Applications. In *The Semantic Web – ISWC 2004*, Lecture Notes in Computer Science, pages 229–243. Springer-Verlag.

[29] Lappin, S. and Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

[30] Loureiro, J. (2007). Reconhecimento de Entidades Mencionadas (Obra, Valor, Relações de Parentesco e Tempo) e Normalização de Expressões Temporais. Master's thesis, Instituto Superior Técnico, Lisboa.

[31] Mamede, N., Baptista, J., Diniz, C., and Cabarrão, V. (2012). STRING: an Hybrid Statistical and Rule-based Natural Language Processing Chain for Portuguese. PROPOR '12 (Demo Session), Coimbra, Portugal. http://www.inesc-id.pt/ficheiros/publicacoes/8578.pdf.

[32] Mamede, N., Baptista, J., and Hagège, C. (2011). Nomenclature of Chunks and Dependencies in Portuguese XIP Grammar 3.0. Technical report, L^{2}F/INESC-ID, Lisboa.

[33] Maurício, A. (2011). Identificação, Classificação e Normalização de Expressões Temporais. Master's thesis, Instituto Superior Técnico, Lisboa.

[34] McCarthy, J. F. and Lehnert, W. G. (1995). Using Decision Trees for Coreference Resolution. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, IJCAI '95, pages 1050–1055, Montreal, Québec, Canada. Morgan Kaufmann Publishers Inc.

[35] Mitkov, R. (1999). Anaphora Resolution: the State of the Art. Technical report, University of Wolverhampton.

[36] Mitkov, R. (2002). *Anaphora Resolution*. Pearson Prentice Hall.

[37] Müller, C. and Strube, M. (2006). Multi-level Annotation of Linguistic Data with MMAX2. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang.

[38] Neves, M. and Leser, U. (2012). Tools for Annotating Biomedical Texts. In *5th International Biocuration Conference*, page 111. Washington D.C., USA.

[39] Ng, V. (2010). Supervised Noun Phrase Coreference Research: the First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

[40] Ng, V. and Cardie, C. (2002). Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 104–111, Philadelphia, PA, USA. Association for Computational Linguistics.

[41] Nobre, N. (2011). Resolução de Expressões Anafóricas. Master's thesis, Instituto Superior Técnico, Lisboa.

[42] O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for Text and Image Annotation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 13–16, Columbus, Ohio, USA. Association for Computational Linguistics.

[43] Ogren, P. V. (2006). Knowtator: a Protégé Plug-in for Annotated Corpus Construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, HLT-NAACL '06, pages 273–275, New York, USA. Association for Computational Linguistics.

[44] Paraboni, I. and Strube-de-Lima, V. L. (1998). Possessive Pronominal Anaphor Resolution in Portuguese Written Texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, COLING '98, pages 1010–1014, Montreal, Québec, Canada. Association for Computational Linguistics.

[45] Pereira, S. (2010). Linguistics Parameters for Zero Anaphora Resolution. Master's thesis, Universidade do Algarve and University of Wolverhampton.

[46] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[47] Rahman, A. and Ng, V. (2009). Supervised Models for Coreference Resolution. In *Proceedings of Empirical Methods in Natural Language Processing*, EMNLP '09, pages 968–977, Singapore. Association for Computational Linguistics.

[48] Ribeiro, R. (2003). Anotação Morfossintáctica Desambiguada em Português. Master's thesis, Instituto Superior Técnico, Lisboa.

[49] Rosário, L. (2007). Resolução de Anáforas e o seu Impacto em Sistemas de Recuperação de Informação. Master's thesis, Universidade de Évora.

[50] Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education.

[51] Santos, D. (1998). Disponibilização de *Corpora* de Texto através da WWW. *Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da APL, FLUL*, pages 323–335.

[52] Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

[53] Stark, M. M. and Riesenfeld, R. F. (1998). *WordNet: an Electronic Lexical Database*. MIT Press.

[54] Talhadas, R. (2013). Semantic Role Labelling. Master's thesis, Universidade do Algarve.

[55] Tapanainen, P. and Järvinen, T. (1997). A Non-Projective Dependency Parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington D.C., USA. Association for Computational Linguistics, Morgan Kaufmann Publishers, Inc.

[56] van Deemter, K. and Kibble, R. (2000). On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26(4):629–637.

[57] Versley, Y. (2006). A Constraint-based Approach to Noun Phrase Coreference Resolution in German Text. In *Konferenz zur Verarbeitung Natürlicher Sprache*, KONVENS '06, Konstanz, Germany. http://www.versley.de/konvens.pdf.

[58] Vicente, A. (2013). LexMan – um Segmentador e Analisador Morfológico com Transdutores. Master's thesis, Instituto Superior Técnico, Lisboa.

