PUCRS

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

EVANDRO BRASIL FONSECA

RESOLUÇÃO DE CORREFERÊNCIA NOMINAL USANDO SEMÂNTICA EM LÍNGUA PORTUGUESA

Porto Alegre
2018

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RESOLUÇÃO DE CORREFERÊNCIA NOMINAL USANDO SEMÂNTICA EM LÍNGUA PORTUGUESA

EVANDRO BRASIL FONSECA

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Renata Vieira
Co-Orientador: Prof. Aline Aver Vanin

Porto Alegre
2018

Ficha Catalográfica

F676r Fonseca, Evandro Brasil

Resolução de Correferência Nominal Usando Semântica em Língua Portuguesa / Evandro Brasil Fonseca . – 2018.
117 p.
Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Renata Vieira.
Co-orientadora: Profa. Dra. Aline Vanin.

1. Resolução de Correferência. 2. Extração de Informação. I. Vieira, Renata. II. Vanin, Aline. 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

Evandro Brasil Fonseca

Resolução de Correferência Nominal Usando Semântica em Língua Portuguesa

Tese apresentada como requisito parcial para obtenção do grau de Doutor em Ciência da Computação do Programa de Pós-Graduação em Ciencia da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado em 19 de março de 2018.

BANCA EXAMINADORA:

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

Profa. Dra. Valéria Delisandra Feltrim (DIN/UEM)

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

Prof. Dra Renata Vieira (PPGCC/PUCRS – Orientadora)

Prof. Dra Aline Vanin (DEH/UFCSPA – Co-orientadora)

AGRADECIMENTOS

Primeiramente agradeço a minha orientadora, Renata Vieira, pelos conselhos, ensinamentos e atenção recebida. À minha co-orientadora, Aline Vanin, pelos ensinamentos e pela motivação. A meus pais, Elemar Leal da Fonseca e Ilda Brasil da Fonseca, meus eternos heróis e fontes de inspiração, pela educação e apoio recebido ao longo de meu caminho. À Faculdade de Informática da PUCRS, excelência em ensino e pesquisa, a quem devo minha formação acadêmica. Ao CNPQ, CAPES e FAPERGS pelo apoio financeiro durante todo o doutorado e, por fim, aos colegas do laboratório de PLN, pelo companheirismo e auxílio no transcorrer deste estudo.

RESOLUÇÃO DE CORREFERÊNCIA NOMINAL USANDO SEMÂNTICA EM LÍNGUA PORTUGUESA

RESUMO

A tarefa de Resolução de Correferência é um grande desafio para a área de Processamento da Linguagem Natural, tendo em vista o conhecimento linguístico exigido e a sofisticação das técnicas de processamento da língua empregados. Mesmo sendo uma tarefa desafiadora, um fator motivador do estudo deste fenômeno se dá pela sua utilidade. Basicamente, várias tarefas de Processamento da Linguagem Natural podem se beneficiar de seus resultados, como, por exemplo, o reconhecimento de entidades nomeadas, extração de relação entre entidades nomeadas, sumarização, análise de sentimentos, entre outras. A Resolução de Correferência é um processo que consiste em identificar determinados termos e expressões que remetem a uma mesma entidade. Por exemplo, na sentença "*A França está resistindo. O país é um dos primeiros no ranking...*" podemos dizer que [*o país*] é uma correferência de [*A França*]. Realizando o agrupamento desses termos referenciais, formamos grupos de menções correferentes, mais conhecidos como cadeias de correferência. Esta tese propõe um processo para a resolução de correferência entre sintagmas nominais para a língua portuguesa, tendo como foco a utilização do conhecimento semântico. Nossa abordagem proposta é baseada em regras linguísticas sintático-semânticas. Ou seja, combinamos diferentes níveis de processamento linguístico utilizando relações semânticas como apoio, de forma a inferir relações referenciais entre menções. Modelos baseados em regras linguísticas têm sido aplicados eficientemente em outros idiomas como o inglês, o espanhol e o galego. Esses modelos mostram-se mais eficientes que os baseados em aprendizado de máquina quando lidamos com idiomas menos providos de recursos, dado que a ausência de corpora ricos em amostras pode prejudicar o treino desses modelos. O modelo proposto nesta tese é o primeiro voltado para a resolu-

ção de correferência em português que faz uso de conhecimento semântico. Dessa forma, tomamos este fator como a principal contribuição deste trabalho.

Palavras-Chave: Resolução de Correferência, Extração de Informação.

NOMINAL COREFERENCE RESOLUTION USING SEMANTICS IN PORTUGUESE

ABSTRACT

Coreference Resolution task is challenging for Natural Language Processing, considering the required linguistic knowledge and the sophistication of language processing techniques involved. Even though it is a demanding task, a motivating factor in the study of this phenomenon is its usefulness. Basically, several Natural Language Processing tasks may benefit from their results, such as named entities recognition, relation extraction between named entities, summarization, sentiment analysis, among others. Coreference Resolution is a process that consists on identifying certain terms and expressions that refer to the same entity. For example, in the sentence " *France is refusing. The country is one of the first in the ranking...* " we can say that [*the country*] is a coreference of [*France*]. By grouping these referential terms, we form coreference groups, more commonly known as coreference chains. This thesis proposes a process for coreference resolution between noun phrases for Portuguese, focusing on the use of semantic knowledge. Our proposed approach is based on syntactic-semantic linguistic rules. That is, we combine different levels of linguistic processing, using semantic relations as support, in order to infer referential relations between mentions. Models based on linguistic rules have been efficiently applied in other languages, such as: English, Spanish and Galician. In few words, these models are more efficient than machine learning approaches when we deal with less resourceful languages, since the lack of sample-rich corpora may produce a poor training. The proposed approach is the first model for Portuguese coreference resolution which uses semantic knowledge. Thus, we consider it as the main contribution of this thesis.

Keywords: Coreference Resolution, Information Extraction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

PLN – Processamento da Linguagem Natural

REN – Reconhecimento de Entidades Nomeadas

EN – Entidades Nomeadas

PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul

SUMÁRIO

1. INTRODUÇÃO

A Resolução de Correferência a partir de textos é uma tarefa útil e também um dos principais desafios da área de Processamento da Linguagem Natural (PLN) . Isso porque esta tarefa depende de diversos níveis de processamento, como análise sintática, morfológica, extração de sintagmas nominais, entre outros. Na literatura, encontramos diversas iniciativas para a língua portuguesa que abordam esse problema, geralmente separados entre a resolução de anáfora [71], [8], [59], [19], [5] e o estudo da correferência nominal [30], [21], [28], [25]. Resolução de Correferência é uma tarefa que consiste em identificar as diferentes formas que uma mesma menção pode assumir em um discurso. Em outras palavras, esse processo consiste em identificar determinados termos e expressões que remetem a uma mesma referência. Na sentença: "*A França resiste como o único país de a União Européia a não permitir patenteamento de genes . A UE...*" podemos dizer que [*o único país de a União Européia a não permitir patenteamento de genes*] é uma correferência de [*A França*], da mesma forma que [*A UE*] é uma correferência de [*a União Européia*]. Agrupando esses termos, formamos grupos de menções referenciais ou, mais conhecidos como cadeias de correferência.

De forma geral, isto é, independente do idioma, muitos trabalhos concentram seus esforços em técnicas de aprendizado de máquina [18, 64, 43, 54, 48, 17, 48, 11, 73, 28, 14, 68]. Dentre eles, Soon et al. [64], um dos pioneiros nesse tipo de abordagem, propõe um modelo baseado em aprendizado supervisionado para o Inglês. No entanto, quando lidamos com técnicas de aprendizado de máquina, a obtenção de bons resultados depende não somente das *features*/características utilizadas, mas da qualidade e quantidade de amostras existentes para treino. Embora a quantidade de recursos para o Português venha aumentando nos últimos tempos ainda existe uma grande carência por corpora ricos em anotação de correferência que contenham anotações suficientes para treinar modelos eficientes. Além disso, quando consideramos empregar o uso da semântica, essa carência é ainda maior, dado que a quantidade de amostras reduz-se drasticamente. Como exemplo de tais afirmações podemos citar os dois principais corpora para o Inglês e para o Português: analisando-os temos, respectivamente, 34290 cadeias para o corpus Ontonotes [53] e 560 cadeias para o corpus Summ-it [12]. Dados esses fatos, o aprendizado de máquina pode não ser a melhor opção para conceber modelos de resolução de correferência, quando exploramos idiomas menos providos de recursos. Para tal cenário, defendemos que abordagens baseadas em regras linguísticas podem prover resultados mais significativos quando comparados a aprendizagem de máquina. Além de optarmos por uma abordagem baseada em regras para a resolução de correferência em Português, outro diferencial deste trabalho reside no fato de a maioria dos trabalhos basear-se somente no conhecimento lexical e sintático. Não estamos descartando tal conhecimento, este certamente é indispensável e

muito utilizado na tarefa, como em: "*o presidente Barack Obama diz que nunca ordenou que cidadãos americanos fossem espionados. Obama se posicionou...*". Note que existe uma relação correferencial entre [*o presidente Barack Obama*] e [*Obama*], relação que pode ser identificada pelo termo "Obama" em comum. Além disso, o processamento em nível lexical e sintático vai além de tal casamento de padrões, pois nem sempre menções parcialmente iguais podem ser tratadas como menções correferenciais. Observe: "*Hoje estarei na Universidade de São Paulo e amanhã na Universidade do Paraná...*". Note que existe uma similaridade lexical nas menções, que é evocada pelo termo "Universidade". No entanto, as menções "Universidade de São Paulo" e "Universidade do Paraná" são entidades totalmente distintas e, portanto, não correferentes. Para casos como este podemos recorrer técnicas de justaposição e identificar termos modificadores, como Lee et al. [39] sugere. Contudo, temos casos em que tal regra nem sempre é válida: "*Adalberto Portugal informou que irá permanecer em Portugal até segunda ordem...*". Assumindo que [*Portugal*] não possui termos que o modifiquem, do ponto de vista sintático poderíamos ter uma relação referencial entre [*Adalberto Portugal*] e [*Portugal*], mas o primeiro refere-se a uma pessoa e, o segundo, a um local. Nesse ponto já é possível perceber a importância do conhecimento semântico para tal tarefa. Contudo, vamos mais além: "*Já se perguntou como as abelhas fabricam mel? Os insetos saem em busca de...*". Note que não temos qualquer evidência sintática ou lexical que estabeleça uma relação de correferência entre os sintagmas nominais [*as abelhas*] e [*os insetos*]. Para esse caso e muitos outros, precisamos fazer uso de conhecimento semântico.

