PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

THIAGO MACHADO LIMA

ANALYSING SEMANTIC RESOURCES FOR COREFERENCE RESOLUTION

Porto Alegre
2019

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

ANALYSING SEMANTIC RESOURCES FOR COREFERENCE RESOLUTION

THIAGO MACHADO LIMA

Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fullfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. Dr. Renata Vieira
Co-Advisor: Dr. Sandra Collovini

Porto Alegre
2019

Ficha Catalográfica

L732a Lima, Thiago Machado

Analysing Semantic Resources For Coreference Resolution / Thiago Machado Lima . – 2019.
69.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Renata Vieira.
Co-orientadora: Profa. Dra. Sandra Collovini.

1. Coreference Resolution. 2. Semantic Knowledge. 3. Corpus Analysis. I. Vieira, Renata. II. Collovini, Sandra. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Salete Maria Sartori CRB-10/1363

Thiago Machado Lima

Analysing Semantic Resources For Coreference Resolution

This Dissertation has been submitted in partial fulfillment of the requirements for the degree of Master of Computer Science, of the Graduate Program in Computer Science, School of Computer Science of the Pontifícia Universidade Católica do Rio Grande do Sul.

Sanctioned on March, 2019

COMMITTEE MEMBERS:

Prof. Dr. Thiago Alexandre Salgueiro Pardo (ICMC/USP)

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS)

Prof. Dr. Renata Vieira (PPGCC/PUCRS - Advisor)

Dr. Sandra Collovini (Co-Advisor)

ACKNOWLEDGMENTS

To my parents, for my education and for having always encouraged me.

To my advisors, Renata Viera and Sandra Collovini, for the confidence they placed in me, their guidance and constant support.

To my colleagues from the Computer Science Graduate Program, especially the Natural Language Processing lab, for being always helpful.

To my work colleagues, for the comprehension when I had to leave during working hours for my academic work.

To PUCRS, to whom I owe my academic background.

Finally, to Hewlett Packard, for their financial support[1].

[1]This thesis was achieved in cooperation with Hewlett Packard Brasil LTDA. using incentives of Brazilian Informatics Law (Law nº 8.2.48 of 1991)

ANÁLISE DE RECURSOS SEMÂNTICOS PARA RESOLUÇÃO DE CORREFERÊNCIA

RESUMO

Resolução de Correferência é uma tarefa que consiste em identificar menções em um discurso que se referem a uma mesma entidade. A tarefa tem o potêncial de aprimorar outras tarefas de Processamento de Linguagem Natural como análise de sentimentos, extração de informação, sistemas de pergunta-resposta, entre outras. Algumas relações de correferência podem ser identificadas utilizando-se regras lexicais e sintáticas, enquanto para outras é necessário conhecimento semântico. No entanto, poucos trabalhos de resolução de correferência focaram em melhorias que possam ser realizadas com conhecimento semântico. O objetivo deste trabalho é aprimorar a tarefa de resolução de correferência utilizando semântica. Para isso, foram revisados os recursos semânticos disponíveis para o Português, dos quais foram selecionados para os experimentos o ContoPT, o ConceptNet e um modelo de word embeddings. Os experimentos foram realizados no CORP, uma ferramenta de correferência para o Português que já utiliza o OntoPT como recurso semântico. A avaliação foi composta pelas métricas MUC, B_3 e $CEAF_e$, utilizando-se os corpora Corref-PT e Summ-it++. Ao comparar com o OntoPT, obtivemos melhores resultados em termos de Medida-F utilizando o ContoPT e o ConceptNet. Nos experimentos com a regra de similaridade semântica que utiliza o modelo de word embeddings não foi posível atingirmos os resultados obtidos com as bases semânticas estruturadas. Textos com mais relações semânticas foram selecionados para análise de erros, na qual observamos algumas dificuldades envolvendo a detecção de relacionamentos semânticos. Para tratar essas dificuldades foram propostas melhorias. Como contribuição este trabalho traz, além da análise das bases, uma nova versão do CORP integrada com três novos recursos semânticos. A nova versão obteve uma maior Medida-F utilizando semântica em relação à versão anterior que utiliza o OntoPT.

Palavras-Chave: Resolução de Correferência, Conhecimento Semântico, Análise de Corpus.

ANALYSING SEMANTIC RESOURCES FOR COREFERENCE RESOLUTION

ABSTRACT

Coreference Resolution is the task that consists of identifying mentions in a discourse that refer to the same entity. The task has the potential to improve other Natural Language Processing tasks such as sentiment analysis, information extraction, question answering, and others. Some coreferent relationships can be identified using lexical and syntactical rules, while others require semantic knowledge. However, few works focus on the possible improvements of using semantic knowledge. This work's objective is to improve the coreference resolution task by using semantic knowledge. For that, we reviewed the semantic resources available for the Portuguese language, and selected ContoPT, ConceptNet and a word embedding model for our experiments. Experiments were performed using CORP, a coreference tool for the Portuguese language which already uses OntoPT as a semantic resource. The evaluation was composed of metrics MUC, B_3 and $CEAF_e$, using Corref-PT and Summ-it++ as corpora. Compared to OntoPT, we obtained better results in terms of F-Measure using ContoPT and ConceptNet. The experiments using the semantic similarity rule based in word embeddings was not able to surpass the results obtained with the structured semantic bases. Texts with more semantic relationships were selected for error analysis, and we were able to observe some difficulties involved in the detection of semantic relationships. To overcome these difficulties improvements are proposed. Besides the analysis of available semantic basis, this work brings as contribution a new CORP version, integrated with three new semantic resources, which obtained a higher F-Measure using semantics than the version that uses OntoPT.

Keywords: Coreference Resolution, Semantic Knowledge, Corpus Analysis.

LIST OF TABLES

Table 2.1 – Variables used for Precision, Recall and F-measure 25
Table 4.1 – Hypernym relation between synsets 37
Table 4.2 – ConceptNet Relations 39
Table 4.3 – Corref-PT - Corpus Statistics 42
Table 4.4 – CORP Rules 43
Table 6.1 – Semantic bases - Corref-PT 52
Table 6.2 – Semantic bases - Summ-it++ 52
Table 6.3 – Hypernym Window Sizes - CorrefPT 53
Table 6.4 – Hypernym Window Sizes - Summ-it++ 53
Table 6.5 – Corpus Relationships - Summ-it++ 54
Table 6.6 – Corpus Relationships - Corref-PT 54
Table 6.7 – Semantic Subset - Semantic Bases - Corref-PT 54
Table 6.8 – Semantic Subset - Semantic Bases - Summ-it++ 54
Table 6.9 – Example Chain 1 55
Table 6.10 – Example Chain 2 56
Table 6.11 – Semantic Similarity Experiments - Corref-PT 58
Table 6.12 – Semantic Similarity Experiments - Summ-it++ 58
Table 6.13 – Positive examples found 58

CONTENTS

1 INTRODUCTION **19**
1.1 MOTIVATION 20
1.2 RESEARCH GOALS 20
1.2.1 GENERAL GOALS 20
1.2.2 SPECIFIC GOALS 21
1.3 DISSERTATION OUTLINE 21

2 BACKGROUND **23**
2.1 REFERRING EXPRESSION 23
2.2 NAMED ENTITIES 23
2.3 NOUN PHRASE 23
2.4 ANAPHORA 24
2.5 COREFERENCE RESOLUTION 24
2.6 EVALUATION OF COREFERENCE RESOLUTION MODELS 25
2.6.1 MUC 26
2.6.2 B^3 26
2.6.3 CEAF 27
2.7 COREFERENCE AND SEMANTIC RELATIONS 27
2.7.1 HYPONYMY AND HYPERONYMY 28
2.7.2 SYNONYMY 28
2.7.3 WORD EMBEDDING MODELS AND SEMANTIC SIMILARITY 28
2.7.4 CHALLENGES 29

3 RELATED WORK **31**
3.1 COREFERENCE RESOLUTION 31
3.1.1 PARSED TREE AND RHETORICAL STRUCTURE 31
3.1.2 CORPUS-BASED AND SUPERVISED LEARNING APPROACHES 32
3.1.3 UNSUPERVISED LEARNING 32
3.1.4 DEEP LEARNING MODELS 33
3.2 COREFENCE RESOLUTION AND SEMANTICS 33
3.2.1 SEMANTICS FOR COREFERENCE IN PORTUGUESE LANGUAGE 35
3.3 SUMMARY 35

4 RESOURCES 37
4.1 SEMANTIC BASES FOR PORTUGUESE 37
4.1.1 OPENWORDNET-PT 38
4.1.2 PORTUGUESE UNIFIED LEXICAL ONTOLOGY (PULO) 38
4.1.3 ONTOPT 38
4.1.4 CONTOPT 38
4.1.5 CONCEPTNET 39
4.1.6 FRAMENET BRASIL 39
4.1.7 BABELNET 39
4.1.8 DBPEDIA 40
4.1.9 NILC WORD EMBEDDINGS 41
4.1.10 SEMANTIC RESOURCES CONSIDERED IN THIS WORK 41
4.2 CORPORA 42
4.2.1 CORREF-PT 42
4.2.2 SUMM-IT++ 42
4.3 CORP 43
4.3.1 HYPONYMY RULE 44
4.3.2 SYNONYMY RULE 45
4.4 CONLL REFERENCE SCORER 45

5 DEVELOPMENT 47
5.1 CONLL SCORER SCRIPT 47
5.2 SEMANTIC BASES INTEGRATION 47
5.2.1 CONTOPT 47
5.2.2 CONCEPTNET 48
5.3 WORD EMBEDDINGS WEB API 48
5.4 EXECUTION PARAMETERS 49

6 EXPERIMENTS AND RESULTS 51
6.1 EVALUATING SEMANTIC BASES 51
6.2 HYPERNYM WINDOW 52
6.3 SEMANTIC SUBSET ANALYSIS 53
6.3.1 ERROR ANALYSYS 54
6.4 SEMANTIC SIMILARITY BASED ON WORD EMBEDDINGS 57
6.4.1 RESULTS 57

6.4.2 ERROR ANALYSIS 58

7 FINAL CONSIDERATIONS 61

7.1 CONTRIBUTIONS 61

7.2 FUTURE WORK 62

REFERENCES 63

1. INTRODUCTION

The recent growth in both generation of and access to information has produced a demand for tools able to analyze this data. One specific area of interest in the study of computational systems pertains to working with natural language. The comprehension necessary for this kind of task is not easily achieved, and current tools need to be improved so that these new necessities might be met. One of the challenges in the area of Natural Language Processing, as this field is called, is enabling a machine to comprehend the semantic meaning underlying a text and from that meaning extract knowledge.

An important task for computational linguistics is coreference resolution, which is the task of identifying all expressions that refer to the same entity in a text. Many of the efforts in this task focus on machine learning techniques [65], but these techniques depend on the quality and quantity of training samples. Systems based on linguistic rules, on the other hand, are an option for low resourced languages, since they wouldn't require a large training dataset.

Most of the linguistic approaches for coreference resolution focus on lexical and syntactic knowledge, which are imperative for the task, but semantic knowledge is necessary in some cases. For instance: "Adalberto Portugal informed he will stay in Portugal until necessary...". From the syntactic point of view, we could have a referential relation between [Adalberto Portugal] and [Portugal], but the former refers to a person, and the latter to a place. On the other hand, in the sentence "Do bees make honey? The insects go in search of...", one can notice that there isn't any lexical or syntactic evidence to establish a coreferential relation between [bees] and [the insects], even though a relation exists. There are cases where semantic knowledge is not enough, and we must take context (pragmatics) into account. "The minister of agriculture has a department dedicated to family farming". Regarding semantic relations, [agriculture] and [farming] have a degree of semantic similarity, but they refer to different entities in this context. These examples showcase the importance of semantic knowledge to this task and how complex it can be.

This work focused on improving the coreference resolution task using semantic knowledge. We will focus on the Portuguese language, which makes for an even greater challenge, as resources for Portuguese are scarcer than for other languages such as English. To complete this task, we will use CORP [14], a coreference resolution system for the Portuguese language, and semantic bases available for Portuguese.