[59] Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *Institute of Electrical and Electronic Engineers (IEEE) Transactions on Information Theory*, 13(2):260–269.

[60] Voorhees, E., editor (1998). *Proceedings of the 7th Message Understanding Conference*. Science Applications International Corporation (SAIC).

[61] Widlöcher, A. and Mathet, Y. (2012). The Glozz Platform: a Corpus Annotation and Mining Tool. In *Proceedings of the 2012 Association for Computational Liguistics Symposium on Document Engineering*, DocEng '12, pages 171–180, Paris, France. Telecom ParisTech, Association for Computational Liguistics.

[62] Witten, I., Frank, E., and Hall, M. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, USA, Second edition.

[63] Zhang, H. and Su, J. (2004). Naïve Bayesian Classifiers for Ranking. In *Proceedings of the 15th European Conference on Machine Learning*, ECML '2004, Pisa, Italy. Springer.

Appendix A

Annotation Directives

Anaphora Annotation Guidelines

João Marques[1], **Jorge Baptista**[2], **Nuno Mamede**[1]

[1]Instituto Superior Técnico, INESC-ID Lisboa
Rua Alves Redol, 9 – Lisboa – Portugal

[2]Universidade do Algarve, FCHS
Campus de Gambelas – Faro – Portugal

May 9, 2013

This document aims at providing a sufficiently detailed set of guidelines to annotate anaphora in *corpora*. The aim is to explicitly providing the knowledge required for the annotator, so that s/he can perform the anaphora annotation task correctly and consistently throughout the corpus. The document assumes basic linguistic knowledge about *anaphora* and anaphoric relations, including the terms *antecedent* and *anaphor*.

These guidelines, therefore, explicitly define the types of anaphora that are to be considered (*target*, §A.2), the ones that should not be considered (*exclusions*, §A.3) and the decisions to make in special/problematic cases (§A.4).

All guidelines are complemented with suitable examples and their annotation.

A.1 Annotation process

General principles of annotation can be put forward:

- The annotation units are the tokens as shown in the Glozz annotation software[1], currently used for this annotation task.
- The anaphora relation is an oriented relation, signalled in the examples below by the symbol '←', which holds between an antecedent and an anaphor: (antecedent ← anaphor). The target anaphors are defined in §A.2.
- When an antecedent consists of more than one token as presented in Glozz, for example, multiword named entities, the last token belonging to the antecedent is chosen to represent it.
- *anaphora takes precedence over cataphora*, that is, if an antecedent is available in the previous discourse (anaphora), this takes precedence over another candidate *antecedent* if this appears after the anaphor (cataphora);
- the closest (to left), explicitely/unreduced antecedent is to be chosen;
- a token can be the antecedent of several anaphors, but an anaphor can only have one antecedent;
- an indefinite can not be anaphor (but it can be an antecedent of an anaphor);
- when one is unsure of whether a token is an anaphor, or whether a anaphora should be marked at all, it is better to use the "anaphora-uncertain" relation and not to introduce noise.

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

A.2 Target anaphors

This section lists the type of anaphoras that should be considered in the annotation process.

1. Personal Pronoun Anaphor;

   Ex. 1: O João deu uma prenda à Maria. Ela gostou muito.

   (Maria ← Ela)

   'João gave a gift to Maria. She liked a lot.'

   (Maria ← She)

   Ex. 2: A Maria Silva finalmente chegou. O Pedro pegou na mochila dela.

   (Silva ← Ela)

   'Maria Silva finally arrived. Pedro took her backpack.'

   (Silva ← She)

2. Possessive Pronoun Anaphor (both as a determinant (Ex. 3) and in headless NPs (Ex. 4));

   Ex. 3: O Pedro pegou na sua mochila.

   (Pedro ← sua)

   'Pedro took his backpack.'

   (Pedro ← his)

   Ex. 4: O Pedro distribuiu os lanches pelos colegas e cada um pegou no seu e foi comê-lo para a sala.

   (colegas ← seu; colegas ← cada um; lanches ← lo)

   'Peter distributed the snacks by his mates and each one took his own and went to eat it in the living room.'

   (mates ← his; mates ← each one; snacks ← it)

3. Demonstrative Pronoun Anaphor (in headless NPs);

   Ex. 5: Passos Coelho falou sobre a situação do país. Esta requer muitos cuidados.

   (situação ← Esta)

   'Passos Coelho spoke about the country's situation. This requires a lot of ministration.'

   (situation ← This)

   Ex. 6: Os adeptos foram ao estádio e cada um bebeu uma cerveja.

   (adeptos ← cada um)

   'The supporters went to the stadium and each_one enjoyed.'