Como veremos nesta tese, a semântica pode prover ganhos significativos para a tarefa de resolução de correferência, pois introduz um novo nível de detecção de menções referenciais. Contudo, esses ganhos podem também dar espaço a ambiguidades, como podemos ver na sentença: "*O ministro da justiça do país informou que irá permanecer... pois isso não fornece igualdade neste processo...*". Se analisarmos a sentença à procura de relações semânticas, temos que [*justiça*] e [*igualdade*] apresentam uma relação de sinonímia[1], mas referem-se a coisas distintas neste contexto. Em poucas palavras, para obtermos bons resultados por meio da semântica, são necessários uma série de cuidados, como uma correta análise sintática, como cada menção está representada no discurso, levando em consideração seu contexto (pragmática), suas relações com outras menções, entre outros.

[1] Considerando as relações semânticas existentes no Onto.PT.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho tem como objetivo geral propor, implementar e avaliar um processo para a resolução de correferência nominal do tipo identidade em textos da Língua Portuguesa, que seja capaz de identificar e agrupar menções referenciais por meio de regras lexicais, sintáticas e semânticas, independente do domínio e em domínios não especializados.

1.1.2 Objetivos Específicos

- Apresentar uma análise detalhada do estado da arte sobre a tarefa de resolução automática de correferência;
- Construir um corpus anotado, para avaliar o processo proposto;
- Definir um modelo para a resolução automática de correferência que incorpore informação semântica;
- Avaliar o modelo proposto.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada a fundamentação conceitual do trabalho proposto; no Capítulo 3 são apresentados sistemas e abordagens computacionais que representam o estado da arte de resolução de correferência e os sistemas para o Português que situam este idioma na tarefa; no Capítulo 4, apresentamos os principais recursos utilizados no decorrer deste trabalho, como corpora utilizados, parser, ferramenta para anotação de correferência, entre outras. No Capítulo 5 apresentamos nosso modelo para a resolução de correferência em Português com informação semântica; no Capítulo 6 apresentamos experimentos realizados com nosso modelo e suas variações, mostrando os principais aspectos de cada abordagem. Por fim, no Capítulo 7 apresentamos as considerações finais, contribuições e direções futuras para este trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo buscamos enfatizar conceitos fundamentais para o entendimento deste trabalho: referentes, entidades nomeadas, sintagmas nominais, tipos de referentes, relações semânticas referenciais, correferência e anáfora.

2.1 Referentes

Referentes, ou menções, podem ser definidos como termos os quais usamos para nos referirmos a determinada entidade em um discurso. Em um texto, essas referências podem aparecer como uma entidade nomeada específica ou como parte constituinte de um sintagma nominal.

2.1.1 Entidades Nomeadas

Entidades nomeadas, a grosso modo, são elementos que podem ser referenciados por meio de nomes próprios [36]. Esses nomes próprios podem configurar-se em classes específicas, tais como: Pessoa (nomes de pessoas), Organização (nomes de empresas), Local nomes de (lugares), entre outras. Por meio dos exemplos abaixo, podemos identificar diversas entidades nomeadas (ENs), como Banco Nacional de Desenvolvimento Econômico e Social (a), Apple (b), nomes de bandas musicais (c).

a) O Banco Nacional de Desenvolvimento Econômico e Social (BNDES), empresa pública federal, é hoje o principal instrumento de financiamento de longo prazo...

b) A Apple informou que vendeu 5 milhões de iPhone 5 só em um fim de semana...

c) Várias bandas de black metal tiveram influências do punk, tais como Venom, Celtic Frost, Bathory, Sarcófago, Darkthrone, Impaled, Nazarene, Mayhem, Hellhammer, Behemoth, entre outras...

2.1.2 Sintagmas Nominais

São unidades formadas por uma ou mais palavras que, juntas, desempenham uma função sintática específica na frase. A natureza de um sintagma depende diretamente do

elemento que constitui seu núcleo. Nesta tese damos foco a menções expressas por sintagmas nominais. Dito isso, temos então os sintagmas nominais, cujos núcleos podem configurar-se em nome comum, próprio ou um pronome. Os pronomes podem apresentar-se, basicamente, nas formas de pronome pessoal, demonstrativo, indefinido, possessivo ou relativo. Um sintagma nominal geralmente é composto por um determinante seguido de um substantivo, por exemplo: "**O especialista** não respondeu todas as perguntas.". Na sentença, "**O especialista**" é um sintagma nominal. O artigo "**O**" chamamos de seu determinante. Por meio do determinante de um sintagma é possível extrair informações valiosas. Isto é, a palavra "especialista", por si só, pode assumir diferentes papéis. Contudo, "**O especialista**" qualifica uma pessoa do sexo masculino, além de informar quem é o especialista (apenas um, não dois ou mais). Note que o determinante carrega informações úteis para o processamento linguístico. Contudo, sintagmas nominais podem configurar-se em apenas substantivos: "Rio de Janeiro, cidade maravilhosa". Na sentença temos dois sintagmas nominais sem determinante: "Rio de Janeiro" e "cidade maravilhosa". Respectivamente um nome próprio e um substantivo comum seguido de seu adjunto adnominal. Algumas vezes esse adjunto pode ser predicativo. Para entendermos a diferença entre adjunto adnominal e predicativo basta observarmos que ora um termo pode exercer a função de adjunto, ora de predicativo. Ou seja, enquanto o adjunto adnominal representa o termo acessório da oração, o predicativo se revela como um termo essencial, de modo a deixá-la compreensível, dotada de sentido. Por exemplo:

- A cidade que é referência em saúde e segurança.

"referência em saúde e segurança", nesse caso, representa parte essencial à constituição do enunciado, pois sem a presença desses termos o entendimento estaria comprometido. Assim, consideramos que se trata de um predicativo, visto que atribui uma característica ao sujeito, cujo núcleo é representado por "cidade".

- "A cidade limpa que é referência em saúde e segurança.".

No exemplo acima constatamos que o termo "limpa" pode perfeitamente ser retirado do contexto oracional sem que isso cause nenhum dano ao perfeito entendimento do discurso. Logo, trata-se de um termo acessório da oração ou adjunto adnominal.

2.1.3 Tipos de Referentes

Existem três tipos de referentes: referentes específicos, referentes não-específicos e referentes abstratos.

Referentes específicos: Quando a menção de uma entidade, basicamente, identifica-a por meio de um nome comum ou próprio.

d) Microsoft informou que irá resolver o bug que reinicia o Windows Phone em dezembro.

e) Luiz Inácio Lula da Silva sancionou nesta quarta-feira, 29, a lei que regulamenta as atividades de moto-taxista e motoboy de todo país...

f) Roger Waters faz seu segundo show em São Paulo.

Em (d) temos um referente específico, isto é, a menção da entidade refere-se diretamente a algo específico, à empresa Microsoft. O referente específico, nesse caso, ainda pode ser classificado como uma entidade do tipo Organização. Existem outros tipos de referentes específicos, como Pessoa (e), Local (f) , entre outros. Note que em (f) temos dois tipos de referentes específicos, “Roger Waters” e “São Paulo”, respectivamente entidades do tipo Pessoa e Local.

Referentes não-específicos: Quando as menções referem-se a uma entidade de forma não específica. (autoridades, funcionários, policiais...), como mostram os exemplos “g”, “h” e “i”.

g) Policiais invadiram a casa, porém os bandidos já haviam fugido....

h) Funcionários estão descontentes. Eles afirmam ainda não terem recebido o seu décimo terceiro salário.

i) Autoridades disseram que estão cansados de fazer as mesmas declarações.

Referentes abstratos: como o próprio nome sugere, são entidades abstratas, “não físicas”. Tratam de estados e qualidades, sentimentos e ações, como: medo, viagem, coragem, felicidade, esforço... Exemplos “j” e “k”.

j) O medo é algo que deve ser superado. Para isso, concentre-se em seus objetivos.

k) A viagem foi ótima, porém o tempo podia estar melhor.

2.1.4 Relações Semânticas Referenciais

Como um dos focos desta tese é propor o uso do conhecimento semântico para auxiliar na tarefa de resolução de correferência é importante tornar claro os tipos de relações semânticas que podem indicar uma relação de correferência.