1.1 Motivation

Many researchers [30] [43] [16] [59] remark on the possibility of using semantic knowledge to improve coreference resolution systems. Relationships seen in (1.1.1) and (1.1.2) cannot be identified using solely lexical and grammatical rules. To find these relationships, we need to know that "plane" and "aircraft" are synonyms and that Muttur is a city in Sri Lanka.

(1.1.1) Ao menos 17 pessoas morreram após a queda de um avião de passageiros... A aeronave se chocou com uma montanha...
At least 17 people died after a passenger plane crash... The aircraft crashed into a mountain...

(1.1.2) Quinze voluntários da ONG francesa Ação Contra a Fome foram assassinados no nordeste do Sri Lanka... Os crimes aconteceram na cidade de Muttur...
Fifteen volunteers from French NGO Action Against Hunger were killed in northeastern Sri Lanka... The crimes happened in the city of Muttur...

Although some studies explore these situations, most of them focused on the coreference resolution task itself, using semantic knowledge as an additional feature and exploring one or two semantic bases. Our work focused on the improvement offered by the use of semantic knowledge and in exploring different semantic bases and relations contained into them. Moreover, we worked with the Portuguese language, which is less studied than many other languages. Any possible improvements we identify could also aid in other Natural Language Processing tasks that use coreference resolution, such as document summarization or information extraction.

1.2 Research Goals

1.2.1 General Goals

This work's goal is to analyse available Portuguese semantic basis to improve coreference resolution using semantic knowledge for the Portuguese language. To do so, we used an existing coreference tool, CORP, detailed in Section 4.3, and semantic bases available for Portuguese, described in Section 4.1. Since CORP currently only identifies nominal coreference, we focused on this category.

1.2.2 Specific Goals

- Review the current semantic bases available for the Portuguese language;
- Select useful relationships for the task of coreference resolution task;
- Improve CORP to allow the use of different semantic bases;
- Experiment with different semantic bases and relations;
- Evaluate results.

1.3 Dissertation Outline

This work is organized in the following way: chapter 2 focuses on the theoretic foundations behind the coreference resolution task; chapter 3 presents the related work found in the literature, including efforts for the Portuguese language; chapter 4 describes the resources used in the research, including semantic resources, the coreference tool, and corpora; chapter 5 presents improvements implemented in CORP in order to perform our experiments; chapter 6 shows the experiments performed and a discussion about the results; chapter 7 presents the contributions of this work and proposals for future work.

2. BACKGROUND

This chapter aims to cover the key topics related to our research. We describe coreference resolution and its related concepts. We also give an overview of the relationship between coreference and semantics. Examples are in Brazilian and European Portuguese.

2.1 Referring expression

A referring expression, or mention, is a natural language expression used to make reference to a real-world entity. The entity that is referred to by the expression is called the referent [28]. These references can be a named entity or be part of a noun phrase. For instance, in the sentence "John was visited by an admirer who was interested in the books he was reading", "John" and "he" are referring expressions, and "John" is their referent.

2.2 Named Entities

Named Entities are elements used to refer to objects or entities in a discourse or domain [74] that could be the target of a referring expression. Entities can be people, companies, places, a specific area's terminology like genes and proteins, among others. Examples (2.2.1) and (2.2.2) show named entities underlined.

(2.2.1) A PF (Polícia Federal) prendeu na manhã desta sexta-feira...
The FP (Federal Police) arrested this Friday morning...

(2.2.2) A seleção brasileira masculina de vôlei conseguiu...
The Brazil men's national volleyball team achieved...

2.3 Noun Phrase

A noun phrase is a unit formed by one or more words that, together, play a syntactic role in a sentence. These words revolve around a head, the central noun in the noun phrase. The head of a noun phrase can be a common noun, a proper noun or a pronoun. For instance: "The Brazilians are in third place". In that sentence, "The Brazilians" is a noun phrase with the determiner "The", and the head "Brazilians".

2.4 Anaphora

To understand coreference, we must first define Anaphora. Anaphora is a reference to an entity that has been previously introduced into the discourse. The previous reference is called the antecedent. For instance, in the sentence: "Daniel went to Bill's car dealership. He perused for about an hour", "He" is the anaphora, and "Daniel" is the antecedent.

2.5 Coreference Resolution

There are anaphora cases where the anaphoric term and its antecedent refer to the same entity. In that case, they are said to be coreferents. In the example (2.5.1) "o animal" evokes the referential value of its antecedent, "o touro", therefore these terms are coreferent. There are anaphora cases without coreference. As shown in example (2.5.2), the expression "no ano seguinte" has its referential value built upon its antecedent "em 2002", but it refers to the next year, 2003. There are also cases of coreference without anaphora, as in (2.5.3), where we can remove "Dilma Vana Rousseff" and still understand the sentence.

(2.5.1) Os outros galegos lançam-se, imediatamente, sobre o touro segurando-o pelas pernas, pontas e cauda, ou mesmo montam sobre ele, até que o animal...
The other galicians immediately throw themselves on the bull, holding it by his legs, ends and tail, or even riding on it, until the animal...

(2.5.2) Ele começou o trabalho em 2002 e, no ano seguinte, o estudo começou a ser feito.
He started the work in 2002 and, in the next year, the study began.

(2.5.3) Dilma Vana Rousseff, a primeira mulher a ocupar a presidência brasileira, nasceu em...
Dilma Vana Rousseff, the first woman to hold the Brazilian presidency, was born in...

Coreference resolution is the process of determining whether two expressions are coreferent. We can have two coreference types: identity and appositive [57]. The identity type is used for anaphoric coreferences, which are links between pronominal, nominal and named mentions of specific referents. The appositive type occurs when an appositive construction exists, that is, a noun phrase that modifies an adjacent noun phrase. Appositive coreference serves to rename or further define the entity. In example (2.5.4) we can notice that "a cidade" is coreferent to "Kashiwazaki", and thus is an identity coreference. In example (2.5.5) we can see that "mamão, melancia, abacate e uva" are coreferent to "frutas", and thus is an appositive coreference.

(2.5.4) Kashiwazaki, a cidade mais afetada pelo tremor...
Kashiwazaki, the city most affected by the quake...

(2.5.5) Maria comprou várias frutas: mamão, melancia, abacate e uva.
Maria bought many fruits: papaya, watermelon, avocado and grape.

A group of mentions that refers to the same entity can be grouped together in a set called a coreference chain. These coreference chains serve to group different parts of a discourse together, making it cohesive. For instance, in the following example, we have a chain composed of `[economista Solange Vieira] [economista] [Solange Vieira]`.

(2.5.6) O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias... Ainda não está definida a diretoria que a economista vai assumir. Inicialmente, Solange Vieira, que...
The Defense Minister, Nelson Jobim, should recommend economist Solange Vieira to assume one of the boards... The board to be assumed by the economist has not yet been defined. Initially, Solange Vieira, who...

2.6 Evaluation of Coreference Resolution Models

Most metrics used to evaluate coreference models are based on three basic measures: Precision, Recall and F-Measure. Table 2.1 show us the variables used by these measures.

Table 2.1 – Variables used for Precision, Recall and F-measure

Variable	**Explanation**
T_p	True positive; correctly classified positive
F_p	False positive; incorrectly classified positive
F_n	False negative; incorrectly classified negative

Precision is the fraction of relevant mentions found among all identified mentions. Recall is the fraction of relevant mentions found over the total quantity of mentions that should have been identified. F-measure, also called F1, is the harmonic mean of precision and recall. It is used to provide a single measurement to represent the previous two.

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + F_n}$$

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Each metric has its own calculus for Precision and Recall. Different details, as mention detection and chains generated, should be taken into account when evaluating coreference systems, and there is currently no agreement on a standard measure for evaluation since each metric evaluates a specific aspect of the model. The most important metrics are MUC [75], B^3 [3], CEAF [34] and BLANC [61]. We used CoNLL Average Scorer, described in section 4.4, to evaluate our experiments. CoNLL Average Scorer is composed of an average of MUC, B^3 and $CEAF_e$.

2.6.1 MUC

MUC is a link-based metric. Considering *S* to be the set of the coreference system's output links and *R* to be the set of the reference chain's links, precision and recall are defined as follows:

$$Precision = \frac{|S \cap R|}{S}$$

$$Recall = \frac{|S \cap R|}{R}$$

The biggest shortcoming of this metric is that, since it only verifies the links between mentions, it does not distinguish between entity detection errors and linking errors, thus favoring systems which produce fewer entities.

2.6.2 B^3

To address MUC's shortcomings, Bagga and Baldwin [3] proposed the B^3 metric. This metric attempts to represent clustering effectiveness, and thus calculates precision and recall by comparing the entities. Precision and recall are first computed for each entity, then the weighted sum of those entities' precision and recall is used as the final scores for precision and recall. Figure 2.1 shows three reference chains, and figures 2.2 and 2.3 show two different outputs generated by the coreference system. MUC returns the same F1 for both outputs. B^3, however, considers output (a) to be better than output (b), since the latter has merged two big entities. Because of that, B^3 output will be higher for (a) than for (b).

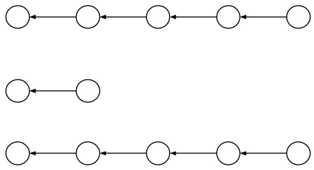

Figure 2.1 – Reference

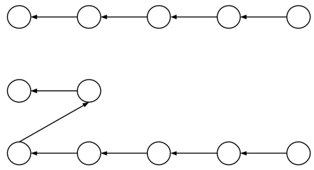

Figure 2.2 – Output (a)

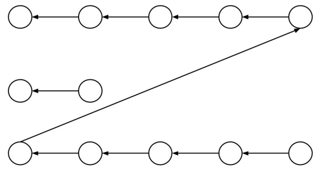

Figure 2.3 – Output (b)

2.6.3 CEAF

Luo [34] proposed the Constrained Entity-Aligned F-Measure (CEAF) to overcome some B^3 and MUC shortcomings. The metric aligns the system output and reference entities by maximizing the total entity similarity between them. There are two variants, $CEAF_m$ and $CEAF_e$: the former represents the percentage of mentions that are in the correct entities, and the latter the percentage of correctly recognized entities.

This metric avoids counterintuitive situations that can happen with B^3. For instance, considering the reference chains from figure 2.1 and the two outputs from figures 2.4 and 2.5, B^3 precision for output (d) is 1.0. B^3 recall for output (c) is also 1.0, since all entities are found after intersecting the reference chains with the system's response.

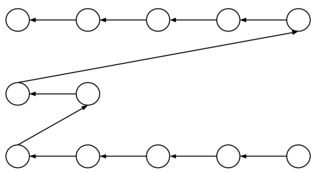

Figure 2.4 – Output (c)

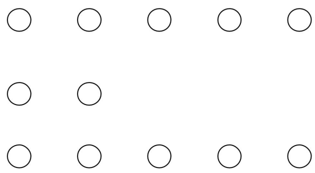

Figure 2.5 – Output (d)

2.7 Coreference and Semantic Relations

Some coreference cases can be identified because both noun phrases are exactly the same, and others can be identified using lexical or syntactical techniques. In some cases, however, semantic knowledge is required. This section describes how semantic relations can be used in coreference resolution. We also describe two relation types that can infer a coreference relation and are already being used by CORP (described in section 4.3).

2.7.1 Hyponymy and Hyperonymy

Hyponymy is a relation that happens when a word's semantic meaning is included within another word, its hyperonym. For instance, "dog" and "cat" are hyponyms of "animal" (their hyperonym). Hyperonymy and hyponymy relations are important because hyperonyms are often used in discourse to refer to a previously mentioned entity. This is done in order to avoid unnecessary repetitions that could make reading tiresome. In the sentence "John is happy with his dog. The animal is a loyal companion", "animal" was used to avoid repetition.