   (supporters ← each one)

A similar case involves the so-called demonstrative pronoun o that can also be an anaphor, in headless noun phrases with relative clauses (see also Ex. 65, below).

4. Other Headless NPs

Other demonstrative and indefinite pronouns, and different types of numerals used as determiners can become the head of headless NPs (and PPs as well). A comprehensive list is provided below[2]. In these headless NPs, the determiner is considered an anaphor. The following situations can be considered:

- __ **[N]** (without any other determiner before): um, ambos, aquele, esse, este, tal, todos (just plural), algum, bastantes (almost only in plural), cada (and compound pronoun cada um), diversos, muito, nenhum, outro, pouco, tanto, uns quantos, uns tantos, vários, cardinal numerals.

  Ex. 7: O Pedro comprou vários livros. Um estava riscado.

  (livros ← Um)

  'Peter bought several books. One was scrawled.'

  (books ← One)

  Ex. 8: O Pedro comprou vários livros. Pagou por cada 10 euros.

  (livros ← cada)

  'Peter bought several books. [he] payed by each 10 euros.'

  (books ← each)

- **o** __ **[N]** (preceded by definite determiner, mostly the definite article o 'the'): o mesmo, o próprio, o restante, o tal, o outro, and cardinal numerals (excluding um 'one', and preceded by os: os três 'the three') and ordinal numerals (o primeiro 'the first'); consider also (and instead of the definite article o), combinations of some of these determiners, namely, mesmo, outro, and cardinal numerals, preceded by the definite, demonstrative pronouns este, esse and aquele, *e.g.* este mesmo, esse outro, aqueles três.

  Ex. 9: O Pedro comprou dois livros. Por este livro pagou 10 euros e pelo outro 15 euros.

  (livros ← outro)

  'Peter payed for this book 10 euros and for the other 15 euros.'

  (books ← other[3])

  Ex. 10: O Pedro e o João passeavam. O primeiro ia à frente e o segundo atrás.

  (Pedro ← primeiro; João ← segundo)

  'Pedro and João gave a walk. The first lead the way and the second went behind.'

  (Pedro ← fisrt; João ← second)

[2]For further information on the use of determiners in headless NPs, please refer to Baptista (2011) *Reclassification of Determinants in the L2F Lexicon*. Technical Report. Lisboa: L2F/INESC-ID Lisboa.

[3]The anaphora is established with the plural NP livros 'books' in the previous sentence and not with the singular NP livro 'book' in the same sentence, because there can be no referential identity between the later and the anaphor.

- **um __ [N]** (preceded by indefinite determiner, mostly the indefinite article um 'a'): um idêntico, um único (just singular), um qualquer, uns and cardinal numerals (just plural: uns três 'some three'), and um plus ordinal numeral (only singular: um terceiro); consider also (and instead of indefinite article) combinations of cardinal numerals and some of these determiners, *e.g.* três idênticos, três quaisquer.
- Complex partitive NPs with the form *NP1 de NP2* can have the head of the first NP zeroed, and in this case, there remaining determiner of NP1 is considered an anaphor, and its antecedent is the head of NP2; in this case, the reconstruction of the zeroed head is possible, though highly redundant (v.g. algumas ações das ações que detinha):

  Ex. 11: O Pedro comprou algumas das ações que detinha.

  (ações ← algumas)

  'Peter bought several of the bonds that [he] owned.'

  (bonds ← several)

  In other cases, these partitive NPs do not allow these reconstruction, hence, they are not considered anaphora:

  Ex. 12: O Pedro comprou **95**% das **ações**.

  'Peter bought **95**% of the **bonds**'

  Ex. 13: O Pedro e eu comprámos dois livros. O Pedro pagou pelo seu livro 10 euros mas eu paguei por um idêntico 15 euros.

  (livros ← idêntico)

  'Peter and I bougth two books. Peter payed for his book 10 euros but I payed for one/an identical 15 euros.'

  (books ← identical)
- The so-called demonstrative pronoun o:

  Ex. 14: As duas universidades, a_1 de Lisboa e a_2 de Évora, suspenderam as aulas.

  (universidades ← a_1; universidades ← a_2)

  'The two universities, the one_1 of Lisbon and the one_2 of Évora, suspended the classes.'

  (universities ← the one_1; universities ← the one_2)

  Ex. 15: A situação pode ser a de rotura de stocks.

  (situação ← a)

  'The situation can be the one1 of rupture of stock '

  (situation ← the one)

5. Reflex Pronoun Anaphor;

   Ex. 16: A Carolina penteou-se.

   (Carolina ← se)

'Carolina combed herself.'

(Carolina ← herself)

Ex. 17: A receita do sucesso deve-se a um processo financeiro.

(receita ← se)

'The recipe for success is due [itself] to a financial process.'