Hiperonímia e Hiponímia

Hiperonímia é uma relação semântica que expressa um sentido amplo entre dois termos, partindo de uma classe mais ampla para uma subclasse mais específica, por exemplo: (inseto – abelha). Neste caso, o termo “inseto” é um hiperônimo de “abelha”. Já Hiponímia representa uma relação contrária. Ou seja, parte de uma classe mais específica para uma classe mais abrangente. Para o exemplo previamente dado temos que “abelha” é um hipônimo de “inseto”. Os hiperônimos e hipônimos são importantes no campo semântico, pois são muito usados na retomada de elementos em um texto, a fim de evitar repetições desnecessárias. No que diz respeito à identificação de menções referenciais em um discurso, na língua portuguesa é muito mais comum partirmos de termos específicos para termos mais abrangentes. Dessa forma, a relação de Hiponímia geralmente ocorre com maior frequência.

l) João e Maria estão muito felizes com o seu cão. O animal é fiel e companheiro.

m) Nada disso vai fazê-los mudar de carro. O pequeno veículo parece suprir todas as necessidades deles.

Sinonímia

Trata-se de uma relação entre dois termos, em que estes, mesmo sendo distintos lexicalmente, possuem significados muito próximos, por exemplo: (menino – garoto). É importante referir que muitas vezes os sinônimos podem ter conotações diferentes, dependendo do contexto. Como por exemplo: (gato – bichano) e (gato – atraente). Em um texto, a utilização de sinônimos de uma palavra é importante para evitar repetições. Assim, um sinônimo é uma palavra que, apesar de ser diferente, tem o mesmo significado (ou semelhante) e por isso a sua inclusão não altera o sentido do texto em questão.

n) Esse carro é maravilhoso. Também, estamos falando de um automóvel de 100 mil reais.

o) Ana comprou um gato. O bichano adora dormir no sofá.

2.2 Correferência, Anáfora e Catáfora

Para o entendimento sobre o que é correferência é relevante também definirmos anáfora, já que seus conceitos estão relacionados. Anáfora pode ser definida como a retomada de uma expressão apresentada anteriormente em um texto. Quando uma entidade é mencionada pela primeira vez textualmente, temos a evocação da entidade. Durante a

leitura da sequência do texto, quando essa entidade é novamente mencionada, temos a realização do acesso a essa entidade. A expressão que faz o acesso é dita como anafórica e a expressão anterior é dita como seu antecedente [70]. Há casos de anáfora em que o termo anafórico e o antecedente são correferentes, isto é, remetem a uma mesma entidade (como os Exemplos “p” e “q” ilustram), mas há também casos de anáfora sem correferência, como podemos ver em “r”.

p) A Ana comprou um cão. O animal já conhece todos os cantos da casa. – Nesse exemplo, o termo anafórico é o grupo nominal “o animal”, que retoma o valor referencial do antecedente, “o cão”. É a relação entre “cão” e “animal” que suporta a correferência.

q) Maria está com febre. Acho que ela está doente. – Note que a interpretação referencial do sintagma nominal “ela” depende da sua relação anafórica com o sintagma nominal “Maria”.

r) O João faz 18 anos no dia 2 de Julho de 2001. No dia seguinte, parte para uma grande viagem pela Europa. – Já nesse caso, o valor referencial da expressão sublinhada constrói-se a partir da interpretação do antecedente, a expressão adverbial “temporal no dia 2 de Julho de 2001”. Assim, “No dia seguinte” designa o dia 3 de Julho de 2001.

Catáfora: semelhante à anáfora mas em ordem oposta, uma relação catafórica ocorre quando um termo se refere a outro que vem à frente e lhe dá, a partir deste, o seu sentido. Conforme podemos ver no exemplo “s”:

s) A mãe olhou-o e disse: - Meu filho, estás com um olhar cansado.

Correferência: é um fenômeno que ocorre quando duas ou mais menções em um discurso referem-se a uma mesma entidade. O conjunto de menções a uma mesma entidade no texto é denominado de cadeia de correferência.

t) O João está doente. Vi-o na semana passada. – Neste caso, o pronome “o” é uma anáfora de “João”, pois, para ser compreendido, necessita resgatar a frase anterior para que seu significado seja construído. Já o tipo aposto ocorre quando o termo da oração se relaciona a uma entidade para esclarecê-la ou explicá-la.

u) Cubatão, a cidade mais poluída do Brasil, localiza-se na Baixada Santista.

v) Maria comprou várias frutas: mamão, melancia, abacate e uva. – Normalmente, o aposto aparece isolado por sinais de pontuação, sendo mais comum aparecer entre vírgulas ou então introduzido por dois pontos. Nos exemplos acima podemos notar que “cidade” é correferente de “Cubatão”, e “mamão, melancia, abacate e uva” são correferentes de “frutas”.

w) (extraído do corpus Summ-it [12]) “A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia. A opinião é de Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina). Guerra participou do debate “Biotecnologia para uma Agricultura Sustentável”... Para o agrônomo, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local...”

No fragmento de texto acima, as expressões “Guerra” e “o agrônomo” fazem referência à entidade “Miguel Guerra”, já mencionada anteriormente no texto. Para não repetir a mesma expressão faz-se uso de outra diferente, mas que retoma a mesma entidade mencionada previamente. Esse é um método muito utilizado no processo de escrita para não deixar o texto repetitivo e está diretamente relacionado a coesão referencial e sequencial. Note que a coesão referencial é responsável por criar um sistema de relações entre as menções dentro de um texto, permitindo que o leitor identifique termos e expressões que remetem a uma mesma entidade. Junto a isso temos também a coesão sequencial, responsável por criar condições que auxiliam na progressão textual. De forma geral, as flexões de tempo, de modo dos verbos e as conjunções são mecanismos responsáveis pela coesão sequencial e auxiliam na coesão referencial.

Esses fatores inferem diretamente nas dificuldades existentes na tarefa de resolução de correferência, dado que estão relacionados diretamente a questões linguísticas e a habilidades cognitivas humanas complexas, de difícil reprodução por sistemas computacionais. Em poucas palavras, o desafio é: como inferir computacionalmente que a palavra “agrônomo”, que está sendo citada dois parágrafos abaixo da expressão “o agrônomo Miguel Guerra”, está se referindo a esta entidade e não a uma outra?

Portanto, o conjunto dessas expressões referenciais relativas a uma mesma entidade de mundo denomina-se **cadeia de correferência**. Esse conjunto é responsável pela construção coesa de um texto e por isso sua importância, já que a coesão é responsável pela compreensão textual. No exemplo acima, podemos afirmar que “Miguel Guerra” é o antecedente e “Guerra” é a anáfora. Dessa forma, expressões correferentes fazem referência à mesma entidade, enquanto expressões anafóricas e catafóricas podem retomar uma referência ou ativar um novo referente. Anáfora e catáfora pressupõem um par ordenado, enquanto que a correferência remete à ideia de conjunto [50].

2.3 Referências Endofóricas e Exofóricas

Quando lidamos com relações referenciais é importante deixarmos claro que estas podem configurar-se em dois tipos: endofóricas e exofóricas. Referências endofóricas são aquelas que antecedem ou sucedem informação dentro de um texto. Essas comumente

ocorrem na forma de anáforas ou catáforas, conforme visto nos exemplos anteriores. Já relações exofóricas referem-se a relações que ocorrem fora de um dado texto e necessitam de um prévio conhecimento de mundo, local ou momento para serem identificadas, como em:

x) O Bruxo do Cosme Velho foi homenageado em nossa cidade.

Note que dentro do texto não existem referências para os termos “Bruxo do Cosme Velho” e “nossa cidade”. É necessário recorrermos ao conhecimento de mundo para inferirmos que “Bruxo do Cosme Velho” refere-se a Machado de Assis. Da mesma forma a referência do termo “nossa cidade” não está no texto, mas pode estar na memória do leitor ou na memória do escritor. Nesta tese damos foco às referências endofóricas. Ou seja, referências que ocorrem dentro do texto.

2.4 Coerência e Coesão

Quando lidamos com a resolução de correferência existem características muito importantes a serem consideradas. Características que geralmente ficam implícitas em textos bem escritos e estruturados, mas que merecem atenção, dado que inferem diretamente na obtenção de bons resultados. Dentro desse contexto temos a coerência e a coesão textual. De acordo com Koch et al. [38] a coerência textual é algo que tem a ver com a boa formação do texto, não em um sentido gramatical, mas sim em nível de interlocução. A coerência é algo que se estabelece na interação, na interlocução ou em uma situação comunicativa entre duas pessoas. Em poucas palavras, a coerência é o que faz com que o texto tenha sentido, devendo ser vista como um princípio de interpretabilidade do texto e também com a capacidade que o leitor possui para calcular seu significado. A coerência é vista também como uma continuidade de sentidos perceptíveis no texto, a qual resulta em uma conexão conceitual cognitiva entre os elementos do texto. Como podemos perceber, a coerência é, ao mesmo tempo, semântica e pragmática aplicadas, pois a forma como construímos nossas ideias pode variar, de acordo com nosso conhecimento de mundo.