2.7.2 Synonymy

Synonymy is the relation between a word or phrase with the exact or near exact same meaning as another word or phrase in the same language. For instance, "boy" and "kid" are synonyms in English, "menino" and "garoto" are synonyms in Portuguese. Since a word can have more than one meaning, synonymy depends on the context around the words. As with hyponymy and hyperonymy relations, synonymy can be used in discourse to avoid repetition: "The youth is intelligent. However, the boy lives a difficult life".

2.7.3 Word Embedding Models and Semantic Similarity

Word Embeddings are distributed representations of words as vectors, which can be generated in a unsupervised way. It was found that the similarity between these word representations goes beyond simple syntactic regularities [40]. The vectors contain subtle semantic relationships between words. It is possible to perform simple algebraic operations on these word vectors. For example, `vector("rei") - vector("homem") + vector("mulher")`[1] results in a vector similar to `vector("rainha")`[2].

Semantic similarity is a metric used to define the relatedness between the meanings of word vectors. Relatedness is based on some common properties which vectors may share. Synonymy and hypernymy relations may be, among others, encoded into them.

Coreference resolution may benefit from these word embedding models by using them as semantic resources. Also, coreference resolution systems based on end-to-end neural networks can use the vectors directly for training.

---

[1] vector("king") - vector("man") + vector("woman")

[2] vector("queen")

2.7.4 Challenges

Semantic knowledge may be obtained using semantic bases, described in section 4.1. Although this knowledge may improve the coreference resolution task, the relations contained in these semantic bases are more susceptible to ambiguity, and therefore can reduce precision. Regarding the Portuguese language, semantic resources are scarcer. Two noun phrases can refer to different entities and yet have a semantic relation between them, as shown in following sentence.

(2.7.1) Todos estamos familiarizados com a mosca doméstica, a qual alimenta-se de matéria orgânica em decomposição, e é praticamente inofensível para outros insetos.
"We are all familiar with the housefly, which feeds on decaying organic matter, and is pretty much harmless to other insects"

3. RELATED WORK

This chapter presents the literature related to Coreference Resolution and the use of semantics for the task.

3.1 Coreference Resolution

Coreference resolution approaches can be divided into two main groups. The earliest works, which date back from the late seventies, adopted linguistic approaches based on parsed tree and rhetorical structure. In the mid-nineties, a shift towards a more corpus-based approach began, alongside the adoption of machine learning strategies.

3.1.1 Parsed Tree and Rhetorical Structure

The Hobbs algorithm [26] was one of the earliest approaches to resolve anaphoric pronouns. The algorithm starts in the pronoun's node of the syntactic tree of a sentence, doing a breadth-first left-to-right search which went through several conditional choices whenever a noun phrase iss found. If an antecedent is not found in the same sentence, then the search continues in the document's previous sentence. This algorithm achieved performance about 90% when searching for antecedents for anaphoric pronouns, indicating 90% of pronouns in a document can be resolved by solely using morphological and syntactic information.

Centering Theory [21] tries to model the relationships by using a focus of attention. It is based on the idea that the speaker or writer wants to keep a main entity, the center, in focus. During discourse, the center usually shifts, but that does not tend to occur when using anaphoric pronouns. When the center changes, it has to be easily perceived by the reader or listener (for instance, using a noun phrase). Its principles have been widely used directly or indirectly in much later work [11].

Another theory is the Discourse Representation Theory [29]. It proposes that each sentence is represented in a Discourse Representation Structure (DRS). A DRS is a semantic representation of the sentence, which can be translated into a first-order logic structure.

3.1.2 Corpus-based and Supervised Learning Approaches

Most corpus-based approaches appeared after MUC-6 [20], in the mid-nineties, when the first annotated coreference documents were published.

Ge et al. [19] tried to resolve pronominal resolution using a statistical approach. Cardie and Wagstaff [6] developed an algorithm that views coreference resolution as a clustering task. Luo et al. [35] proposed a pairwise classification based on a Bell Tree.

Harabagiu et al. [24] presented a knowledge-minimalist methodology of mining coreference rules from annotated corpora.

Conditional Random Fields [37] [69] were applied to the task. However, enumerating all possible configurations in order to find the most probable one can result in intractable combinatorial growth [65].

The first machine learning systems developed for coreference resolution were based on decision trees [65]. The coreference resolution problem is cast then as a pairwise classification problem, in which the question is whether two markables corefer or not. Soon et al [70] proposed a system based on a decision tree using 12 features. They noticed that three features (STRING_MATCH, ALIAS and APPOSITIVE) were significantly more informative than the others. The decision tree learned using only these three features was just 2% worse than the one with all 12 features.

Haghini and Klein [23] developed a rule-based, deterministic system, based on hierarchical sieves, known as (H and K model). Standford CoreNLP's deterministic system, described by Lee et al. [31][30], is a product of extensive investigations conducted on deciding the precise rules to govern the task of Coreference Resolution [73].

Garcia and Gamallo presented a rule-based, entity-centric system for Portuguese, Spanish and Galician [18]. The entity-centric approach determines if a mention belongs to an entity by comparing it to the existing entity mentions, differently from a mention-pair approach. The system uses the multi-pass sieve architecture proposed by Lee et al. [31] and it can only resolve mentions for mentions of the "person" category.

3.1.3 Unsupervised Learning

Since the lack of annotated corpora is a major obstacle for supervised learning approaches, some researches experimented with unsupervised learning. Bean and Riloff [5] developed a system that learns relations between words and the different contexts in which they can appear in an unsupervised manner. They represented context by a case frame. They found that contextual role knowledge was more beneficial for pronouns than for noun phrases.

Muller et al. [44] investigated a co-training approach to build a classifier for German texts, but obtained mostly negative results. Haghighi and Klein [22] presented a nonparametric Bayesian approach.

Unsupervised learning presents some limitations, as described by Pierce and Cardie [53], and their performance is not comparable to that of a fully supervised coreference resolver [47].

3.1.4 Deep Learning models

More recently, deep learning models have been applied to coreference resolution. Some works include Wiseman et al. [76] and Clark et al. [7]. In 2017 Lee et al. [32] presented the first end-to-end coreference resolution model, which outperformed all previous work.

3.2 Corefence Resolution and Semantics

The coreference system presented by Harabagiu et al. [24] verifies semantic consistency between nouns using information retrieved from WordNet. The resource was used to obtain, among other relationships, synonyms, hypernyms and meronyms.

The model by Soon et al. [70] included a semantic class feature, which obtains information from WordNet. Strube and Ponzetto [54] [55], based on their work, used Wikipedia and WordNet, at different times, to disambiguate mentions and compare the results obtained with each base, showing that Wikipedia is competitive with WordNet for the coreference resolution task. They also noticed that, although WordNet and Wikipedia can increase performance on common nouns, neither affects performance for proper nouns, where rules such as string matching and alias suffice.

Ji et al. [27] added semantic relations to refine decisions taken by a pair classifier. The classifier uses a set of hand-written rules and maximum entropy models to determine coreferential pairs. Once classification is done the system searches for semantic relations between mentions that seem to be non-coreferential.

Ng et al. [46] introduced new linguistic features in their learning-based coreference resolution system, including semantic class agreement, semantic ACE class, and semantic similarity, in which they achieved limited gains. For semantic similarity, they used a dependency-based thesaurus, which was constructed using a distributional approach.

Rahman and others [59] evaluated world knowledge using Yago [72] and FrameNet [4]. They used relations such as "Means" and "Type". The "Means" relation is analog to synonymy, and provides a way to disambiguate, for instance, between `(Einstein, Means, AlbertEinstein)` and `(Einstein, Means, AlfredEinstein)`, since Einstein could refer to the physicist or the musician. From FrameNet they used verb's semantic roles. Each relation is represented by a triple `(AlbertEinstein, Type, physicist)`. They concluded that semantics can provide small gains to the coreference resolution task that, globally, could become substantial.

Recasens et al. [60] proposed an unsupervised method to resolve coreferent mentions with very different words (opaque mentions). They built a dictionary of opaque coreferent mentions and integrated it into the Stanford coreference resolution system. To build the dictionary they used a comparable corpus of tech news, calculating the distributional similarity in a set of articles that discussed the same event. Their results showed about 1% F-measure increase for all metrics.

Prokofyev et al. [58] improved the Stanford Coref[1] pipeline adding a semantic layer which tries to associate each mention with a DBpedia entity. They used YAGO to associate the mentions with semantic types. They were able to improve precision, recall and F-measure. Word embedding models were experimented with, but ultimately lead to worse overall results.

Simova and Uszkoreit [68] researched ways to use word embedding models as features in a system based on supervised learning. They created three types of features based on word embeddings: Embedding Cluster, Dense Embedding Features and Cosine Similarity Features. For the Embedding Cluster, they performed clustering of word embeddings and assigned a cluster label to each head word. Dense Embedding Features were created by embedding the head word vector directly, creating a numeric feature for each vector dimension. Cosine Similarity features are based on the cosine similarity between the mention's head and governing words. They noticed that overall the results improved with their features, although 12% to 36% of their results were below the baseline, depending on the feature set. They also achieved better results by merging the embedding models than by using just one model, which indicates that the knowledge encoded by these vectors could be complementary to each other. They also noticed a decrease in precision for situations involving named entities.

[1] https://nlp.stanford.edu/projects/coref.shtml

3.2.1 Semantics for coreference in Portuguese Language

Silva [9] proposed a coreference resolution model using semantic tags from the HAREM corpus [17]. As a knowledge base, TeP 2.0 – Electronic Thesaurus for Brazilian Portuguese [36] was used. It stores synonymous and antonymous word forms.

Coreixas [43] evaluated and proposed coreference resolution methods for Portuguese, focusing on named entities and using semantic categories. The author showed that the use of semantic categories can improve the results, and highlights the necessity to use more structured semantic bases such as WordNet.

Evandro et al. [16] evaluated the impact of semantic knowledge on coreference resolution for Portuguese. This work was implemented using the tool that will be used for our work, CORP, which is further explained in section 4.3.

3.3 Summary

This chapter presented the main proposed models for coreference resolution found in the literature. The tool used for our experiments is based on the rule-based model proposed by Lee et al. [30]. An advantage of this model is that it does not need a large corpus for training which, unfortunately, is not available for the Portuguese language yet.

Regarding the works using semantics, there are works using rule-based or machine learning models. Some efforts tried to resolve semantic links using semantic class agreement, word embedding models, WordNet, DBPedia, YAGO and FrameNet. For the Portuguese language, there are works using thesaurus, semantic class agreement and OntoPT. These works obtained mixed results or small gains, which shows the difficulties involved in using semantics for coreference resolution.

Most works which explored semantics focused in a specific resource. Our work compared different resources, analysing the impact of each one. The resources analysed in our work were generated from various sources, including other semantic bases. Nevertheless, our work focused on the Portuguese language.

4. RESOURCES

This section will give an overview of the resources used for this work's experiments. These include semantic bases for the Portuguese language, the coreference resolution system and the metric used to evaluate the models.

4.1 Semantic Bases for Portuguese

We examined the available semantic resources for Portuguese. Some resources were discarded because they were discontinued (TeP[1], Port4Nooj[2], PAPEL[3]), not available for download (WordNet.PT[4]), seem unfinished (WordNet.BR[5]), are a paid resource (MWN.PT[6]) or are in a preliminary state (Ufes WordNet[7]). Further details about the semantic bases available for Portuguese can be found in Santos et al. [64], Oliveira et al. [51] and Oliveira [50].

Many resources listed here are based on WordNet [42]. Its basic structure is the Synset, which represents a word sense formed by a set of synonyms. Synonymy is considered, therefore, a symmetric relation between word forms. The word "house" contains, among others, these different senses, each one being a synset:

- A dwelling that serves as living quarters for one or more families;
- An official assembly having legislative powers;
- Aristocratic family line.

Synsets are linked by different relations, including hypernymy, antonymy and meronymy. Table 4.1 shows a synset containing the word "car" and it's hypernym synset, which contains the word "motor vehicle".

Table 4.1 – Hypernym relation between synsets

Synset	Words	Description
02961779	car, auto, automobile, machine, motorcar	a motor vehicle with four wheels ...
03796768	motor vehicle, automotive vehicle	a self-propelled wheeled vehicle ...