(recipe ← itself)

This last example illustrates a pronominal passive construction, v.g. deve-se a = é devida a, which is to be considered, for the purpose of anaphora annotation, as an ordinary reflexive pronominal construction.

6. Relative Pronoun Anaphor.

Ex. 18: O livro de que eu falara era horrível.

(livro ← que)

'The book that I had talked about was awful.'

(book ← that)

Ex. 19: O livro cujo autor se desconhecia era horrível.

(livro ← cujo)

'The book whose author was unknown was awful.'

(book ← whose)

Ex. 20: A sala em que estava era linda.

(sala ← que)

'The room in which I was was beautiful.'

(room ← which)

Ex. 21: A sala onde estava o Pedro era linda.

(sala ← onde)

'The room where Peter was beautiful.'

(room ← where)

Ex. 22: O modo como o Pedro falava era fascinante.

(modo ← como)

'The way as Peter talked was fascinating.'

(way ← as)

7. Cataphora

Cataphora, that is, phoric relations where the *anaphor* precedes the antecedent, is to be treated exactly as anaphora (but only if no prior instance of the antecedent occurs in the previous discourse) . The following examples illustrate situations of cataphora.

Ex. 23: Cansado de esperar por ela, o Pedro deixou uma mensagem à Maria.

(Maria ← ela)

'Tired of waiting for her, Pedro left a message to Maria.'

(Maria ← her)

Ex. 24: Embora a sua casa fosse longe, o Pedro convidou a Maria para jantar.

(Pedro ← sua)

'Though his home was far away, Pedro invited Maria for diner.'

(Pedro ← his)

Ex. 25: Depois desta, o Pedro já não se mete noutra alhada.

(alhada ← esta)

'After this one, Pedro will not get into another trouble.'

(trouble ← this one)

Ex. 26: Não se revendo na decisão, o Pedro abandonou a reunião.

(Pedro ← se)

'Not recognising himself in the decision, Pedro abandoned the meeting.'

(Pedro ← himself)

A.3 Exclusions

This section contains the types of anaphora that should not be considered in the annotation process (which signalled by the symbol '↚').

1. Zero Anaphora;

   Ex. 27: O João levantou-se. $\emptyset_{[Ele]}$ Acordara não há muito tempo.

   (João ↚ Ele)

   'João got up. $\emptyset_{[He]}$ woke up not long ago.'

   (João ↚ He)

   Ex. 28: Portugal deve renegociar a dívida tal como fez a Alemanha quando precisou $\emptyset_{[de\ o\ fazer]}$

   (renegociar a dívida ↚ o fazer)

   'Portugal should renegotiate the debt as did Germany when they needed $\emptyset_{[to\ do\ it]}$'

   (renegotiate the debt ↚ do it)

2. Lexical Anaphora;

Ex. 29: Cavaco Silva está em Belém, onde o Presidente da República vai falar ao país.

(Cavaco Silva ↚ Presidente da República)

'Cavaco Silva is in Belém, where the President of the Republic will speak to the nation.'

(Cavaco Silva ↚ President of the Republic)

3. Verb Phrase Anaphora;

Ex. 30: Portugal deve renegociar a dívida tal como fez a Alemanha.

(renegociar a dívida ↚ fez)

'Portugal should renegotiate the debt as did Germany.'

(renegotiate the debt ↚ did)

Ex. 31: Portugal deve renegociar a dívida. Para tal/tanto, precisa de convencer a UE.

(Portugal deve renegociar a dívida ↚ tal/tanto)

'Portugal should renegotiate the debt. For that, it needs to convince the EU.'

(Portugal should renegotiate the debt ↚ that)

4. Indirect Anaphora;

Ex. 32: Harry subiu a escada, saltando três degraus de cada vez.

(escada ↚ degraus)

'Harry climbed the stairs, leaping three steps at a time.'

(stairs ↚ steps)

5. Locative Anaphora;

Ex. 33: Fernando Pessoa viveu em Lisboa e lá conheceu Almada Negreiros.

(Lisboa ↚ lá)

'Fernando Pessoa lived in Lisbon and there he met Almada Negreiros.'

(Lisbon ↚ there)

6. Temporal Anaphora.

Ex. 34: Antes do 25 de Abril de 1974 não se podia dizer nada. As coisas eram bem diferentes então.

(Antes do 25 de Abril de 1974 ↚ então)

'Before April 25th 1974, one could say nothing. Things were quite different then.'

(Before April 25th 1974 ↚ then)

7. Interrogative and indefinite pronouns (not used as determiners).

Ex. 35: Diz-se que o país está em crise

(? ← se)

'One says/People say that the country is going through a crisis'

(? ← one/people; literally: itself)

Ex. 36: O supermercado já só tinha 10 laranjas. O Pedro comprou tudo (o que restava.)