Paralelamente ao conceito de coerência temos a coesão. Ao contrário da coerência, a coesão é explicitamente revelada por meio de marcas linguísticas, sendo de caráter linear, dado que manifesta-se na organização sequencial de um texto. Em poucas palavras, a coesão está muito mais ligada à sintaxe e à gramática. Note que esses conceitos são muito importantes para a tarefa de Resolução de Correferência, dado que a correferência de um termo e seu antecedente é guiada por essa construção de ideias. No Capítulo 5 veremos que as regras propostas em nosso modelo e a forma com que nosso método de agrupamento lida com as menções é uma tentativa de representar computacionalmente esses conceitos.

2.5 Métricas de Avaliação

A tarefa de resolução de correferência é complexa e envolve diferentes níveis de processamento. Logo, avaliar um modelo de correferência não é uma tarefa simples, dado que existem muitos detalhes a serem considerados, como a detecção de menções, agrupamentos realizados, agrupamentos não realizados. Na literatura encontramos cinco métricas propostas para avaliar esses modelos: MUC[72], B-CUBED[3], Ceaf$_e$, Ceaf$_m$ [41] e BLANC [57]. Cada uma dessas métricas visa avaliar uma característica específica de cada modelo. Anualmente, competições como a CoNLL[52] são realizadas, visando motivar o desenvolvimento de sistemas. Nos anos de 2011 e 2012 essas competições foram voltadas à tarefa de Resolução de Correferência. Com o objetivo de avaliar os modelos participantes por meio de uma pontuação única, a conferência propôs uma nova métrica, chamada CoNLL[51]. A métrica CoNLL consiste na média da medida-F de três outras métricas da literatura, como veremos nessa seção.

2.5.1 MUC:

Baseada em cadeias, mede quantos agrupamentos de menções são necessários para cobrir as cadeias padrão. Por exemplo, considere que o conjunto K (cadeia de referência) seja composto pelas seguintes ligações(*links*) de correferência {A–B, B–E, C–D} e que o conjunto R (cadeia predita pelo modelo) seja composto por {A–B, C–D}. Para este caso podemos ver que falta uma ligação no conjunto R. teremos então Abrangência= $\frac{2}{3} = 0,67\ (67\%)$ e Precisão = $\frac{2}{2} = 1\ (100\%)$. De forma mais geral, o cálculo da métrica MUC pode ser obtido por meio das seguintes fórmulas:

$$Abrangência = \frac{\sum_{i=1}^{N_k}(\|K_i\| - \|p(K_i)\|)}{\sum_{i=1}^{N_k}(\|K_i\| - 1}$$

$$Precisão = \frac{\sum_{i=1}^{N_r}(\|R_i\| - \|p'(R_i)\|)}{\sum_{i=1}^{N_r}(\|R_i\| - 1)}$$

Onde: K$_i$ é i-ésima entidade padrão (*key entity* – referência) e p(K$_i$) é o grupo de partições criado por meio da intersecção de K$_i$ e os *links* preditos pelo modelo; R$_i$ é a i-ésima entidade predita pelo modelo (*Response entity*) e p'(R$_i$) é o conjunto de partições

criadas por meio da intersecção de R_i e K_i. N_k e N_r representam a quantidade de menções padrão e resposta, respectivamente.

2.5.2 B-CUBED:

Baseada em menções, gera resultados considerando as menções presentes e ausentes de cada entidade em dada cadeia. Basicamente, a métrica B-Cubed atribui um peso para as menções, baseando-se na quantidade total de menções existentes. Sua abrangência e precisão são obtidas por:

$$Abrangência = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \frac{\|K_i \cap R_j\|^2}{K_i}}{\sum_{i=1}^{N_k} K_i}$$

$$Precisão = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \frac{\|K_i \cap R_j\|^2}{R_j}}{\sum_{i=1}^{N_k} R_j}$$

Onde: K representa o conjunto das *key entities* (menções padrão) e R o conjunto de menções preditas pelo modelo. Por exemplo, dadas as cadeias de referência :

C_{K1} = {A, B, C, D, E};

C_{K2} = {F, G};

C_{K3} = {H, I, J, K, L}.

E as cadeias preditas pelo modelo:

C_{R1} = {A, B, C, D, E};

C_{R2} = {F, G, H, I, J, K, L}.

Cada menção possuirá o peso de $\frac{1}{12}$, dado que o total de menções existente é 12. Dito isso temos então:

Abrangência = $\frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{2} + \frac{2}{2} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7}] = 1\ (100\%)$

Precisão = $\frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{7} + \frac{2}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7}] = \frac{16}{21} = 0,76\ (76\%)$

2.5.3 CEAF:

Baseada no alinhamento de menções e entidades, possui duas variações: CEAF$_m$ (Φ_3) e CEAF$_e$ (Φ_4).

$$\Phi_3(K, R) = \|K \cap R\|$$

$$\Phi_4(K, R) = \frac{2\|K \cap R\|}{\|K\| + \|R\|}$$

$$Abrangência = \frac{\Phi_x}{\sum_{i=1} \|K_i\|}$$

$$Precisão = \frac{\Phi_x}{\sum_{i=1} \|R_i\|}$$

Por exemplo, dadas as cadeias de referência:

C$_{K1}$ = {A, B, C, D, E};

C$_{K2}$ = {F, G}.

E as cadeias preditas pelo modelo:

C$_{R1}$ = {A, B, C, D, E}.

As métricas CEAF utilizam o alinhamento entre as entidades ou menções para calcular seus resultados, dessa forma C$_{K1}$ será alinhado com C$_{R1}$ e C$_{K2}$ não possuirá um alinhamento, dado que o modelo não obteve tal cadeia. Note que o número de menções alinhadas é 5. Portanto Φ_3 = 5. Dito isso, temos:

CEAF$_m$: Abrangência = $\frac{5}{7} = 0,71\ (71\%)$ e Precisão = $\frac{5}{5} = 1\ (100\%)$

Para CEAF$_e$, dado que Φ_4 = $\frac{2 * 5}{5 + 5} = 1$, temos:

Abrangência = $\frac{1}{2} = 0,5\ (50\%)$ e Precisão = $\frac{1}{1} = 1\ (100\%)$

Note que para a métrica CEAF$_m$ o calculo de precisão e abrangência é realizado considerando a quantidade de menções, para a métrica CEAF$_e$ esse valor é considerado em Φ_4. Contudo, para obtenção dos valores de precisão e abrangência, são usados os valores referentes a quantidade de entidades/cadeias.

2.5.4 BLANC:

BiLateral Assessment of NounPhrase Coreference avalia tanto *links* de correferência quanto os de não correferência. Basicamente, um *link* de não correferência é formado por duas menções que não são correferentes entre si. A métrica BLANC tem como objetivo recompensar as cadeias de correferência corretas, de forma proporcional ao seu tamanho. Temos, então, C$_K$ e C$_R$ respectivamente como: *links* de correferência padrão e preditos automaticamente e; N$_K$ e N$_R$ como grupo dos *links* de não correferência padrão e preditos automaticamente; Abrangência$_C$ e Precisão$_C$ remetem ao cálculo de abrangência e precisão dos *links* de correferência, e Abrangência$_N$ e Precisão$_N$, aos *links* de não correferência.

$$Abrangência_C = \frac{\|C_k \cap C_r\|}{C_k}$$

$$Precisão_C = \frac{\|C_k \cap C_r\|}{C_r}$$

$$Abrangência_N = \frac{\|N_k \cap N_r\|}{N_k}$$

$$Precisão_N = \frac{\|N_k \cap N_r\|}{N_r}$$

Por fim, a precisão e abrangência da métrica BLANC são calculadas respectivamente por meio das médias de Precisão e abrangência, obtidos entre os *links* de correferência e de não correferência:

$$BLANC_{Precisao} = \frac{Precisão_C + Precisão_N}{2}$$

$$BLANC_{Abrangencia} = \frac{Abrangência_C + Abrangência_N}{2}$$

Por exemplo: dados os seguintes links de correferência:

C$_{K1}$ ={A–B, B–C, C–D, D–E};

C$_{K2}$ ={F–G}.

E os seguintes links preditos pelo modelo:

C_{R1} = {A–B, B–C, C–D, D–E};

C_{R2} ={F–G, F–I}.