[1] http://www.nilc.icmc.usp.br/tep2/
[2] http://www.linguateca.pt/Repositorio/Port4NooJ/
[3] http://www.linguateca.pt/PAPEL/
[4] http://www.clul.ulisboa.pt/clg/wordnetpt/index.html
[5] http://www.nilc.icmc.usp.br/wordnetbr/
[6] http://mwnpt.di.fc.ul.pt/features.html
[7] https://sites.google.com/site/ufeswordnet/

4.1.1 OpenWordNet-PT

OpenWordNet-PT is a syntatic projection from the Universal WordNet being developed since 2010. It is based on synsets and strongly coupled with the Princeton WordNet. The project is still active and was created by Valeria de Paiva, Alexandre Rademaker and Gerard de Melo [10]. They used machine learning to build relations among graphs which represent information from multilingual Wikipedia versions and open electronic dictionaries. It is freely available in RDF/OWL format.

4.1.2 Portuguese Unified Lexical Ontology (PULO)

PULO aims to be a free WordNet for the Portuguese language, aligned with the Multilingual Central Repository (MCR) project. Work began on 2014 by Simões and Guinovart [67] with the translation of the English, Spanish and Galician WordNets. It is available online and has the same ontological structure as the Princeton WordNet. It is freely available in the SQL format.

4.1.3 OntoPT

OntoPT is a lexical ontology for Portuguese [52]. It was built automatically using Portuguese textual resources such as OpenThesaurus.PT, Wikicionário.PT, Dicionário Aberto, TeP, OpenWordNet-PT and the PAPEL project. Its structure is based on synsets, similar to WordNet, although it is not aligned with it. It possesses several relevant relations, such as hyperonymy, hyponymy, synonymy and meronymy.

4.1.4 ContoPT

ContoPT is the continuation of the work done on OntoPT [49]. Fuzzy synsets were generated automatically from seven Portuguese lexical-semantic resources: PAPEL, Dicionário Aberto, Wikitionary.PT, TeP 2.0, OpenThesaurus.PT, OpenWordNet-PT and PULO. It contains a confidence measure, which is calculated based on the redundancy between these resources.

Authors performed experiments with Portuguese native speakers to evaluate the fuzzy synsets correctness. Experiments were done using different confidence cut-points, where they perceived a correlation between the confidence measure and correctness.

4.1.5 ConceptNet

ConceptNet [71] is a product of Open Mind Common Sense, an MIT Media Lab crowdsourcing project from 1999. Currently, it is maintained by Luminoso Techonologies[8]. It has support for hundreds of languages from different families. The base uses various data sources, ranging from DBpedia, Wiktionary, Open Multilingual WordNet, UMBEL, among others. The sources are combined with word embedding models such as Word2Vec. The resource is freely available.

Information is presented as a graph in which edges connect two terms. Each object contains a URI, an example of this being: `/c/en/house` is the word "house" in English. Each edge contains a relation, which can be of many types. Table 4.2 shows some relation types.

Table 4.2 – ConceptNet Relations

Relation URI	**Description**	**Example**
/r/RelatedTo	The most general relation. There is some relation between A and B.	Learn <-> erudition
/r/IsA	A is a subtype or a specific instance of B. WordNet hyponym relation.	Car -> vehicle Chicago -> city
/r/PartOf	A is a part of B. WordNet meronym relation.	Gearshift -> car
/r/HasA	B belongs to A	Bird -> wing
/r/Synonym	A and B have very similar meanings. They may be translations.	Sunlight -> sunshine
/r/Antonym	A and B are opposites in some relevant way.	Hot <-> cold

4.1.6 FrameNet Brasil

FrameNet Brasil is a lexical database maintained by Juiz de Fora Federal University[63]. The project started on 2007 as a Portuguese variation of the original FrameNet project being developed by Berkley University[4].

FrameNet is based on Frame Semantics theory, derived from the work of Charles and colleagues [12]. This theory claims that the meaning of most words can be best understood on the basis of a semantic frame. A semantic frame is a description of a type of event, relation, entity and the participants in it. A frame is composed of frame elements, and words that evoke the frame are called lexical units.

4.1.7 BabelNet

BabelNet [45] is a multilingual dictionary and semantic network. It has about 14 million entries, called Babel synsets, each one representing a given meaning and containing

[8] https://luminoso.com/

all synonyms which express that meaning in different languages. It was built automatically from the following sources: Wordnet, Wikipedia, Wikitionary, Wikidata, VerbNet, GeoName, ImageNet, FrameNet, among others. It has a SPARQL endpoint and an online tool, but the endpoint requests are limited per user. It is possible to download the entire index for research purposes, but that requires approval from the BabelNet staff. Knowledge is represented in a graph format, similarly to WordNet. Given a specific synset and a relation type, it is possible to obtain the set of related synsets.

4.1.8 DBpedia

DBpedia [2] is a crowdsourcing project which aims to extract structured knowledge from Wikimedia projects, Wikipedia being the most popular one, and make it freely available on the Web. The project began in 2007 and is still active. The project extracts information in more than 110 languages, including Portuguese.

An open-source extractor[9] is used for extracting information from Wikimedia dumps. The extracted information is stored in RDF statements. Wikimedia information structures, such as infoboxes, are mapped to an ontology. This ontology unifies these template structures across different languages. DBpedia dumps are freely available on their site.

They also created NLP Datasets [38] to support entity recognition and disambiguation tasks. The Lexicalization Data Set contains mappings between surface forms (for instance, names and nicknames of people) and URIs, therefore providing access to alternative names for entities and concepts. Page titles can be seen as community-approved surface forms, redirects can indicate synonyms or acronyms. Disambiguation links can resolve ambiguous surface forms. It also calculates the co-occurrence between entities and their name variations using Anchor Texts, which generates a strength value between the surface form and URI. The Topic Signatures Data Set extracts paragraphs that contain links to DBpedia resources and aggregates them in a vector space model of terms, weighted by their co-occurrence with the target entity. The Grammatical Gender Data Set can be used to verify the gender of a person.

DBpedia Spotlight [39] is a tool for automatically annotating text documents with DBpedia URIs. It is freely available as a web service, which can be downloaded to be executed in a specific machine.

[9]https://github.com/dbpedia/extraction-framework

4.1.9 NILC Word Embeddings

Aside from the structured semantic bases presented so far, there are approaches based on distributed word representations. Although these representations have been studied for a long time, with early works dating back to the eighties [41], neural network based approaches have seen a recent resurgence and obtained outstanding results in terms of word-prediction [41].

The Word2Vec model, composed of its Skip-Gram and CBOW variants, was proposed by Mikolov et al. [40]. It differs from the previous work because it is less computationally expensive, making it possible to learn vectors from huge data sets with billions of words containing millions of words in the vocabulary. Previous architectures trained a vocabulary composed of a few hundred million words with 50 to 100 dimensions.

Models trained for the Portuguese language are available at the NILC Word Embeddings Repository[10]. The vectors were generated using 17 corpora composed of Brazilian and European Portuguese texts from different genres. Vectors were trained using Word2vec, FastText, Wang2vec and Glove algorithms[25].

4.1.10 Semantic Resources Considered in This Work

In this work we considered for experimental analysis the following semantic resources: ConceptNet, ContoPT and NILC Word Embeddings.

We decided to not use PULO, described in section 4.1.2, because the project seems to have been discontinued. FrameNet presents a different paradigm and would thus require a lot more effort to integrate into CORP. OpenWordNet-PT was not used directly because OntoPT, ContoPT and ConceptNet already use it as a base resource.

From the DBpedia projects, DBpedia Spotlight, a tool for entity recognition which uses DBpedia datasets, is the most interesting resource for our task. This tool associate mentions in a text to DBpedia resources. If the tool associates two or more mentions with the same DBpedia resource, they may be coreferent. We noticed from our investigations that, although many mentions were correctly labelled by the tool, just a few of them were associated with the same resource (or any of them, depending on the text). DBpedia is also used by ConceptNet as a base resource.

BabelNet is a very interesting and extensive resource. We implemented a proof of concept for it. The performance was slower when compared to the other resources, probably due to its large amount of data. Since CORP can require hundreds of queries

[10] http://nilc.icmc.usp.br/embeddings

for a single text, that problem could escalate easily. Considering that some indexes, such as image indexes, are not relevant for our task, we could discard them in order to improve performance, but since the resource dumps are strongly coupled with the Lucene Framework (for the indexes) and BabelNet API, that task would require more time and effort which were not possible for this work.

4.2 Corpora

4.2.1 Corref-PT

Corref-PT is a Portuguese corpus annotated with coreference information. It was created as a collective effort during IBEREVAL-2017[11] (Evaluation of Human Language Technologies for Iberian Languages), in which PUCRS's Natural Language Processing Group proposed the task "Collective Elaboration of a Coreference Annotated Corpus for Portuguese Texts" [15], with the objective to collectively create a corpus annotated for coreference for Portuguese. To do so, each team presented texts which they considered interesting to annotate. The task involved twenty one annotators from seven teams, all native Portuguese speakers. The corpus is composed of journalistic texts, books, magazine articles, Wikipedia articles, among others. Table 4.3 presents some corpus details.

Table 4.3 – Corref-PT - Corpus Statistics

Corpus	Texts	Tokens	Mentions	Coreferent Mentions	Coreference Chains	Biggest Chain	Avg. Chain Size($)
CST-News	137	54445	14680	6797	1906	25	3.6
Le-Parole	12	21607	5773	2202	573	38	3.8
Wikipedia	30	44153	12049	4973	1308	53	3.8
Fapesp Magazine	3	3535	1012	496	111	33	4.5
Total	182	123740	33514	14468	3898	53	3.7

Corref-PT is available in four formats: TXT, XML, HTML and SemEval [62]. The resource is free and can be obtained from PUCRS's Natural Processing Language group's page[12].

4.2.2 Summ-it++

Summ-it++ [1] is an evolution of the Summ-it corpus [8]. The new version contains the coreference annotation in the SemEval format, along with two new semantic annotation

[11] http://sepln2017.um.es/ibereval.html
[12] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/corref-pt/

layers. The corpus consists of fifty journalistic texts from the Folha de São Paulo newspaper's Science section. Summ-it++ contains 560 coreference chains, with an average of three noun phrases per each chain, the largest one being composed of 16 noun phrases. This resource is available at [13].

4.3 CORP

CORP (Coreference Resolution for Portuguese) [14] is a coreference resolution tool for Portuguese. It was implemented based on the sieve architecture proposed by Lee et al. [30], uses the CoGrOO [66] API to extract noun phrases from the text and removes mentions that begin with numerical entities as the percent, money, cardinals or quantifiers during pre-processing. It then generates noun phrase pairs and extracts features from them. The rules for each feature are listed in Table 4.4, and are further explained in [13]. The feature extraction task generates a vector, which is submitted to a classifier that decides whether the noun phrases are coreferent or not. Finally, the coreference chains are generated from the classified pairs. CORP is only trained to resolve identity type nominal coreference (proper and common nouns), independently of semantic category or domain. It does not treat personal pronouns. Currently, CORP has two rules regarding semantic knowledge: Synonymy and Hyponymy. It uses OntoPT to extract these semantic features. Fonseca [13] performed experiments to evaluate the semantic relations' impact. He observed that the addition of semantic rules increased recall and F-measure, but reduced the precision of most models. In the following sections, we will explain how these relations are being used.