(laranjas ↚ tudo)

'The supermarket only had 10 oranges. Peter bought everything that remained.'

(oranges ↚ everything)

Ex. 37: O Pedro perguntou aos amigos quem tinha feito isso.

(amigos ↚ quem)

'Peter asked his friends who had done it.'

(friends ↚ it)

Notice that the compound pronoun cada um 'each one' is to be treated as an indefinite pronoun that can not be used as a determinant:

Ex. 38: O Pedro perguntou isso a **cada um** dos **amigos**.

'Peter asked that to **each one** of the **friends**'

The pre-determiner todo 'all,entire' is not considered an anaphor when the NP it determines is present, irrespective of its' being placed at the left, or the right, or next to or distant from that NP:

Ex. 39: O Pedro comprou **todas** as **ações**.

'Peter bought **all** the **bonds**'

Ex. 40: O Pedro comprou as **ações todas**.

'Peter bought the **bonds all**'

Ex. 41: As **ações** caíram **todas**.

'The **bonds** fell **all**.'

Focalizers próprio and mesmo 'very/same/himself' are also not marked.

Ex. 42: O Pedro disse que ele **mesmo** faria isso

'Peter said that he **himself** would do that'

A.4 Special or problematic cases

This section introduces cases that are specially ambiguous and, therefore, problematic. They should be carefully read for the annotator to make the right decisions along the process.

1. Two candidates are the same: in this situation, choose the most recent unreduced antecedent:

   Ex. 43: O Pedro$_1$ gosta muito da Maria. O Pedro$_2$ é namorado dela$_1$. Ele vai a casa dela$_2$.

   (Pedro$_2$ ← Ele; Maria ← ela$_1$; Maria ← ela$_2$)

   'Pedro$_1$ is very fond of Maria. Pedro$_2$ is her$_1$ boyfriend. He is going to her$_2$ place.'

   (Pedro$_2$ ← He; Maria ← ela$_1$; Maria ← ela$_2$)

2. Coordinated antecedents: a relation is set between the anaphor and each of the NPs composing the antecedent, either in simple coordinations with e ('and') or ou ('or') or in enumerations where all but the last of a sequence of NPs are conjoined by comma (Ex. 44-45); the same also happens with coordinated PPs (Ex. 46-47):

   Ex. 44: O José e o Mário encontraram-se. Eles foram juntos a casa do Alberto.

   (José ← se; Mário ← se; José ← Eles; Mário ← Eles)

   'José and Mário met. They went to Alberto's place together.'

   (José ← They; Mário ← They)

   Ex. 45: O António, o José e o Mário encontraram-se. Eles foram juntos a casa do *Alberto*.

   (António ← Eles; José ← Eles; Mário ← Eles)

   'António, José and Mário met. They went to Alberto's place together.'

   (António ← They; José ← They; Mário ← They)

   Ex. 46: A Maria conversou com o José e com o Mário. Depois, foi com eles a casa do Alberto.

   (José ← eles; Mário ← eles)

   'Maria talked with José and Mário. Then, [she] went with them to Alberto's place.'

   (José ← them; Mário ← them)

   Ex. 47: A Maria conversou com o José, o Mário e com o Pedro. Depois, foi com eles a casa do Alberto.

   (José ← eles; Mário ← eles; Pedro ← eles)

   'Maria talked with José, Mário and Pedro. Then, [she] went with them to Alberto's place.'

   (José ← them; Mário ← them; Pedro ← them)

   Ex. 48: A vida humana é constituída de muitas coisas: amor, trabalho, ação. Estes são ...

   (amor ← Estes; trabalho ← Estes; ação ← Estes)

   'The human life is made of such things: love, work, action. These are ...'

   (love ← These; work ← These; action ← These)

   However, in the following case, an explicit antecedent can be found:

Ex. 49: A vida humana é constituída de muitas coisas: amor, trabalho, ação. Estas são ...

(coisas ← Estas)

'The human life is made of such things: love, work, action. These are ...'

(things ← These)

3. Appositions, including the relative pronouns of appositive-relative sub-clauses, can not be the antecedent of pronouns in the main clause:

Ex. 50: Massoud Barzani, líder do PDC, que foi eleito em maio passado, voltou-se para Saddam e disse umas boas

(Barzani ← se; Barzani ← que; líder ↚ que; líder ↚ se)

'Massoud Barzani, leader of PDC, who was elected last May, turned himself to Saddam and told some few good things'

(Barzani ← himself; Barzani ← who; leader ↚ who; leader ↚ himself)

Ex. 51: O subinspetor da Judiciária de Tomar, Manuel Silva, encontrou-se com a imprensa.