Temos então: $Abrangência_C = \frac{5}{5} = 1\ (100\%)$, $Precisão_C = \frac{5}{6} = 0,83\ (83\%)$

Considerando que os links de não correferência representam ligações entre todas as menções que não são referenciais, teremos então:

N_K = {F–A, F–B, F–C, F–D, F–E, G–A, G–B, G–C, G–D, G–E};

N_R ={F–A, F–B, F–C, F–D, F–E, G–A, G–B, G–C, G–D, G–E, I–A, I–B, I–C, I–D, I–E}.

$$Abrangência_N = \frac{10}{10} = 1\ (100\%)\ , Precisão_C = \frac{10}{15} = 0,67\ (67\%)$$
$$BLANC_{Precisao} = \frac{0,83 + 0,67}{2} = 0,75\ (75\%)$$
$$BLANC_{Abrangencia} = \frac{1+1}{2} = 1\ (100\%)$$

2.5.5 CoNLL:

Amplamente utilizada para avaliar modelos de resolução de correferência, a métrica CoNLL calcula um score único, baseando-se no cálculo da medida-f das métricas MUC, B^3 e $CEAF_e$:

$$CoNLL = \frac{(F(MUC) + F(B^3) + F(CEAF_e))}{3}$$

2.6 Definição da Tarefa de Resolução de Correferência no Contexto desta Tese

Devido a variedade de abordagens para a tarefa de resolução de correferência e os diferentes níveis de escopo, conforme será visto no Capítulo 3, definimos nesta Seção o escopo específico desta tese. Esta tese de doutorado baseou-se principalmente no trabalho de Lee et al. [39], em que um modelo baseado em regras é aplicado para a resolução de correferência em textos da língua inglesa. Os autores estudam os fenômenos linguísticos e propõem regras bem definidas que consistem em agrupar duas menções caso uma regra ou conjunto de regras seja satisfeito. Além deste, os trabalhos de [49, 67, 51, 12, 63, 60] mostraram-se relevantes para o contexto desta tese, os quais são apresentados nos Capítulos 3 e 4. A partir do trabalho de Lee et al., é proposto um processo para a resolução

de correferência entre sintagmas nominais em textos da língua portuguesa, aplicando regras que visam agrupar menções referenciais em níveis lexicais, sintático, justaposição e semântico. Nesse contexto definiu-se a tarefa de resolução de correferência da seguinte forma:

- Consideramos como menções passíveis de serem agrupadas: todo tipo de sintagmas nominais (substantivos comuns e próprios, específicos, não específicos e abstratos), com exceção de pronomes pessoais;
- Não tratamos domínios específicos, nossa abordagem objetiva resolver correferência para qualquer tipo de texto que esteja escrito em Português, desde que este esteja bem escrito e estruturado;
- Toda menção que satisfaça uma ou mais regras linguísticas possui chance de ser agrupada a outras menções, respeitando nosso critério de agrupamento proposto, contido na Seção 5;
- Em nossa abordagem consideramos dois tipos de relações semânticas: Sinonímia e Hiponímia.

2.7 Considerações do Capítulo

Este capítulo apresentou os fundamentos conceituais sobre a tarefa de Resolução de Correferência. Conforme pode-se ver, a correferência é um fenômeno que pode ocorrer de diversas formas em um discurso. Além disso, vimos também que as relações semânticas Hiponímia, Hiperomínia e Sinonímia podem indicar, em alguns casos, uma relação de correferência entre dois termos. Contudo, precisamos levar em consideração o contexto de cada menção, dado que é possível existir uma relação semântica entre dois termos sem existir uma relação de correferência, algo que ocorre com certa frequência quando lidamos com casos ambíguos, como por exemplo: (gato – bichano) e (gato – atraente). Basicamente uma relação referencial não depende apenas da semântica, mas também de seu contexto. Apresentamos também as principais métricas utilizadas pela literatura, as quais possuem como objetivo avaliar características específicas em modelos de resolução de correferências. No próximo Capítulo apresentamos uma descrição sobre os principais tipos de abordagens encontrados na literatura.

3. TRABALHOS RELACIONADOS

3.1 Abordagens Computacionais para Resolução de Correferência

Na literatura, encontramos uma grande variedade de abordagens que propõem resolver correferência em diversos idiomas, como: Inglês, Chinês, Árabe, Espanhol, Galego, Português, entre outros [64, 48, 73, 54, 18, 11, 14, 28, 21, 24, 43]. Essas abordagens, em sua maioria, são voltadas para a língua Inglesa e baseadas em aprendizado de máquina. Contudo, é possível encontrarmos alguns modelos baseados em regras linguísticas [39, 31, 35]. Veremos que, diferente dos modelos baseados em regras, o aprendizado de máquina pode se ramificar em diferentes propostas, como *Mention-Pair*, *Entity-Mention*, *Mention-Ranking* e *Antecedent-Trees*.

3.1.1 Modelos Baseados em Machine Learning

Mention Pair

A primeira delas, e a mais popular, provavelmente por sua simplicidade, é a abordagem baseada em pares de menções. Basicamente, modelos que lidam com essa abordagem optam por efetuar seu treino por meio de pares de menções, de forma a determinar se duas menções são correferentes ou não. Os modelos baseados em pares de menções têm influenciado significativamente os trabalhos que propõem a resolução de correferência utilizando técnicas de *Machine Learning* nos últimos dezesseis anos [64]. Modelos baseados em pares de menções visam responder se devem ou não classificar como correferente uma menção m_j com um candidato antecedente m_i. Inicialmente, para treinar um modelo baseado em pares, é necessário extrair características/*features* que possibilitem obter alguma informação proveniente da comparação entre m_i e m_j. Na Tabela 3.1, podemos observar as *features* mais utilizadas pela literatura.

Um dos grandes desafios ao utilizar uma abordagem baseada em pares de menções se dá no desbalanceamento entre as classes positiva (pares correferentes) e negativa (pares não correferentes). Ou seja, todo modelo requer, além de amostras positivas, amostras de pares negativos. Dada essa premissa, é necessário realizar a construção de pares. Nessa etapa, ao cruzarmos essas menções, consequentemente teremos muito mais amostras negativas do que positivas. Objetivando minimizar esse desbalanceamento entre as classes, alguns trabalhos propõem diferentes técnicas para geração de pares.

Soon et al. [64] realizam um pareamento distinto para cada uma das classes: para os pares positivos, dado o conjunto de menções C={m_i,m_j,m_k, m_l} (todas correferentes

Tabela 3.1 – *Features* mais comuns na literatura.

Feature	**Descrição**
Casamento de Padrões	Se m_i e m_j são lexicalmente iguais.
Casamento de Núcleos	Se m_i e m_j possuem o mesmo núcleo.
Alias	Se m_j é sigla de m_i ou vice-versa.
I_Pronome	Se m_i é um pronome.
J_Pronome	Se m_j é um pronome.
Número	Se m_i e m_j concordam em número (singular/plural).
Gênero	Se m_i e m_j concordam em gênero (masculino/feminino).
Nome Próprio	Se m_i e m_j são nomes próprios.
Aposto	Se m_j é aposto de m_i
Distância entre Sentenças	Distância em sentenças entre mj e m_i.
Distância entre Sintagmas	Distância em menções à entre mj e m_i.
Classe Semântica	Se m_i e m_j possuem mesma classe semântica.
Hiponímia	Se m_i e m_j possuem uma relação de hiponímia.
Hiperonímia	Se m_i e m_j possuem uma relação de hiperonímia.
Sinonímia	Se m_i e m_j possuem uma relação de sinonímia.

entre si), apenas as menções imediatamente adjacentes formam pares (Tabela 3.2): P_p= {(m_i,m_j), (m_j,m_k), (m_k,m_l)}. Para gerar os pares negativos, considere o conjunto de menções M={m_m, m_n, m_o, m_p, m_q} em que apenas as menções m_m e m_q são correferentes. Dentro desse contexto, a última menção deste conjunto, m_q, faz par com todas as anteriores, exceto com m_m: P_n={ (m_q,m_p), (m_q,m_o) e (m_q,m_n)} (Tabela 3.3).

Note que na Tabela 3.2, o conjunto de menções considerado é uma cadeia de correferência. Já na Tabela 3.3, o conjunto de menções não consiste de uma cadeia. Apenas as menções m_m e m_q são correferentes. Logo, não formam par.

Tabela 3.2 – Geração de Pares Positivos proposta por Soon et al.

	Pareamento de amostras Positivas
Conjunto de menções	m_i,m_j,m_k, m_l
Pares	m_i, m_j m_j, m_k m_k,m_l

Tabela 3.3 – Geração de Pares Negativos proposta por Soon et al.

	Pareamento de amostras Negativas
Conjunto de menções	m_m,m_n,m_o,m_p, m_q
Pares	m_q, m_p m_q, m_o m_q,m_n

Martschat et al. [43] propõem uma mesma metodologia para geração de pares positivos e negativos: dado documento D_x, que possua um conjunto de menções M={m_i, m_j,m_k, m_l, m_m, m_n} onde apenas m_l e m_i são correferentes, o conjunto de pares (positivos e negativos) será: P= {(m_l,m_k), (m_l,m_j), (m_l,m_i)}. Basicamente, para cada par correferente (m_x,m_y), a geração de amostras negativas será realizada com as menções entre (m_x,m_y). Note que essa construção é efetiva pelo fato de não gerar uma grande quantidade de amostras negativas. No entanto, devido a esta restrição, pode-se perder pares negativos que possuam informações relevantes.

Tabela 3.4 – Geração de Pares Positivos e Negativos proposta por Martschat et al.

	Pareamento de amostras Positivas e Negativas
Conjunto de menções	m_i,m_j,m_k,m_l, m_m, m_n
Pares	m_l, m_k m_l, m_j m_l, m_i

Em Fonseca et al. [24], para um dado conjunto de menções M={m_i,m_j,m_k}, temos: P={(m_i,m_j), (m_i,m_k), (m_j,m_k)}. Basicamente, cada menção faz par com a próxima, independente de esta ser correferente ou não. Note que a quantidade de pares será muito maior que em [43]. Para minimizar o desbalanceamento entre as classes foi utilizado *random undersampling*, que consiste na seleção aleatória de *n* pares negativos, em que *n* é a quantidade de pares positivos. Por meio de experimentos, foi visto que os níveis de balanceamento “1 para 1” (um par positivo para cada par negativo) e “1 para 2” (um par positivo para cada dois pares negativos) foram os que obtiveram melhores resultados.

Tabela 3.5 – Geração de Pares Positivos e Negativos proposta por Fonseca et al.