Table 4.4 – CORP Rules

Rule	Description	Example
Exact Match	Both mentions are equal	[the Japan], [the Japan]
Partial Matching	Both mentions' heads are equal	[the supposed scheme], [scheme]
Appositive	Appositive construction	[Defense minister], [Nelson Jobim]
Appositive Role	Appositive role construction	[Heloísa Helena], [candidate for presidency]
Acronym	A mention is an acronym of the other	[the European Union], [EU]
Predicate Nominative	m_y completes a linking verb and renames m_x	[France] is [the unique country]
Relative Pronoun	m_y is a relative pronoun of m_x	[Spaghetti], [which] many of us enjoy
Strict Head Match	Any of their head words match the same restrictions of Lee et al. [30]	[the second road] [the road]
Proper Noun Matching	Mentions must be proper nouns, some words are equal and are not embedded	[European Union] [Union]
Partial Proper Noun Matching	m_y must be equal to any word in m_x	[Heloísa Helena] [Heloísa]
Hyponymy	Hyponymy relation between m_x and m_y	[the dog], [the animal]
Synonymy	Synonymy relation between m_x and m_y	[the house], [the residence]

CORP's outputs a XML and, to make visualization easier, an HTML file for each text. Figure 4.1 shows an example of a XML file, and Figure 4.2 shows an an example of an HTML file. The XML file's contents are:

- The original text;
- Identified sentences;

[13] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/summ-it/

- Token list with POS markup;
- Coreference chains;
- Unique mentions list. An mention is called unique when it doesn't have any referent.

```
<Cadeia_161>
  <sn id="11" tokens="47...50" nucleo="torneio" sintagma="o torneio mais antigo" Categoria="OUTRO" sentenca="5">
    <word_47 token="o" lemma="o" pos="art" features="M=S" ></word_47>
    <word_48 token="torneio" lemma="torneio" pos="n" features="M=S" ></word_48>
    <word_49 token="mais" lemma="mais" pos="adv" features="" ></word_49>
    <word_50 token="antigo" lemma="antigo" pos="adj" features="M=S" ></word_50>
  </sn>
  <sn id="14" tokens="68...70" nucleo="política" sintagma="a política local" Categoria="ABSTRAÇÃO/DISC" sentenca="5">
    <word_68 token="a" lemma="o" pos="art" features="F=S" ></word_68>
    <word_69 token="política" lemma="política, político" pos="n" features="F=S" ></word_69>
    <word_70 token="local" lemma="local" pos="adj" features="F=S" ></word_70>
  </sn>
  <sn id="21" tokens="93...95" nucleo="competição" sintagma="a competição continental" Categoria="OUTRO" sentenca="5">
    <word_93 token="a" lemma="o" pos="art" features="F=S" ></word_93>
    <word_94 token="competição" lemma="competição, competir" pos="n" features="F=S" ></word_94>
    <word_95 token="continental" lemma="continental" pos="adj" features="F=S" ></word_95>
  </sn>
  <sn id="26" tokens="126...127" nucleo="disputa" sintagma="a disputa" Categoria="OUTRO" sentenca="7">
    <word_126 token="a" lemma="o" pos="art" features="F=S" ></word_126>
    <word_127 token="disputa" lemma="disputa" pos="n" features="F=S" ></word_127>
  </sn>
  <sn id="55" tokens="227...230" nucleo="confronto" sintagma="o confronto mais tradicional" Categoria="OUTRO" sentenca="13">
    <word_227 token="o" lemma="o" pos="art" features="M=S" ></word_227>
    <word_228 token="confronto" lemma="confronto" pos="n" features="M=S" ></word_228>
    <word_229 token="mais" lemma="mais" pos="adv" features="" ></word_229>
    <word_230 token="tradicional" lemma="tradicional" pos="adj" features="M=S" ></word_230>
  </sn>
```

Figure 4.1 – CORP - XML output

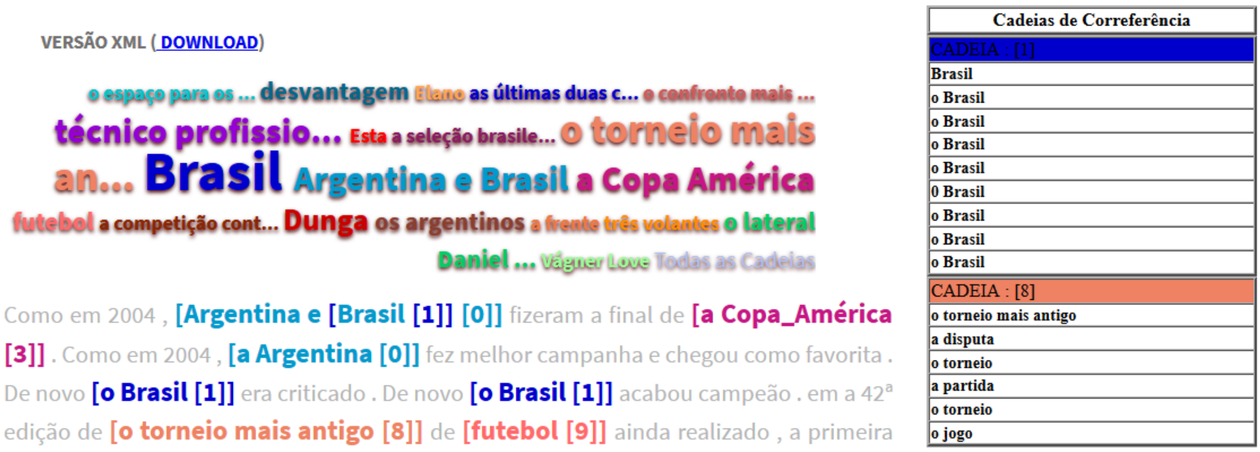

Figure 4.2 – CORP - HTML output

It is also possible to convert the output to the SemEval format. SemEval is the input format for the tool that calculates CoNLL scores, which will be described in section 4.4. CORP is not restricted to a specific domain. It can be used with any text written in Portuguese, as long as it is coherent and well-written.

4.3.1 Hyponymy Rule

Two mentions will be grouped if a hyponymy relation exists between a referent and it's antecedent. Hyperonymy relations are not considered since it is more common to introduce an entity in a specific form and, in following mentions, use more generic terms for it. There is a clause, called Modifying Word, that restricts the order of pronouns. This clause was created because it is not common, after a definite pronoun, to use an indefinite one. For instance, is not common to use "the car" and, after it, "a car". They must also agree in number. Therefore, considering two mentions m_i and m_j, the following conditions must be satisfied if they are to be grouped together:

- Core mention lemma from m_i and m_j must have a hyponymy relation;
- If m_i has a definite pronoun, m_j can't have a indefinite pronoun;
- m_i and m_j must agree in number (singular/plural);
- Other words that modify the mentions cannot occur (Modifying Word clause).

4.3.2 Synonymy Rule

Synonymy relations are considered when the following conditions are satisfied:

- Core mention lemma from m_i and m_j must have an synonymy relation;
- If m_i has a definite pronoun, m_j can't have a indefinite pronoun;
- m_i and m_j must agree in number (singular/plural);
- Other words that modify the mentions cannot occur;
- Each mention in a coreference chain must have a synonymy relation with all other mentions in the chain.

4.4 CoNLL Reference Scorer

CoNLL Scorer is an API to evaluate coreference models. It was developed for the CoNLL 2011 Shared Task [56], as a way of provide an automatic and fair evaluation, since each metric favors a specific model, as can be seen in [56]. CoNLL Average is an average between respective F-Measures from MUC [75], B^3 [3] and $CEAF_e$ [34]. The API output is composed of the following metrics: MUC, B^3, $Ceaf_e$, $Ceaf_m$ and BLANC [61].

Its input is composed of two files, both of them in the SemEval format, which is a well-known format and used for most corpora. The first file contains the reference annotations, and the second contains the annotations provided automatically by the model being evaluated.

5. DEVELOPMENT

In this section, the implementations created to improve CORP and to perform the experiments are described.

5.1 CONLL Scorer Script

The CoNLL Reference Scorer script, first presented in section 4.4, was written in Perl and is available on GitHub[1]. To avoid having to manually run the script after each experiment we embedded the script into CORP.

If a SemEval file is included in the CORP input folder, CORP will execute the script automatically and generate a text file containing the script results for MUC, B^3, $Ceaf_e$ and the CoNLL average.

5.2 Semantic Bases Integration

CORP was already using Onto.PT for synonym and hyperonym relationships. The data extracted from this resource includes 79425 hyperonyms and 156566 synonyms. We decided to integrate CORP with ConceptNet and ContoPT.

5.2.1 ContoPT

We ran a script on the ContoPT dump, filtering for synsets of the type name and relations of the hypernym type. The script resulted in 21517 synsets and 258194 hypernym relations.

ContoPT's structure is based on synsets. Each word in a synset contains a confidence measure which indicates how much this word is related to the concept. Also, each relationship is associated with a confidence measure. We decided to use those confidence measures to avoid unlikely relationships that could generate false-positives. To use those confidence measures, two cut-off point values were necessary, one for synonyms and one for hypernyms. These cut-off points can be set in the parameters file.

[1] https://github.com/conll/reference-coreference-scorers/

For the synonym rule, considering two candidate mentions, CORP searches for a synset which contains both mentions. If both mentions' confidence measures are above the cut-off point, they will be considered synonyms.

For the hypernym rule, CORP searches for synsets containing the first candidate, which will compose set S_1, and for synsets containing the second candidate, which will compose set S_2. If there is a relationship R between any synset in S_1 and any synset in S_2, and R's confidence measure is above the cut-off point, the hypernym rule will be activated.

5.2.2 ConceptNet

To integrate ConceptNet, we performed first some data preparation. In the ConceptNet dump file, there were 791502 records, from various languages. All records that were not a Portuguese-Portuguese relation were deleted, leaving 280915 records. From the remaining records, 3757 were hyponymy relations and 13727 were synonymy relations.

Each relation contains a weight, but it was not used since most did not seem to be accurate, with default values such as 1 or 0.

5.3 Word Embeddings Web API

We decided to do some experiments using Word Embedding models as a semantic resource. With these experiments, we aim to verify the usefulness of the simple application of semantic similarity calculus to our task.

To use word embeddings in CORP, we created a semantic similarity rule. The rule weight is 0.177, the same as the synonym rule. To define a pair of mentions as semantically similar, the semantic similarity between the head of both mentions must be above a specified threshold characterized in the parameters file.

To calculate the semantic similarity between two mentions we used the Gensim[2] library. Since CORP was developed in Java and Gensim is a Python library, we decided to implement a semantic similarity Web API[3]. CORP uses the API to acquire the semantic similarity between all mention pairs generated by CORP for a specific text. The Web API architecture avoided some architectural obstacles that could arise, such as interoperability issues between Java and Python, concurrence problems (CORP is multi-thread) and performance problems. The 300 dimension models we used take about three minutes to load into memory, and consume about 3GB of memory after loading. The Web API can be run in another server if necessary.

[2]https://radimrehurek.com/gensim/

[3]https://github.com/tmlima/SemanticSimilarity

When processing the semantic similarity rule, CORP sends a JSON file containing the candidate mentions to the API. The API calculates the similarity between each pair and returns the JSON updated with a semantic similarity result for each candidate pair. When CORP receives the response, it selects all mentions with a semantic similarity score above the defined threshold and marks them as semantically similar.

5.4 Execution Parameters

Execution parameters are defined in a JSON file in the CORP root folder. These parameters are the semantic base, specific semantic base parameters such as ContoPT's cut-off points and hypernym window. When using semantic similarity, the user can also set the threshold and the Web API endpoint.

For each execution, a parameters file is generated in the output folder. This file contains the provided parameters and CORP version used. This was developed to easily debug the experiments.

6. EXPERIMENTS AND RESULTS

This chapter presents the experiments performed using different semantic resources. The Baseline is defined as CORP without any semantic resource. The first experiments evaluated the synonymy and hypernymy relationships contained in ConceptNet, ContoPT and OntoPT.

Since the hypernym rule only considers relationships within a specific window, we performed experiments with different window sizes in order to verify how results might be impacted.

Experiments using ContoPT were performed considering 1 as the hypernym cut-off point and 0.75 as the synonym cut-off point. These values were based on experiments performed by Oliveira [49]. It was found that using a 0.75 cut-off point, 87% of all synonym nouns pair were labelled correctly. Cut-off points greater than 0.75 resulted in a small gain in accuracy when considering the quantity of synsets above the cut-off point. Using a cut-off point of 1.0 for relationships between synsets (not restricted to hypernyms) resulted in 72.1% of all sysnsets being labelled correctly. Experiments using a relationship cut-off point greater than 1.0 resulted in a steep decrease in the number of synsets.