(subinspetor ← se; Silva ↚ se)

'The subinspetor of Tomar Police, Manuel Silva, met himself with the press.'

(subinspector ← himself; Silva ↚ himself)

Ex. 52: Pedro e João, três e cinco anos, miúdos reguilas que não sabiam ler, magoaram-se.

(miúdos ← que; Pedro ← se; João ← se; miúdos ↚ se)

'Peter and John, aged three and five, irreverent kids who did not know how to read, hurted themselves.'

(kids ← who; Peter ← themselves; John ← themselves; kids ↚ themselves)

Ex. 53: Seus olhos, indagadores holofotes, fixaram-se por muito tempo na baía anoitecida.

(olhos ← se; holofotes ↚ se)

'His eyes, inquisitive spotlights, settled themselves for long on the dusked bay.'

(eyes ← themselves; spotlights ↚ themselves)

Ex. 54: O Presidente da República (PR) declarou . . .

(Presidente da República ↚ PR)

'The President of the Republic (PR) declared . . .'

(President of the Republic ↚ PR)

The reverse situation, however, is to be considered. For example:

Ex. 55: Os feridos, todos militares, foram levados para o hospital.

(feridos ← todos)

'The injured, all military, were taken to the hospital'

(injured ← all)

4. When a pronoun can be linked to a set of candidates that belong to the same anaphoric chain, the antecedent should be the first instance (that is, the antecedent referring to the same entity and in an explicit, not reduced/zeroed form) that appears immediately to the left of the anaphor[4]. However, another antecedent must found elsewhere if the nearest candidate antecedent is a `PREDSUBJ` (see 5) or an `APPOSIT` (see 3).

   Ex. 56: O Pedro Santos, que$_1$ foi um grande actor, sempre se destacou pela sua$_1$ postura, pela sua$_2$ calma que$_2$ lhe valeram um Óscar em 2000.

   (Santos ← que$_1$; Santos ← se; Santos ← sua$_1$; Santos ← sua$_2$; actor ↚ sua$_2$; calma ← que$_2$; postura ← que$_2$; Santos ← lhe; actor ↚ lhe; actor ↚ se; actor ↚ sua$_1$; actor ↚ sua$_2$; actor ↚ lhe;)

   'Pedro Santos, who was a great actor, always stood himself out for his$_1$ posture, for his$_2$ calm that earned him an Oscar in 2000.'

   (Santos ← who; Santos ← himself; Santos ← his$_1$; Santos ← his$_2$; calm ← that; posture ← that; Santos ← him Santos ↚ himself; actor ↚ his$_1$; actor ↚ his$_2$; actor ↚ him)

5. When a headless `NP` has its antecedent candidate in an attributive position (in a `PREDSUBJ` dependency), no anaphoric relation should be extracted. The headless `NP` can appear both before or after the copula verb.

   Ex. 57: Essa é a razão desta carta.

   (razão ↚ Essa)

   'That is the reason for this letter.'

   (reason ↚ That)

   Ex. 58: Eram várias as sombras que se erguiam sobre este projeto.

   (sombras ↚ várias)

   'Were several the shadows that had risen over this project'

   (shadows ↚ several)

6. For headless NPs or PPs, when the head of an NP or PP is not a noun, and only when no previous instance of the reduced head noun can be found in the prior discourse, a relation is still set between the anaphor and the head of the NP/PP, even if the head of the NP/PP is a determiner, like a numeral (Ex. 59) or an adjective (60);

[4]This sentence is presented here as it appeared in a corpus, though it shows two errors. It should read pela sua postura **e** pela sua calma, que lhe valeram [...], that is, with explicitely coordinated PPs (no asyndeton) and a comma before the relative pronoun of the appositive-relative clause.

Ex. 59: As duas foram baratas e ela conseguiu integrá-las muito bem na decoração.

(duas ← las)

'The two were cheap and she managed to integrate them very well in the decor.'

(two ← them)

Ex. 60: Ela comprou periquitos amarelos e periquitos azuis e ofereceu à irmã os azuis. A irmã adorou-os.

(azuis ← os)

'She bought yellow parakeet and blue parakeet and [she] offered the blue ones to her sister. Her sister loved them.'

(blue ← them)

7. The partitive, complex NPs, of the form NP_1 *de* NP_2, where the first NP can be headless, *e.g.* um dos alunos ('one of the students'), the determiner of the NP_1 is considered an anaphor, and a cataphora relation is established between the former and the head of the NP_2:

   Ex. 61: Um dos alunos desistiu.

   (alunos ← um)

   'One of the students gave up.'