	Pareamento de amostras Positivas e Negativas
Conjunto de menções	m_i,m_j,m_k
Pares	m_i, m_j m_i, m_k m_j, m_k

Entity-Mention

Diferente do tradicional *Mention Pair*, o *Entity-Mention* [73] explora a propriedade de representação do discurso, tendo em vista o conhecimento de quando uma entidade é nova no discurso ou anafórica (semelhante à nossa metodologia de agrupamento proposta). Para conceber os pares, assume-se que uma instância de treino positiva consiste em i{e_x,m_y}, na qual m_y é uma menção ativa e e_x é uma entidade parcial, encontrada antes

de m_y. Para cada menção anafórica m_y, uma única instância de treinamento positivo é criada para a entidade parcial a qual m_y pertence. Para os pares negativos, é criado um grupo de instâncias para cada entidade cuja última menção ocorra entre m_y e o antecedente mais próximo de m_y. Por exemplo: considere o conjunto de menções M={e_i,m_j,e_k,m_l,m_m,m_n}. Assumindo que neste conjunto tenhamos duas cadeias: C1={e_i,m_j, m_m} e C2={e_k,m_n} e m_n é a menção ativa. Teremos então, como conjunto dos pares positivos, P_p={(m_n,e_k)} e como conjunto de pares negativos P_n={(m_n,m_m), (m_n,m_j), (m_n,e_i), (m_n,m_l)}. Basicamente, assumindo que m_n representa a menção ativa e e_k representa seu antecedente, note que temos duas menções entre elas (m_l e m_m). Nesse caso, toda menção ou cadeia pertencente a m_l e m_m forma par com a menção ativa. Nas Tabelas 3.6 e 3.7 temos os pares gerados, considerando m_n como menção ativa. Note que a cada iteração a menção ativa será outra e com isso novos pares serão gerados, sempre utilizando o mesmo critério.

Tabela 3.6 – Geração de Pares Positivos Proposta por Yang et al.

	Pareamento de Amostras Positivas
Conjunto de Menções	e_i, m_j, e_k, m_l, m_m, m_n
Pares	m_n, e_k

Tabela 3.7 – Geração de Pares Negativos Proposta por Yang et al.

	Pareamento de Amostras Negativas
Conjunto de Menções	e_i, m_j, e_k, m_l, m_m, m_n
Pares	m_n, m_m m_n, m_j m_n, e_i m_n, m_l

Outro diferencial deste modelo focado em entidades está na forma de representar suas *features*: os autores definem três tipos de instâncias, que representam como as menções se relacionam: "*link*(e_x,m_y)", onde m_y representa uma menção ativa e e_x representa uma entidade parcial; *has_mention*(e,m), descrevendo todas as menções as quais determinada menção está ligada. Por exemplo, para a cadeia previamente mencionada, C1={e_i,m_j, m_m}, teremos então *has_mention*(e_i,m_j), *has_mention*(e_i,m_m); e,o último, denota as características de cada par de menções, seguindo a seguinte estrutura: nome_da_feature(m_x,m_y, 0), representando respectivamente: o nome da feature, o par de menções e um valor binário (0 para falso e 1 para verdadeiro).

Mention-Ranking

No *Mention-Ranking model*, assim como o *Mention pair*, cada instância de treino i(m_x,m_y) representa m_y e sua menção precedente m_x. Basicamente, as *features* que representam uma instância e um método para criar uma instância de treino são idênticas às utilizadas pelo *Mention Pair model*. A diferença reside em rotular as instâncias de treino, assumindo que I_y é um conjunto de instâncias de treino, criadas para a menção anafórica m_y, o *rank* para i(m_x,m_y) em I_y é o rank de m_y entre os candidatos antecedentes, que é 2 se m_x é o antecedente mais próximo de m_y ou 1 caso contrário. Em poucas palavras, o antecedente mais próximo de sua anáfora recebe um ranking maior em relação as demais menções. Considere o seguinte conjunto de menções M={m_i, m_j, m_k, m_l, m_m, m_n}, contendo as seguintes cadeias C1={m_i, m_k, m_n,} C2={m_j, m_l,}. Note que, para m_n, teremos as seguintes instâncias:

Tabela 3.8 – Instâncias de treino geradas para m_n

$\mathbf{I_n}$		***Ranking***
Pares	m_m, m_n	1
	m_l, m_n	1
	m_k, m_n	2

Note que m_i não faz par com m_n, mas sim com m_k, dado que m_i é antecedente de m_k. Dado que m_k é o antecedente mais próximo de m_n, o par recebe ranking 2. Já os demais pares (considerando m_n) recebem valor 1. Mesmo o *Mention-Ranking* não sendo muito popular, seus resultados são superiores às abordagens baseadas em pares de menções, como podemos ver em [55] e [43].

Antecedent-Tree

Na proposta de Fernandes et al. [18], também baseada em pares de menções, os autores propõem um conjunto de regras, as quais objetivam reduzir a quantidade de pares menos propensos a serem correferentes. Assim, para um dado par de menções, caso pelo menos uma das regras da tabela 3.9 seja satisfeita, o par é considerado válido para utilizar em seu treinamento (seja ele um par positivo ou negativo). No Capítulo 5 veremos que o conjunto de regras utilizadas para seleção das amostras de treino, proposto por Fernandes et al. possui semelhança às regras de nosso modelo. Contudo, utilizamos regras menos abrangentes e mais precisas, dado que nossas regras, quando aplicadas, objetivam estabelecer uma relação de correferência entre duas menções, não apenas criar uma amostra de treino.

Tabela 3.9 – Conjunto de regras para seleção de pares proposto por Fernandes et al.

Regra	**Descrição – Considera um par como válido se:**
Distância	a quantidade de menções entre m_i e m_j não ultrapassar um determinado *threshold*.
Classe Semântica	m_i e m_j possuírem mesma classe semântica.
Combinação de Núcleos	o núcleo de m_i combinar com o núcleo de m_j.
Concordância em atributos de discurso	os atributos de discurso combinam para m_i e m_j. Esta regra consiste de um conjunto de regras proposto por [39], o qual baseia-se em atributos de uma menção e seu falante.
Pronome J	m_j for um pronome e m_i concordar em gênero, número ou fala.
Pronome e Entidade Nomeada	m_j for um pronome e m_i for um pronome compatível ou uma entidade nomeada.

Referente ao motivo dos autores nomearem sua abordagem como *Antecedent-Tree model*, reside na forma de representar o agrupamento de suas menções: para representar o agrupamento de menções correferentes entre si são utilizadas estruturas chamadas de árvores. Uma árvore de correferência é uma árvore cujos nós são dirigidos às menções e os arcos representam alguma relação entre elas. Basicamente, para cada documento é gerado um conjunto de arvores e sub-árvores, em que cada sub-árvore representa uma menção e seus referentes. Ou seja, cada anáfora pode ser considerada uma raiz ou nodo-pai e seus antecedentes podem ser considerados nodos-filhos.

Note que cada abordagem possui uma forma distinta para concepção de suas amostras de treino, assim como que para representar suas estruturas de agrupamento. Martschat et al. [43] propuseram uma forma unificada de representar tais estruturas, os autores a chamam de estrutura latente. Basicamente, uma estrutura latente é representada por um conjunto de *arrays* “V”, “A” e “L”. Analisando-a, podemos verificar que uma estrutura latente pode ser abstraída à forma de um grafo, o qual “V” representa um conjunto de nós/menções; “A” representa o conjunto de arestas, e “L”(label), representa um sinal, positivo ou negativo, informando se dada menção é correferente de outra.

Em nossa abordagem utilizamos uma estrutura muito similar a esta proposta por Matschat el al [43], mas, em vez de aprendizado de máquina, utilizamos regras. Adicionalmente em nossa abordagem propomos um novo método para o agrupamento de menções, o qual consiste em decidir se uma menção m$_j$ é referente de um antecedente m$_i$ ou se é nova no discurso. Com nosso método proposto conseguimos melhorar a precisão de nosso modelo em aproximadamente 10%, como veremos no Capítulo 5.

3.2 Modelos Baseados em Regras

Diferente dos modelos baseados em Aprendizado de Máquina, modelos baseados em regras consistem em uma série de passos que definem se duas menções são correferentes entre si. Abordagens baseadas em regras requerem um conhecimento prévio mais aprofundado referente ao idioma e ao domínio a serem tratados. Por exemplo, ao lidarmos com aprendizado de máquina, durante a implementação e seleção de *features*, caso sejam selecionadas *features* irrelevantes, a maioria dos algoritmos de treino consegue detectar e desconsiderar tal característica. Já em abordagens voltadas à regras, não temos essa flexibilidade. Cada regra deve ser elaborada cuidadosamente, pois não temos um modelo estatístico como apoio. Outra característica forte desses modelos é a forma como as menções são agrupadas. Ou seja, em abordagens baseadas em regras não existe etapa de treinamento: definidas as regras, uma menção m_j é comparada com todas as menções que a antecedem e, caso alguma das regras seja satisfeita, essas menções são agrupadas. Esse tipo de método é o mais utilizado pelos modelos de regras atuais [39, 31]. Contudo, embora nossa proposta lide com regras, nesta tese propomos uma nova metodologia de agrupamento de menções, a qual exploramos a propriedade de discurso, semelhante ao trabalho de [73].