To highlight the semantic resources' impact, experiments using a semantic subset were performed. This semantic subset is composed of three texts, for each corpus, which contain more semantic relationships than the corpus average.

Finally, we performed experiments using the semantic similarity rule, described in section 5.3. The results for these will also be discussed below.

Experiment results, CORP version, corpora, and SemEval files are available at GitHub[1]. The repository includes also instructions on how to reproduce the experiments.

6.1 Evaluating Semantic Bases

Tables 6.1 and 6.2 show the results obtained using semantic bases for the Corref-PT and Summ-it++ corpora. The baseline, as expected, obtained the lowest recall and highest precision. OntoPT obtained, for both corpora, the highest recall in all metrics, but at the same time the largest loss in precision, resulting in the lowest CoNLL average. Although ConceptNet obtained the highest CoNLL average after the baseline, it is important to note the small size of that base compared to the others, and how that produced recall results noticeably similar to the baseline ones. Unfortunately, the addition of semantic bases were unable to achieve CoNLL average scores comparable to those obtained by the baseline, althought it is important to notice that all bases increased the recall.

[1] https://github.com/tmlima/masters_dissertation

The experiment using ContoPT for the Summ-it++ corpus obtained a higher F1 for MUC than the baseline. MUC, for our purposes, can be considered the most relevant metric, since it is based on links between mentions, which is exactly what we are trying to find using these semantic resources.

Table 6.1 – Semantic bases - Corref-PT

Base	MUC			B_3			$CEAF_e$			CoNNL Avg
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
Baseline	49.40	**63.42**	**55.54**	41.71	**60.04**	**49.22**	51.27	**50.35**	**50.81**	**51.857**
OntoPT	**51.69**	54.67	53.14	**44.76**	51.09	47.71	**52.96**	46.25	49.38	50.077
ContoPT	50.03	61.78	55.29	42.45	58.25	49.11	51.84	49.28	50.53	51.643
ConceptNet	49.69	62.55	55.39	42.09	59.17	49.19	51.63	49.95	50.77	51.783

Table 6.2 – Semantic bases - Summ-it++

Base	MUC			B_3			$CEAF_e$			CoNNL Avg
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
Baseline	41.79	**53.61**	46.97	39.42	**54.07**	**45.60**	51.32	**50.55**	50.93	**47.833**
OntoPT	**43.60**	45.71	44.63	**41.56**	45.77	43.56	**52.47**	44.63	48.23	45.473
ContoPT	42.44	53.02	**47.14**	39.89	53.21	45.59	51.27	49.75	50.50	47.743
ConceptNet	41.93	52.91	46.79	39.69	53.42	45.54	51.64	50.33	**50.98**	47.770

6.2 Hypernym Window

The existing hypernym rule in CORP restricted the relationship for cases in which both mentions were in adjacent sentences. We did some experiments using different window sizes and without that restriction. A window size of 3 means that mentions will be linked by the hypernym rule only if both mentions are in a range of 3 sentences. Tables 6.3 and 6.4 show the results obtained.

A decrease in the recall is noticeable when not using the hypernym window, probably due to the incorrect identification of relationships between distant mentions. The best results obtained using Corref-PT were with a window size of 1 and, while for Summ-it++ they were obtained using a window size of 3. That can be explained by differences between both corpora. Summ-it++ is composed of short and concise journalistic texts. CorrefPT, on the other hand, is composed of texts from different domains, some of them very long.

The hypernym window can be a helpful feature, but at the same time, some straightforward relationships may not be found because of that, and the usefulness of this restriction will depend on the text.

Table 6.3 – Hypernym Window Sizes - CorrefPT

Base	Window	MUC			B_3			$CEAF_e$			CoNNL Avg
		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
OntoPT	1	51.69	54.67	53.14	44.76	51.09	47.71	**52.96**	46.25	49.38	50.077
ContoPT	1	50.03	61.78	55.29	42.45	58.25	49.11	51.84	49.28	50.53	51.643
ConceptNet	1	49.69	62.55	**55.39**	42.09	59.17	**49.19**	51.63	49.95	**50.77**	**51.783**
OntoPT	3	**51.74**	53.62	52.66	**44.86**	49.97	47.27	52.77	45.80	49.04	49.657
ContoPT	3	50.04	61.76	55.29	42.46	58.20	49.1	51.81	49.24	50.49	51.627
ConceptNet	3	49.71	62.52	55.38	42.11	59.12	49.18	51.65	49.92	**50.77**	51.777
OntoPT	-	51.44	55.68	53.48	44.38	52.10	47.93	**52.96**	46.52	49.54	50.317
ContoPT	-	50.04	61.82	55.31	42.46	58.29	49.13	51.87	49.30	50.55	51.663
ConceptNet	-	49.60	**62.63**	55.36	41.98	**59.27**	49.15	51.59	**49.97**	**50.77**	51.760

Table 6.4 – Hypernym Window Sizes - Summ-it++

Base	Window	MUC			B_3			$CEAF_e$			CoNNL Avg
		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
OntoPT	1	43.60	45.71	44.63	41.56	45.77	43.56	**52.47**	44.63	48.23	45.473
ContoPT	1	42.44	**53.02**	**47.14**	39.89	53.21	45.59	51.27	49.75	50.50	47.743
ConceptNet	1	41.93	52.91	46.79	39.69	53.42	45.54	51.64	50.33	50.98	47.770
OntoPT	3	**43.67**	44.37	44.02	**41.73**	44.34	43.00	52.18	43.81	47.63	44.883
ContoPT	3	42.44	**53.02**	**47.14**	39.89	53.24	**45.60**	51.27	49.82	50.54	47.760
ConceptNet	3	42.01	52.81	46.79	39.79	53.32	45.57	51.75	**50.36**	**51.04**	**47.800**
OntoPT	-	43.38	47.43	45.31	41.11	47.35	44.01	52.05	45.21	48.39	45.903
ContoPT	-	42.44	**53.02**	**47.14**	39.89	53.24	45.60	51.27	49.82	50.54	47.760
ConceptNet	-	41.86	52.92	46.75	39.65	**53.46**	45.53	51.64	50.26	50.94	47.740

6.3 Semantic Subset Analysis

Since the results differed little regardless of the semantic resource used, we decided to select a small subset of each corpus that contained texts with more semantic relationships than the average, so that we further analyze the effect of the semantic resources where they would be theoretically most impactful. To select those texts, we implemented two scripts, one for each corpus.

Correft-PT and Summ-it++ keep their reference chains in XML files, but the information in them is not exactly the same. Summ-it++ reference chains contain information about whether or not a mention is indirect, therefore we implemented a script that selects texts with the most indirect mentions. For CorrefPT that information was not available, so we implemented a script that obtains the texts wherein most mention heads are distinct from one another. Table 6.5 shows the total quantity of mention and indirect mentions found in Summ-it++, as well as these quantities for the three texts with the most indirect mentions. Table 6.6 shows the total mentions for CorrefPT and the quantity of heads found which were distinct from the other heads in the chain, as well as these quantities for the three texts with the most distinct heads.

The subset experiments were performed using the same configuration as the experiments presented in section 6.1. Tables 6.7 and 6.8 show the results obtained. Overall results were lower with both subsets when compared to the results with the entire corpus, which is understandable considering that the subset is composed of texts with more complex

Table 6.5 – Corpus Relationships - Summ-it++

	Mentions	Indirect Mentions
Corpus	1982	330
CIENCIA_2000_17088	40	19
CIENCIA_2003_24212	63	14
CIENCIA_2004_26423	85	13

Table 6.6 – Corpus Relationships - Corref-PT

	Mentions	Distinct Heads
Corpus	14470	5018
dn81201	239	119
mcarv1_2	234	109
dn81219	210	103

relationships. The CorrefPT subset's results drop by about 20%. The Summ-it++ subset's results drop by about 7%. That can be related to the fact that Summ-it++ texts are related to a specific domain, unlike CorrefPT texts. We expected that the semantic bases would improve the results more than they actually did. There was, however, a pattern in that ContoPT achieved a higher CoNLL average than other bases, including the baseline, which could not be clearly perceived in results using the entire corpus. Table 6.8 also shows that, contrary to what was seen in previous experiments, the usage of semantics improved recall without a loss in precision.

Table 6.7 – Semantic Subset - Semantic Bases - Corref-PT

Base	MUC			B_3			$CEAF_e$			CoNNL Avg
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
Baseline	24.44	**44.16**	31.46	19.65	**42.57**	26.89	29.22	**35.90**	**32.22**	30.190
OntoPT	**26.46**	31.04	28.57	**22.91**	29.86	25.93	**33.18**	29.42	31.19	28.563
ContoPT	25.85	41.42	**31.84**	20.92	38.92	**27.21**	29.97	33.54	31.65	**30.233**
ConceptNet	24.44	42.01	30.90	19.83	40.42	26.61	29.47	33.99	31.57	29.693

Table 6.8 – Semantic Subset - Semantic Bases - Summ-it++

Base	MUC			B_3			$CEAF_e$			CoNNL Avg
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
Baseline	39.20	45.79	42.24	33.47	44.77	38.31	38.08	36.89	37.48	39.343
OntoPT	**43.20**	40.60	41.86	**37.62**	39.05	38.32	**39.27**	34.78	36.89	39.023
ContoPT	42.40	**47.32**	**44.72**	35.70	**45.53**	**40.02**	38.13	36.37	37.23	**40.657**
ConceptNet	39.20	45.37	42.06	34.01	44.83	38.68	38.89	**37.09**	**37.97**	39.570

6.3.1 Error Analysys

So that we could propose new strategies to improve recall, we performed an error analysis based on the results presented in the previous sections in order to comprehend why CORP, using the semantic bases, was not able to identify certain relationships. For that purpose, we compared the system's output with the reference chains. Some named entities from CORP's output and the reference chains were not exactly the same, but this is a known limitation, since this task is performed by the parser. The configuration selected for the comparison was ContoPT, which obtained the best results for the semantic subset, with the default cut-off points and no hypernym window. Some cases in which semantic knowledge helped are presented in Table 6.9.

Table 6.9 – Example Chain 1

Reference	Found
numeroso sibilante e barulhento rapazio a garotada	numeroso sibilante e barulhento rapazio a garotada
carros de metal os carros os objectos metálicos com rodas os automóveis automóveis os automóveis	carros os carros os automóveis automóveis os automóveis
os pacientes os pacientes com a doença os doentes	os pacientes os pacientes os doentes

Some mentions could have been linked by simple relationships, but were not because the relationships were not in the semantic base. In example (6.3.1), the link between [os cientistas] and [os pesquisadores] was not created because the weight for the hypernym relationship between both entities is below the cut-off point. The relationship between [células] and [neurônios], presented in example (6.3.2), is not present in ContoPT.

(6.3.1) [os paleontólogos] [os cientistas] [os pesquisadores]
[the paleontologists] [the scientists] [the researchers]

(6.3.2) [células nervosas] [os neurônios]
[nerve cell] [the neurons]

There were cases where CORP generated two chains when the reference claims they should have been one. This is likely caused by CORP being unable to identify all of the relationships between the mentions, thus being unable to unite the chains.

(6.3.3) [os paleontólogos] [os cientistas] [os pesquisadores]
[the paleontologists] [the scientists] [the researchers]

(6.3.4) [células nervosas] [os neurônios]
[nerve cell] [the neurons]

Table 6.10 – Example Chain 2

<table>
<tr><th>Reference</th><th>Found</th></tr>
<tr><td rowspan="2">cães em fase de envelhecimento (a partir dos 9 anos de vida)
os animais
os bichos
os cães
eles</td><td>os animais
os bichos</td></tr>
<tr><td>cães
os cães</td></tr>
</table>

One of the types of relationship found was related to people and places. Example (6.3.5) is related to a region. Examples (6.3.6) and (6.3.7) are related to people, where one can notice that different occurrences of the mention [Jaspers] are related to two different chains.