   (students ← one)

   The headless NP allows for the reconstitution of the zeroed head noun, which, theoretically, is considered the source of the NP, though the phrase may be deemed awkward because of its high redundancy v.g. um (aluno) dos alunos 'one student of the students'. This does not happen with clear-cut nominal determiners: uma dezena de alunos /*uma dezena de alunos de alunos 'a dozen of students'; in this case, no anaphora is considered.

   Another case involves complex determinants expressing ranges: Though a relatively acceptable reconstitution may be proposed, only the last numeral of the range may take role of anaphor:

   Ex. 62: De todos os livros disponíveis, o Pedro comprou entre **dois** e quatro.

   (livros ← quatro; livros ↚ dois)

   'From all the available books, Peter bough between **two** and four.'

   (books ← four; books ↚ two)

8. When the pronoun is a relative, a relation is set between the relative pronoun and its antecedent (Ex. 63-64).

   Ex. 63: O jogador cujo cabelo era branco.

   (jogador ← cujo)

   'The player whose hair was white.'

   (player ← whose)

Ex. 64: O jogador que tem o cabelo branco.

(jogador ← que)

'The player who has white hair.'

(player ← who)

9. We consider the so-called *relative clauses without antecedent* (Ex. 65) as headless NPs, that is, noun phrases whose head has been zeroed (zero anaphora). In this NPs, we consider that the relative pronoun refers to an antecedent, which is indicated by the article or by any another determiner of that headless NP. This determiner can vary in gender-number, and it agrees with the (zeroed) head noun of the NP. Like in all other cases involving relative pronouns, in this special case of relative clause, the relative pronoun should be linked to its antecedent, even though it is necessary to find it, not immediately before the pronoun, from where it has been zeroed, but at a prior instance of the antecedent, further to left:

   Ex. 65: O Pedro comprou vários jornais mas leu apenas os $\emptyset_{[\text{jornais}]}$ que falavam desse assunto.

   (jornais ← os; jornais ← que)

   'Pedro bought several newspapers but only read the ones $\emptyset_{[\text{newspapers}]}$ that talked about this subject.'

   (newspapers ← ones; newspapers ← that)

10. A similar case involves an indefinite (and invariable) pronoun o as the antecedent of the relative pronoun (Ex. 66), often determined by quantifiers like tudo (Ex. 67):

    Ex. 66: O Pedro vendeu o que tinha.

    (o ← que)

    'Pedro sold everything $_{\text{that}}$ he had.'

    (everything ← that)

    Ex. 67: O Pedro vendeu tudo o que tinha.

    (o ← que)

    'Pedro sold everything that he had.'

    (everything ← that)

    Other invariable indefinite quantifiers like tudo include algo and nada.

11. The interrogative pronouns in general – particularly in the so-called *pseudo-relatives*, which are in fact, subordinate completive (partial interrogative) subclauses –, are not considered as anaphors, as they are indefinites:

    Ex. 68: **O que** fizeste?

    '**What** did [you] do.'

Ex. 69: **Quem** está aí?

'**Who** is there?'

Ex. 70: Não sei **o que** fazer.

'I do not know **what** to do.'

Ex. 71: Perguntei **quem** fizera isso.

'I asked **who** did that.'

12. In so-called *cleft sentences* with ser ... que, which formally resemble relative clauses, the *relative* pronoun que is not considered an anaphor:

    Ex. 72: O Pedro é **quem**/**que** fizera isso.

    'It was Pedro **who**/**that** did that.'

    Ex. 73: Foi o Pedro **quem**/**que** fez isso.

    'It was Pedro **who**/**that** did that.'

    Ex. 74: Foi este livro **que** o Pedro leu.

    '[It] was this book **that** Peter read.'

    Ex. 75: Foi neste livro **que** o Pedro leu isso.

    '[It] was in this book **that** Peter read that.'

13. Certain multiword, time expressions constitute a single token, hence they should not be annotated even if they include anaphor-like elements. However, in some cases, a word may not be tokenised correctly in the corpus, *e.g.* no ano que vem, but in such cases no anaphor should not be considered also.

14. When the pronoun quantifies or otherwise determines the head of noun phrase, no anaphora relation should be established;

    Ex. 76: O Pedro passou a **algumas** disciplinas.

    'Pedro passed **several** subjects.'

    Ex. 77: **Esta** semana, o Pedro está de férias.

    '**This** week, Pedro is on vacancies.'

    This also happens in certain uses of demonstrative pronouns (*e.g.* esse 'that, such'), which are equivalent to classifiers (v.g. esse tipo de 'that type of', desse tipo 'of such type'):

    Ex. 78: Podia lá eu dizer uma coisa **dessas**!

    'I would not say **such** a thing.'

15. In the case of generic (nós, eles) and indefinite (se) personal pronouns, no co-reference is established:

Ex. 79: Quando **nós** vemos a política deste governo...