Um dos principais modelos de regras contido na literatura foi proposto por Lee et al. em 2011 [40]. Denominado *Stanford Multi-Pass Sieve*, é um sistema para a resolução de correferência puramente baseado em regras linguísticas. Seu modelo possui dez *Sieves*/filtros, cujo objetivo é agrupar menções correferentes, caso cada regra ou conjunto de regras sejam satisfeitos. O modelo de Lee et al. foi proposto para o inglês, durante a CoNLL [1] (Conference on Natural Language Learning), ficando em primeiro colocado no ranking de melhores modelos. Os modelos foram avaliados por meio do corpus Ontonotes [53], em conjunto do CoNLL Scorer [51]. Alguns anos após surgiram outras abordagens semelhantes com o mesmo propósito, como o trabalho de Garcia et al. [31], voltado ao Português, Espanhol e Galego. Embora abordagens baseadas em regras possam ser de custoso planejamento, dado que cada idioma possui suas características, estas podem provar-se eficazes e competitivas, principalmente quando há carência por corpora anotados.

3.3 Semântica Aplicada à Resolução de Correferência

Conforme visto na seção anterior, na literatura encontramos muitos trabalhos voltados à resolução de correferência. Em sua grande maioria esses trabalhos fazem um uso mais restrito da semântica, focando em categorias de entidades nomeadas e deixando de

[1] http://conll.cemantix.org/2011/

lado relações importantes, que poderiam trazer ganhos à tarefa. Nesta Seção, relatamos os principais trabalhos voltados à resolução de correferência para os idiomas Português e Inglês. Veremos que os níveis de semântica utilizados variam de acordo com o escopo e idioma de cada trabalho.

O trabalho de Lee et al. [39], para a língua inglesa, faz uso de semântica para identificar menções que remetem a entidades do tipo 'Pessoa', objetivando resolver correferência pronominal. Isto é, os autores utilizam semântica de forma mais simples, fazendo uso de apenas uma categoria de entidade, sem explorar quaisquer outras possíveis relações semânticas. Existem trabalhos que fazem um uso mais elaborado da semântica, como o de Rahman et al. [54]. Rahman et al. avaliaram a utilidade do conhecimento de mundo usando duas bases de conhecimento: Yago [66] e FrameNet [4]. Utilizando os recursos citados, os autores fazem a identificação de relações semânticas como: "*Means*" (significa) e "*Type*" (tipo de). Cada relação semântica é representada por uma tripla *(AlbertEinsten, Type, physicist)*. Essa instância denota o fato de que Albert Einstein é um físico. A relação *"Means"*, análoga à sinonímia, provê as diferentes formas de expressar uma entidade. Portanto, permite tratar casos ambíguos, como: *(Einstein, Means, AlbertEinstein)* e *(Einstein, Means, AlfredEinstein)*, pois denotam o fato de que "Einstein" pode referir-se ao físico Albert Einstein e ao músico Alfred Einstein. Do FrameNet foram utilizados os papéis semânticos dos verbos, como por exemplo:

- Peter Anthony condena o programa de negociação, limitando o jogo para alguns, mas ele não tem certeza se quer denunciá-lo, porque...

Note que o papel semântico pode ajudar a estabelecer um *link* de correferência entre "programa de negociação" e o pronome pessoal oblíquo "lo", uma vez que com o FrameNet é possível recuperar a relação entre "condena" e "denuncia", pelo fato dessas duas palavras aparecerem no mesmo *frame* e os sintagmas estarem adjacentes a estes. Como resultado, os autores constataram que a semântica pode prover pequenos ganhos para a tarefa de resolução de correferência e, mesmo que pequenos, se acumulados, podem tornar-se algo substancial.

Hou et al. [35] propôs um modelo baseado em regras para a resolução de anáfora direta e indireta (*bridging*). Diferente da tarefa de resolução de correferências, a qual busca por relações de identidade, a resolução de anáfora indireta consiste em reconhecer e criar um elo entre duas menções por meio de uma relação de "não identidade". Um bom exemplo de tal relação é a meronímia (parte de), como em: "a casa" e "a chaminé". Para identificar tais relações, os autores utilizaram o WordNet [46]. Em nossa abordagem proposta focamos nas relações do tipo identidade e utilizamos as relações semânticas de sinonímia e hiponímia, contidas no Onto.PT[49].

3.3.1 Modelos Voltados à Língua Portuguesa

Para a língua portuguesa, Silva [62], propôs um modelo para a resolução de correferência utilizando o conjunto de etiquetas semânticas providas pelo corpus do HAREM [29]. Para detectar tais categorias, Silva utilizou o parser PALAVRAS [7] e o reconhecedor de entidades nomeadas Rembrandt [10]. Como base de conhecimento semântico, o autor utilizou o TEP2.0 [45],um *thesaurus* contendo relações de sinonímia e antonímia para a língua portuguesa. Coreixas [14] propôs a resolução de correferência, focando-se nas categorias 'Pessoa', 'Local', 'Organização', 'Acontecimento', 'Obra', 'Coisa' e 'Outro'. Como recursos, foram utilizados o corpus do HAREM [29], o parser Palavras e o corpus Summ-it.

Em nossa abordagem também utilizamos o corpus Summ-it[12]. Contudo, o utilizamos apenas para estudo de caso. Referente ao parser utilizado, optamos pelo CoGrOO [63], um parser de código aberto, desenvolvido para o Português. Para validação dos resultados utilizamos os corpora Summ-it++[2] e Corref-PT[20]. Para o tratamento de entidades, optamos por recursos distintos: Repentino [60] e listas genéricas de entidades nomeadas [21]. O motivo desta escolha se dá pela intenção de tratarmos, inclusive, nomes comuns. Esses recursos são descritos no Capítulo 4.

De forma a demonstrar que o uso de categorias semânticas pode auxiliar na tarefa de resolução de correferência, Coreixas [14] compara duas versões de seu sistema: a primeira, sem fazer o uso de categorias semânticas; e a segunda, fazendo uso dessas categorias. Como resultado, a autora mostrou que o uso de categorias pode prover melhorias significativas, dado que seu uso pode auxiliar a determinar se dado par de menções é correferente ou não. Em nossa abordagem utilizamos um método contrário. Ou seja, utilizamos a informação de correferência para inferir categorias de entidades à menções não classificadas no processo de REN (Reconhecimento de Entidades Nomeadas) , melhor descrito na Seção 4.4.3.

A autora também mostrou a importância do conhecimento de mundo para esta linha de pesquisa. Garcia e Gamallo [31] propõem um modelo baseado em regras (semelhante ao de Lee et al. [39], mas para múltiplos idiomas (Português, Espanhol e Galego). Em seu trabalho, os autores focam apenas na categoria semântica 'Pessoa'.

Em trabalhos anteriores [21] propusemos uma abordagem baseada em aprendizado de máquina, com foco em nomes próprios e nas categorias de entidades 'Pessoa', 'Local' e 'Organização'. Para detectar as entidades, utilizamos o Repentino [60] e NERP-CRF [15]. Adicionalmente, para casos mais genéricos de entidades, utilizamos listas, contendo substantivos comuns, que remetem a determinadas entidades, tais como: [advogado, agrônomo, juiz] para a categoria 'Pessoa', e [avenida, rua, praça, cidade] para 'Local'.

Em trabalhos recentes [27] introduzimos conhecimento semântico por meio do Onto.PT [33], de forma a estudar o impacto que a semântica pode ter no âmbito da lín-

gua Portuguesa. Esse trabalho remete à continuação de estudos prévios [24], realizados, com o propósito de avaliar o impacto do desbalanceamento entre as classes na concepção de modelos de resolução de correferência. Como fruto desses estudos pudemos concluir a relevância de muitas regras propostas, bem como os ganhos que a semântica pode prover à tarefa no âmbito da língua Portuguesa, dado que o modelo treinado com semântica obteve uma melhoria de 0.6% em precisão, 3.4% em abrangência e 2.4% em medida-F, para a classe positiva (correferente). Nesses experimentos consideramos as categorias semânticas providas pelo PALAVRAS [7].

De forma geral, modelos baseados em aprendizado de máquina, para serem eficazes, requerem uma grande quantidade de dados anotados, o que tomamos como limitação para o Português, dado que o corpus referência [12] utilizado nesse experimento possui apenas 50 textos anotados. Essa foi a principal motivação para conduzirmos uma abordagem baseada em regras. Inicialmente avaliamos a eficácia de nosso modelo de regras [25]. Nossa avaliação objetivou verificar valores de precisão, abrangência e medida-F para cada categoria de entidade nomeada. Para tal, utilizamos a coleção dourada do segundo HAREM[29] e, como resultado, nosso modelo obteve 80.8% de precisão, 62.3% de abrangência e 70% de medida-F.

3.4 Avaliação da Tarefa de Resolução de Correferência

Nas tabelas 3.10 e 3.11 mostramos os principais resultados da conferência CoNLL em 2011 e 2012, os dois anos em que a conferência promoveu a tarefa de Resolução de Correferência. “P”, “A” e “F” representam respectivamente Precisão, Abrangência e Medida-F. Na Tabela 3.11 apresentamos um resumo dos resultados da conferência de 2012. “Aberto” remete para a competição em que é permitido aos participantes utilizarem recursos externos; “Fechado” remete à competição em que o uso de recursos é restrito aos dados fornecidos pelos organizadores. A coluna “CoNLL Score” denota a média da métrica CoNLL na modalidade “Fechado”, para os três idiomas da tarefa (Inglês, Chinês e Árabe). Resultados completos estão contidos em [53, 52].