(6.3.5) [a região de a Chapada do Araripe , Ceará] [o nordeste brasileiro]
[Chapada do Araripe region, in Ceará] [the Brazilian Northeast]

(6.3.6) [Jaspers][o filósofo germânico]
[Jaspers] [the german philosopher]

(6.3.7) [certos pensadores] [Jaspers] [Ingenieros] [Aldous Huxley] [Orwell]
[certain thinkers] [Jaspers] [Ingenieros] [Aldous Huxley] [Orwell]

For some cases, Wikipedia redirect links could be helpful. Considering examples (6.3.8) and (6.3.9), DBpedia Spotlight associated [o mal de Alzheimer] and [Alzheimer] to the same resource, as well as [os Estados Unidos] and [EUA]. However, since CORP considers only a mention's head, querying both mentions in DBpedia would not be enough for the first example, as the head of [o mal de Alzheimer] is [mal].

(6.3.8) [o mal de Alzheimer] [Alzheimer] [a doença] [a moléstia]
[Alzheimer disease] [Alzheimer] [the disease] [the illness]

(6.3.9) [os Estados Unidos] [EUA]
[the United States] [US]

Examples (6.3.10), (6.3.11), (6.3.12) and (6.3.13) show situations too complex to be identified by just querying a semantic base. The method proposed by Recasens et al. [60] and further described in section 3.2 could be used to resolve those mentions. This method, however, requires a comparable corpus, and their experiments were restricted to the tech news domain.

(6.3.10) [o Tyrannossaurus Rex] [o ilustre réptil norte-americano]
[the Tyrannossaurus Rex] [the well-known North American reptile]

(6.3.11) [os objectos metálicos com rodas] [automóveis] [carros de metal]
[metal objects with wheels] [automobiles] [metal cars]

(6.3.12) [Este mal de o século] [o stress]
[The health epidemic of the 21st century] [the stress]

(6.3.13) [precursores de mérito] [Kohler] [Zukermann] [Yerkes]
[talent pioneers] [Kohler] [Zukermann] [Yerkes]

6.4 Semantic Similarity based on Word Embeddings

Most semantic bases are built manually or semi-automatically. Word embedding models, on the other hand, are built automatically and include a large vocabulary. We decided to experiment with those models, using the entire corpora, in order to verify their usefulness compared to structured semantic bases.

Here we present the experiments which we considered most informative. These experiments, as well as their configurations, are based on previous ones, but using different models and threshold values. The objective is to compare the usefulness of a word embedding model compared to a structured semantic base for the coreference task, as well as verify if the knowledge obtained from both sources could be complementary. These expriments were performed using the same configuration as the experiments presented in section 6.1. For the experiments using both the semantic similarity and the semantic base rules we selected ContoPT as semantic base, which obtained the best results in the experiments presented in section 6.3.

The model used for these experiments is a Word2Vec CBOW with 300 dimensions. Experiments using that model, Glove and Word2Vec Skip-Gram showed us that, although the difference in results was very small (below 0.1% in all metrics), CBOW performed better for both recall and precision. We used a 0.7 threshold because a higher threshold would result in a very small recall gain compared to the baseline without a semantic base.

6.4.1 Results

Tables 6.11 and 6.12 show the results obtained. The semantic similarity rule was able to detect some semantic relationships, but the gain obtained in the recall was equal or lower than the obtained using the semantic bases. The experiments using both CBOW and ContoPT show us that the knowledge contained in them was not complementary, considering that a small gain in the recall was obtained at the cost of a much higher decrease in precision.

Table 6.11 – Semantic Similarity Experiments - Corref-PT

Base	MUC			B_3			$CEAF_e$			CoNNL Avg
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
Baseline	49.40	**63.42**	**55.54**	41.71	**60.04**	**49.22**	51.27	**50.35**	**50.81**	**51.857**
OntoPT	**51.69**	54.67	53.14	**44.76**	51.09	47.71	**52.96**	46.25	49.38	50.077
ContoPT	50.03	61.78	55.29	42.45	58.25	49.11	51.84	49.28	50.53	51.643
ConceptNet	49.69	62.55	55.39	42.09	59.17	49.19	51.63	49.95	50.77	51.783
CBOW	49.61	61.59	54.95	42.01	57.91	48.70	50.61	49.44	50.02	51.223
CBOW + ContoPT	50.17	60.11	54.69	42.68	56.30	48.55	51.09	48.45	49.74	50.993

Table 6.12 – Semantic Similarity Experiments - Summ-it++

Base	MUC			B_3			$CEAF_e$			CoNNL Avg
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
Baseline	41.79	**53.61**	46.97	39.42	**54.07**	**45.60**	51.32	**50.55**	50.93	**47.833**
OntoPT	**43.60**	45.71	44.63	**41.56**	45.77	43.56	**52.47**	44.63	48.23	45.473
ContoPT	42.44	53.02	**47.14**	39.89	53.21	45.59	51.27	49.75	50.50	47.743
ConceptNet	41.93	52.91	46.79	39.69	53.42	45.54	51.64	50.33	**50.98**	47.770
CBOW	41.93	52.87	46.77	39.60	53.27	45.43	51.36	49.98	50.66	47.620
CBOW + ContoPT	42.58	52.35	46.96	40.07	52.50	45.45	51.32	49.21	50.24	47.550

6.4.2 Error Analysis

In order to comprehend both the situations in which the semantic similarity rule was useful and those in which it was not, we performed an error analysis of the results. Some positive examples found by the semantic similarity rule are presented in Table 6.13.

Table 6.13 – Positive examples found

Head	**Antecedent Head**	**Included in ContoPT**
maneira	forma	yes
confronto	embate	no
começo	início	no
cachorros	cães domésticos	yes
renúncia	demissão	no
incremento	acréscimo	no
acréscimo	aumento	yes
veículos	carros	yes
hábitos	costumes	yes

By our analysis, we noticed that most examples found by the semantic similarity rule were already found in the semantic base. Additionally, most examples found in the semantic base were not found by the semantic similarity rule (considering our threshold). More importantly, we found many false-positives which, although semantically similar, are not helpful for our purposes. Some examples are listed below:

- Months: [junho] [outubro]
- Time period: [ano] [mês]

- Geographic coordinates: [sul] [oeste]
- Metals: [ouro] [prata]
- Antonyms: [redução] [aumento]
- Units: [dezenas] [centenas]
- Family members: [filha] [pai]

To overcome those false-positives, strategies to obtain certain relationships for these models could be implemented. Kim et al. [48] proposed a vector representation which integrates lexical contrast information from WordNet and Wordnik to distinguish between synonyms and antonyms. In the resultant model, antonyms were less similar whereas synonyms were more similar. This distinction, however, performed better between adjectives and verbs, while the improvement for nouns was rather small.

7. FINAL CONSIDERATIONS

The main objective of this work was to improve the coreference resolution task for Portuguese using semantic bases. For this, we reviewed the literature related to the task, including different strategies to resolve coreferent mentions and works which attempted to address the use of semantics for the task. We reviewed the semantic resources available for the Portuguese language, selected three semantic resources for our experiments and implemented the changes necessary in CORP to integrate it with these resources. Experiments using the different resources were performed, as well as experiments using different parameters, in order to verify how those could improve the results. The results obtained were compared to the existing CORP version, which used OntoPT already, as well as a baseline composed of CORP without any semantic base. Although the results were ambiguous, they revealed some perspectives on the problem of using semantics for coreference resolution. It was found that semantic relationships contained in corpora which could be identified by simple semantic relationships were few, many of them were not contained in any semantic base experimented, and the differences obtained in metrics were insufficient to obtain concrete evidence. To address this problem we selected a few texts containing more semantic relationships than the average and performed experiments. These results showed a higher CoNLL average using ContoPT for both corpora subsets, which may indicate that ContoPT was not highlighted by results obtained using the entire corpora due to the small percentage of semantic relationships present. The experiments performed with the semantic similarity feature showed that simply calculating cosine similarity in order to find coreferent mentions does not improve the coreference resolution task as much as a structured semantic base. Our analysis described the difficulties faced when using those models and some cases which can generate noise. Finally, considering that few works focused on how to use semantics for coreference resolution, we believe that our findings could be of aid to future works in this area.

7.1 Contributions

The main contribution of this work is a new CORP version integrated with three new resources. This new version includes minor improvements, further described in chapter 5, which we believe could assist other researchers in their experiments and investigations. We consider the Word Embeddings web API another contribution, since it allows anyone to easily experiment with different models or even semantic bases without the need to make any change to CORP.

Our analysis exposed some difficulties involved in the detection of semantic relationships. We believe that this analysis could assist in future work, as it highlights some types of relationships that are not addressed yet.

The preliminary results obtained with ConceptNet, alongside the accompanying error analysis, were published in the article “Analysing Semantic Resources for Coreference Resolution” [33].

7.2 Future Work

As future work, we aim to refine the semantic similarity rule. A possibility is to use strategies to filter the cases where mentions are semantically similar, but not synonyms or hypernyms. One possibility of improvement in CORP, already pointed out by Evandro Fonseca [13], is the resolution of pronominal coreference. From our analysis, we detected some chains containing pronouns and we believe resolving pronouns could improve both accuracy and precision.

One drawback of the semantic rules is that only the mentions’ heads are compared. For instance, when comparing [Alzheimer] and [Mal de Alzheimer], only the head [Mal] from the second mention is used, which is not very informative for our purposes. Therefore, it is clear that comparing the entire noun phrase would be helpful in some situations. DBpedia spotlight, for instance, can successfully associate [Alzheimer] and [Mal de Alzheimer] to the same resource. Word embedding models could also compare the entire noun phrase and, theoretically, could resolve even more complex cases, such as [os objectos metálicos com roda] and [os carros].

REFERENCES

[1] Antonitsch, A.; Figueira, A.; Amaral, D.; Fonseca, E.; Vieira, R.; Collovini, S. "Summ-it++: an enriched version of the summ-it corpus". In: International Conference on Language Resources and Evaluation, 2016, pp. 2047–2051.

[2] Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. "Dbpedia: A nucleus for a web of open data". In: International Semantic Web Conference, 2007, pp. 722–735.

[3] Bagga, A.; Baldwin, B. "Algorithms for scoring coreference chains". In: International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, 1998, pp. 563–566.

[4] Baker, C. F.; Fillmore, C. J.; Lowe, J. B. "The berkeley framenet project". In: Annual Meeting of the Association for Computational Linguistics, 1998, pp. 86–90.

[5] Bean, D.; Riloff, E. "Unsupervised learning of contextual role knowledge for coreference resolution". In: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2004, pp. 297–304.

[6] Cardie, C.; Wagstaf, K. "Noun phrase coreference as clustering". In: Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999, pp. 82–89.

[7] Clark, K.; Manning, C. D. "Improving coreference resolution by learning entity-level distributed representations". In: Annual Meeting of the Association for Computational Linguistics, 2016, pp. 643–653.

[8] Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Rino, L.; Vieira, R. "Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática". In: Tecnologia da Informação e da Linguagem Humana, 2007, pp. 1615–1614.

[9] da Silva, J. F. "Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado", Master's Thesis, Universidade de São Paulo, 2011, 108p.

[10] de Paiva, V.; Rademaker, A.; de Melo, G. "Openwordnet-pt: An open Brazilian Wordnet for reasoning". In: International Conference on Computational Linguistics: Demonstration Papers, 2012, pp. 353–360.

[11] Elango, P. "Coreference resolution: A survey", Technical Report, University of Wisconsin-Madison, 2005, 8p.

[12] Fillmore, C. J. “Frame semantics and the nature of language”, *Annals of the New York Academy of Sciences*, vol. 280–1, Dec 1976, pp. 20–32.

[13] Fonseca, E. “Resolução de correferência nominal usando semântica em língua portuguesa”, Ph.D. Thesis, Graduate Program in Computer Science, PUCRS, 2018, 109p.

[14] Fonseca, E.; Sesti, V.; Antonitsch, A.; Vanin, A.; Vieira, R. “Corp: Uma abordagem baseada em regras e conhecimento semântico para a resoluçao de correferências”, *Linguamática*, vol. 9–1, Jul 2017, pp. 3–18.