'When **we** see this Government politics...'

Ex. 80: Quando **se** vê a política deste governo...

'When **one** sees this Government politics...'

1st and 2nd person pronouns, that usually appears in dialogue, should not be considered:

Ex. 81: Eu conhecia os soldados. Inclusivamente, um tio **meu** foi soldado naquele tempo.

'I knew the soldiers. Inclusively, an uncle of **mine** was a soldier at that time.'

3rd person pronouns with 2nd person value such as você e V. Exa. should not be considered as well:

Ex. 82: Não sei se **você** concorda.

'I do not know if **you** agree.'

Ex. 83: Qual é a **sua** música favorita, Jorge?

'What is your favorite song, Jorge?'

16. In the case of intrinsically reflexive pronouns, like suicidar-se ('to commit suicide') the general case of reflexive is applied:

    Ex. 84: O Pedro suicidou-se.

    (Pedro ← se)

    'The Pedro suicide himself, 'Pedro committed suicide'.'

    (Pedro ← himself)

    Exception to the above is made to the reflexive impersonal of tratar-se ('to concern') and the like, as there is no subject.

    In the same way, no anaphora is marked for reflexive pronouns involved in sentential subjects in reflexive verb constructions:

    Ex. 85: Torna**-se** difícil compreender isso.

    '[It] becomes difficult to understand that.'

17. In the case of insoluble ambiguity in co-reference, considering the context, the nearest antecedent is chosen:

    Ex. 86: A Maria finalmente chegou. O Pedro pegou na sua mochila e foi ter com ela.

    (Pedro ← sua chosen over Maria ← sua; Maria ← ela)

    'Maria finally arrived. Pedro took his backpack and went to meet her.'

    (Pedro ← his; Maria ← her)

18. In the case where the relation between the anaphor and the antecedent are not co-referential (same real-life object) but instead refer to an object of similar description (identity-of-sense anaphora), no anaphora should be considered:

    Ex. 87: Uma **mão** na cicatriz e a **outra** estendida no escuro.
    'A **hand** in the scar and the **other** stretched in the dark.'

    Ex. 88: Bebeu um copo de vinho. A seguir a este, vieram **outros**.
    '[He] drunk a cup of wine. After this one, other [cups] came'

    Ex. 89: Eu não quero o **casaco** azul. Eu prefiro **o** preto.
    'I do not want the blue **jacket**. I prefer the black **one**.'

19. In the presence of the so-called *echo complements* such as um Prep outro ('each other'), each pronoun should be linked to its singular antecedent in the *reversed* order in which they appear:

    Ex. 90: A religião e a ciência precisam uma da outra.
    (ciência ← uma; religião ← outra)
    'Religion and science need each other.'
    (science ← each; Religion ← other)

    Ex. 91: A ciência precisa da religião e a religião precisa da ciência. Ou melhor: nenhuma delas precisa uma da outra porque são ambas autónomas.
    (religião ← elas; ciência ← elas; ciência ← uma; religião ← outra; religião ← ambas; ciência ← ambas)
    'Science needs religion and religion needs science. Or rather, none of them needs each other because they are both autonomous.'
    (religion ← them; science ← them; religion ← each; science ← other; religion ← they; science ← they; religion ← both; science ← both)

20. When an anaphor is a part (or member) of the antecedent entity (or collective of entities), an anaphoric (metonymic) relation should be considered:

    Ex. 92: A Joana comprou 10 maçãs. Várias estavam podres.
    (maçãs ← Várias)
    'Joana bought 10 apples. Several were rotten.'
    (apples ← Several)

    The reversed situation is also possible, that is, the anaphor constitutes the set of entities from which the antecedent is a part (or member). Therefore, the two following examples should be represented as shown:

Ex. 93: Este encontro foi o primeiro de vários em que se tratou deste assunto.

(encontro ← primeiro; encontro ← que; encontro ↚ vários)

'This meeting was the first of many in which this matter was discussed.'

(meeting ← first; meeting ← which; meeting ↚ many)

Ex. 94: Este foi o primeiro de vários encontros em que se tratou deste assunto.

(encontros ← este; encontro ← primeiro; encontro ← que)

'This was the first of many meetings in which this matter was discussed.'

(meetings ← This; meetings ← first; meetings ← which)

Notice that in the first example the determinant vários is a headless NP, and therefore it is in fact an anaphor. However, since it has not the same referent as the antecedent of primeiro 'first', no anaphora can established.

A similar situation can be found in the headless NPs in relative-appositive sub-clauses:

Ex. 95: Os colaboradores, um dos quais era muito magro, disseram ...

(colaboradores ← um; colaboradores ← quais)

'The employees, one of which was extremely thin, said ...'

(employees ← one; employees ← which)