[15] Fonseca, E.; Sesti, V.; Collovini, S.; Vieira, R.; Leal, A. L.; Quaresma, P. “Collective elaboration of a coreference annotated corpus for portuguese texts”. In: Workshop on Evaluation of Human Language Technologies for Iberian Languages, 2017, pp. 68–82.

[16] Fonseca, E.; Vieira, R.; Vanin, A. “Improving coreference resolution with semantic knowledge”. In: International Conference on Computational Processing of the Portuguese Language, 2016, pp. 213–224.

[17] Freitas, C.; Mota, C.; Santos, D.; Oliveira, H. G.; Carvalho, P. “Second harem: Advancing the state of the art of named entity recognition in portuguese”. In: International Conference on Language Resources and Evaluation, 2010, pp. 3630–3637.

[18] Garcia, M.; Gamallo, P. “An entity-centric coreference resolution system for person entities with rich linguistic information.” In: International Conference on Computational Linguistics, 2014, pp. 741–752.

[19] Ge, N.; Hale, J.; Charniak, E. “A statistical approach to anaphora resolution”. In: Special Interest Group on Linguistic data and corpus-based approaches to NLP, 1998, pp. 161–170.

[20] Grishman, R.; Sundheim, B. “Message understanding conference-6: A brief history”. In: International Conference on Computational Linguistics, 1996, pp. 466–471.

[21] Grosz, B. J.; Weinstein, S.; Joshi, A. K. “Centering: A framework for modeling the local coherence of discourse”, *Computational Linguistics*, vol. 21–2, Jun 1995, pp. 203–225.

[22] Haghighi, A.; Klein, D. “Unsupervised coreference resolution in a nonparametric bayesian model”. In: Annual Meeting of the Association of Computational Linguistics, 2007, pp. 848–855.

[23] Haghighi, A.; Klein, D. “Simple coreference resolution with rich syntactic and semantic features”. In: Conference on Empirical Methods in Natural Language Processing, 2009, pp. 1152–1161.

[24] Harabagiu, S. M.; Bunescu, R. C.; Maiorano, S. J. “Text and knowledge mining for coreference resolution”. In: North American Chapter of the Association for Computational Linguistics on Language technologies, 2001, pp. 1–8.

[25] Hartmann, N.; Fonseca, E.; Shulby, C.; Treviso, M.; Rodrigues, J.; Aluisio, S. “Portuguese word embeddings: Evaluating on word analogies and natural language tasks”. In: Brazilian Symposium in Information and Human Language Technology, 2017, pp. 122–131.

[26] Hobbs, J. “Pronoun resolution”, *Lingua*, vol. 44-4, Apr 1978, pp. 311–338.

[27] Ji, H.; Westbrook, D.; Grishman, R. “Using semantic relations to refine coreference decisions”. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 17–24.

[28] Jurafsky, D.; Martin, J. H. “Speech and Language Processing”. Prentice-Hall, 2009, 1024p.

[29] Kamp, H.; Reyle, U. “From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory”. Springer Science & Business Media, 2013, 689p.

[30] Lee, H.; Chang, A.; Peirsman, Y.; Chambers, N.; Surdeanu, M.; Jurafsky, D. “Deterministic coreference resolution based on entity-centric, precision-ranked rules”, *Computational Linguistics*, vol. 39–4, Dec 2013, pp. 885–916.

[31] Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M.; Jurafsky, D. “Stanford’s multi-pass sieve coreference resolution system at the conll-2011 shared task”. In: Conference on Computational Natural Language Learning, pp. 28–34.

[32] Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. “End-to-end neural coreference resolution”. In: Conference on Empirical Methods in Natural Language Processing, 2017, pp. 188–197.

[33] Lima, T.; Collovini, S.; Leal, A.; Fonseca, E.; Han, X.; Huang, S.; Vieira, R. “Analysing semantic resources for coreference resolution”. In: Computational Processing of the Portuguese Language, 2018, pp. 284–293.

[34] Luo, X. “On coreference resolution performance metrics”. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 25–32.

[35] Luo, X.; Ittycheriah, A.; Jing, H.; Kambhatla, N.; Roukos, S. “A mention-synchronous coreference resolution algorithm based on the bell tree”. In: Annual Meeting of the Association for Computational Linguistics, 2004, pp. 135–142.

[36] Maziero, E. G.; Pardo, T. A. S.; Di Felippo, A.; Dias-da Silva, B. C. “A base de dados lexical e a interface web do tep 2.0: thesaurus eletrônico para o português do brasil”. In: Brazilian Symposium on Multimedia and the Web, 2008, pp. 390–392.

[37] McCallum, A.; Wellner, B. “Conditional models of identity uncertainty with application to noun coreference”. In: Advances in Neural Information Processing Systems, 2005, pp. 905–912.

[38] Mendes, P. N.; Jakob, M.; Bizer, C. “Dbpedia for nlp: A multilingual cross-domain knowledge base”. In: International Conference on Language Resources and Evaluation, 2012, pp. 1813–1817.

[39] Mendes, P. N.; Jakob, M.; García-Silva, A.; Bizer, C. “Dbpedia spotlight: shedding light on the web of documents”. In: International Conference on Semantic Systems, 2011, pp. 1–8.

[40] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. “Efficient estimation of word representations in vector space”. In: International Conference on Learning Representations, 2013, pp. 12.

[41] Mikolov, T.; Yih, W.; Zweig, G. “Linguistic regularities in continuous space word representations”. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 746–751.

[42] Miller, G. A. “Wordnet: A lexical database for english”, *Communications of the ACM*, vol. 38–11, Nov 1995, pp. 39–41.

[43] Moraes, T. C. “Resolução de correferência e categorias de entidades nomeadas”, Master’s Thesis, Graduate Program in Computer Science, PUCRS, 2010, 76p.

[44] Müller, C.; Rapp, S.; Strube, M. “Applying co-training to reference resolution”. In: Annual Meeting on Association for Computational Linguistics, 2002, pp. 352–359.

[45] Navigli, R.; Ponzetto, S. P. “Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network”, *Artificial Intelligence*, vol. 193, Dec 2012, pp. 217–250.

[46] Ng, V. “Shallow semantics for coreference resolution”. In: International Joint Conferences on Artificial Intelligence, 2007, pp. 1689–1694.

[47] Ng, V. “Unsupervised models for coreference resolution”. In: Conference on Empirical Methods in Natural Language Processing, 2008, pp. 640–649.

[48] Nguyen, K. A.; Walde, S. S. i.; Vu, N. T. “Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction”. In: Annual Meeting of the Association for Computational Linguistics, 2016, pp. 454–459.

[49] Oliveira, H. G. “Conto.pt: Groundwork for the automatic creation of a fuzzy portuguese wordnet”. In: Computational Processing of the Portuguese Language, 2016, pp. 283–295.

[50] Oliveira, H. G. “A survey on portuguese lexical knowledge bases: Contents, comparison and combination”, *Information*, vol. 9–2, Feb 2018, pp. 1–34.

[51] Oliveira, H. G.; de Paiva, V.; Freitas, C.; Rademaker, A.; Real, L.; Simões, A. “As wordnets do português”, *Oslo Studies in Language*, vol. 7–1, Nov 2015, pp. 387–424.

[52] Oliveira, H. G.; Gomes, P. “Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically”, *Language Resources and Evaluation*, vol. 48–2, Jun 2014, pp. 373–393.

[53] Pierce, D.; Cardie, C. “Limitations of co-training for natural language learning from large datasets”. In: Conference on Empirical Methods in Natural Language Processing, 2001, pp. 1–9.

[54] Ponzetto, S. P.; Strube, M. “Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution”. In: Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006, pp. 192–199.

[55] Ponzetto, S. P.; Strube, M. “Knowledge derived from wikipedia for computing semantic relatedness”, *Journal of Artificial Intelligence Research*, vol. 30–1, Sep 2007, pp. 181–212.

[56] Pradhan, S.; Luo, X.; Recasens, M.; Hovy, E.; Ng, V.; Strube, M. “Scoring coreference partitions of predicted mentions: A reference implementation”. In: Annual Meeting of the Association for Computational Linguistics, 2014, pp. 30–35.

[57] Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; Zhang, Y. “Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes”. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning Shared Task, 2012, pp. 1–40.

[58] Prokofyev, R.; Tonon, A.; Luggen, M.; Vouilloz, L.; Difallah, D. E.; Cudré-Mauroux, P. “Sanaphor: Ontology-based coreference resolution”. In: International Semantic Web Conference, 2015, pp. 458–473.

[59] Rahman, A.; Ng, V. “Coreference resolution with world knowledge”. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 814–824.

[60] Recasens, M.; Can, M.; Jurafsky, D. “Same referent, different words: Unsupervised mining of opaque coreferent mentions”. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 897–906.

[61] Recasens, M.; Hovy, E. “Blanc: Implementing the rand index for coreference evaluation”, *Natural Language Engineering*, vol. 17–4, Oct 2011, pp. 485–510.

[62] Recasens, M.; Màrquez, L.; Sapena, E.; Martí, M. A.; Taulé, M.; Hoste, V.; Poesio, M.; Versley, Y. “Semeval-2010 task 1: Coreference resolution in multiple languages”. In: International Workshop on Semantic Evaluation, 2010, pp. 1–8.

[63] Salomao, M. M. M. “Framenet brasil: um trabalho em progresso”, *Calidoscópio*, vol. 7–3, Sep 2009, pp. 171–182.

[64] Santos, D.; Barreiro, A.; Freitas, C.; Oliveira, H. G.; Medeiros, J. C.; Costa, L.; Gomes, P.; Silva, R. “Relações semânticas em português: comparando o tep, o mwn.pt, o port4nooj e o papel”. In: Encontro Nacional da Associção Portuguesa de Linguística, 2010, pp. 681–700.

[65] Sapena Masip, E.; Padró, L.; Turmo Borras, J. “Coreference resolution survey”, Technical Report, TALP Research Center, 2008, 39p.

[66] Silva, W. D. C. d. M. “Aprimorando o corretor gramatical cogroo”, Master’s Thesis, Universidade de São Paulo, 2013, 178p.

[67] Simões, A.; Guinovart, X. G. “Bootstrapping a portuguese wordnet from galician, spanish and english wordnets”. In: Advances in Speech and Language Technologies for Iberian Languages, 2014, pp. 239–248.

[68] Simova, I.; Uszkoreit, H. “Word embeddings as features for supervised coreference resolution”. In: International Conference Recent Advances in Natural Language Processing, 2017, pp. 686–693.

[69] Singla, P.; Domingos, P. “Discriminative training of markov logic networks”. In: American Association for Artificial Intelligence, 2005, pp. 868–873.

[70] Soon, W. M.; Ng, H. T.; Lim, D. C. Y. “A machine learning approach to coreference resolution of noun phrases”, *Computational Linguistics - Special Issue on Computational Anaphora Resolution*, vol. 27–4, Dec 2001, pp. 521–544.

[71] Speer, R.; Chin, J.; Havasi, C. “Conceptnet 5.5: An open multilingual graph of general knowledge.” In: AAAI Conference on Artificial Intelligence, 2017, pp. 4444–4451.

[72] Suchanek, F. M.; Kasneci, G.; Weikum, G. “Yago: A core of semantic knowledge”. In: International Conference on World Wide Web, 2007, pp. 697–706.

[73] Sukthanker, R.; Poria, S.; Cambria, E.; Thirunavukarasu, R. “Anaphora and coreference resolution: A review”, Technical Report, School of Information Technology and Engineering, VIT Universit, 2018, 45p.

[74] Vieira, R.; Lima, V. L. “Lingüística computacional: princípios e aplicações”. In: Congresso da SBC: Jornada de Atualização em Inteligência Artificial, 2001, pp. 47–86.

[75] Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; Hirschman, L. “A model-theoretic coreference scoring scheme”. In: Conference on Message Understanding, 1995, pp. 45–52.

[76] Wiseman, S.; Rush, A. M.; Shieber, S. M. “Learning global features for coreference resolution”. In: North American Chapter of the Association for Computational Linguistics:Human Language Technologies, 2016, pp. 994—1004.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br