Tabela 3.10 – Resultados oficiais da CoNLL 2011 [53]

Modelo	Detecção de Menções			MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CoNLL
	P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
Lee	66,8	75,1	70,1	57,5	61,8	59,6	68,2	68,4	68,3	56,4	56,4	56,4	47,7	43,4	45,5	76,2	70,6	73,0	57,8
Sapena	28,2	92,4	43,2	63,7	56,3	59,5	72,1	62,7	67,1	53,5	53,5	53,5	38,4	44,7	41,3	73,1	69,5	71,1	56,0
Chang	62,0	68,1	64,9	57,1	57,1	57,1	70,5	67,1	68,8	54,4	54,4	54,4	41,9	41,9	41,9	77,1	71,2	73,7	56,0
Nuges	68,1	69,9	69,0	57,1	60,2	58,6	64,2	66,7	65,5	51,4	51,4	51,4	41,1	38,1	39,5	70,3	72,0	71,1	54,5
Santos	63,2	67,8	65,4	54,3	59,2	56,6	62,8	68,8	65,7	49,5	49,5	49,5	40,2	35,9	37,9	66,9	73,4	69,5	53,4
Song	80,4	57,8	67,3	67,8	53,7	59,9	66,0	60,6	63,2	46,3	46,3	46,3	30,7	43,4	36,0	59,7	69,5	61,5	53,0
Stoyanov	65,0	70,8	67,8	54,0	63,6	58,4	53,3	72,6	61,4	46,1	46,1	46,1	40,8	32,0	35,9	58,9	73,2	60,9	51,9
Sobha	62,1	67,8	64,8	49,9	51,1	50,5	65,4	62,6	64,0	49,5	49,5	49,5	41,8	40,6	41,2	68,3	61,4	63,9	51,9
Kobdani	60,0	62,1	61,0	51,5	55,6	53,5	62,4	69,7	65,8	42,7	42,7	42,7	35,4	32,3	33,8	63,5	61,9	62,6	51,0
Zhou	63,6	61,1	62,3	52,8	45,6	49,0	72,9	57,1	64,1	47,5	47,5	47,5	36,8	43,2	39,7	73,9	61,1	64,7	50,9
Charton	62,8	65,9	64,3	50,0	55,1	52,4	58,4	66,3	62,1	46,8	46,8	46,8	39,0	34,3	36,5	62,2	69,9	64,8	50,4
Yang	57,5	71,9	63,9	46,4	59,9	52,3	55,1	71,6	62,3	46,5	46,5	46,5	42,4	30,3	35,3	61,7	71,1	64,6	50,0
Hao	64,1	64,5	64,3	51,4	57,9	54,5	55,4	67,8	61,0	45,1	45,1	45,1	35,8	30,1	32,7	62,4	72,6	65,3	49,4
Xinxin	58,7	65,5	61,9	44,8	48,5	46,6	62,3	61,6	61,9	44,7	44,7	44,7	38,6	35,2	36,8	68,8	63,0	64,3	48,5
Zhang	68,25	55,3	61,1	55,6	42,0	47,9	73,0	52,6	61,1	44,5	44,5	44,5	30,3	42,0	35,2	69,2	62,8	65,2	48,1
Kummerfeld	57,0	69,8	62,7	39,6	46,4	42,7	57,3	63,6	60,3	45,3	45,3	45,3	42,3	35,0	38,3	61,6	58,7	59,9	47,1
Zhekova	37,6	67,5	48,3	20,7	28,9	24,1	56,7	67,1	61,5	40,4	40,4	40,4	41,2	31,6	35,7	57,0	52,8	53,8	40,4
Irwin	61,1	17,1	26,7	50,60	12,4	20,0	89,9	35,1	50,5	31,7	31,7	31,7	17,4	45,8	25,2	56,8	51,5	51,1	31,9

Tabela 3.11 – Resultados oficiais CoNLL 2012 [52]

Participante	Aberto			Fechado			CoNLL
	In	Ch	Ar	In	Ch	Ar	Score
Fernandes				63,4	58,5	54,2	58,7
Chen		63,5		59,7	62,2	47,1	56,3
Stamborg				59,4	56,8	49,4	55,2
Uryupina				56,1	53,9	50,4	53,5
Zhecova				48,7	44,5	40,6	44,6
Li				45,8	46,3	33,5	41,9
Yuan		61,0		58,7	60,7		39,8
Xu				57,5	59,2		38,9
Martschat				61,3	53,1		38,1
Chunyang				59,2	51,8		37,0
yang				55,3			18,4
Chang				60,2	45,7		35,3
Xinxin				48,8	51,8		33,5
Shou				58,25			19,4
Xiong	59,2	44,3	44,4				0,0

3.4.1 Características dos principais trabalhos relacionados

Na Tabela 3.12, temos um resumo em relação às características dos principais trabalhos citados neste Capítulo. No que diz respeito a semântica, alguns autores utilizaram categorias de entidades nomeadas ou recursos como Yago [66], WordNet [46], FrameNet [4], Tep2.0 [45], entre outros. As relações semânticas comumente utilizadas são: Sinonímia, Hiponímia, Hiperonímia, Meronímia. A coluna “Clustering” denota trabalhos que fazem uso de técnicas de agrupamento por meio de aprendizado de máquina [48, 54, 62]; a coluna “Rule-Based”, denota que o modelo é baseado em regras [39, 35, 31, 25]. Alguns trabalhos possuem mais de uma característica, como *Mention pair*, *Mention Ranking*, entre outras. Isso ocorre pelo fato desses trabalhos realizarem experimentos com mais de um modelo e testarem seus diferentes resultados, como em [43].

Tabela 3.12 – Características dos principais trabalhos relacionados

Modelo	Características						Idiomas						Semântica	
	Mention Pair	Entity Mention	Mention Ranking	Antec. Tree	Clus-tering	Rule Based	In	Ch	Ar	Pt	Es	Gl	Categoria de Entidade	Recursos Semânticos
Martschat et al. (2015) [43]	√		√	√			√						√	
Hou et al. (2014) [35]						√	√						√	√
Yang et al. (2008) [73]	√	√					√						√	
Fernandes et al. (2014) [18]	√			√			√	√	√				√	
Ng et al. (2002) [48]	√				√		√						√	
Lee et al. (2013) [39]						√	√						√	
Soon et al. (2001) [64]	√						√						√	
Rahman et al. (2011) [54]	√		√		√		√						√	√
Chang et al. (2012) [11]	√						√						√	
Garcia et al. (2014) [31]						√				√	√	√	√	
Silva (2011) [62]					√					√			√	√
Coreixas (2010) [14]	√									√			√	
Fonseca et al. (2015) [24]	√									√			√	
Fonseca et al. (2016) [27]	√									√			√	√
Fonseca et al. (2016) [25]						√				√			√	
Abordagem Proposta [23]						√				√			√	√

3.5 Considerações do Capítulo

Neste capítulo mostramos os principais modelos para a resolução de correferência encontrados na literatura, envolvendo diferentes abordagens e escopos. Pudemos notar que muitos dos trabalhos apresentados focam-se no uso de categorias de entidades nomeadas, mas poucos preocupam-se em fazer um uso mais elaborado do conhecimento semântico. Podemos notar também, que existe uma grande tendência em abordagens voltadas para a língua Inglesa e aprendizado de máquina, tendo como base modelos baseados em pares de menções. Não é difícil entender a motivação de tais trabalhos, dado que a concepção desses modelos torna-se uma tarefa muito menos onerosa. Contudo, para obtenção de bons resultados, uma grande quantidade de dados anotados se faz necessária; algo que muitos idiomas ainda não possuem. Para o Português existem muitos trabalhos que seguiram essa tendência [14, 21, 24, 27], mas é possível notarmos que seus resultados limitam-se a modelos pouco funcionais ou com resultados bem inferiores, quando comparamos com trabalhos voltados ao Inglês. Dito isso, acreditamos que o caminho para concepção de modelos iniciais eficientes para o Português não encontra-se no aprendizado de máquina, mas sim no estudo de padrões linguísticos e na elaboração de regras, que permitam considerar a semântica a coesão textual, o que veremos no Capítulo 5.

No Capítulo 4 apresentamos os principais recursos utilizados para a concepção do modelo proposto, bem como os recursos gerados durante esta tese de doutoramento.

4. RECURSOS DE PESQUISA

Neste capítulo, apresentamos os recursos utilizados nesta pesquisa para estudo de caso e concepção desta tese. Alguns desses recursos foram concebidos no decorrer deste trabalho, tais como: CorrefVisual [61], Summ-it++[2], Corref-PT [20] e CORP [23] (apresentado na Seção 5.5).

4.1 CoGrOO

CoGrOO [63] é um corretor gramatical de código aberto em uso por milhares de usuários de uma suíte de escritório de código aberto. Ele é capaz de identificar erros como: colocação pronominal, concordância nominal, concordância sujeito-verbo, uso da crase, concordância nominal e verbal e outros erros comuns de escrita em português do Brasil. Para tal, o CoGrOO realiza uma análise híbrida: inicialmente, o texto é anotado usando técnicas de PLN e, em seguida, um sistema baseado em regras é responsável por identificar os possíveis erros gramaticais.

Além das funcionalidades já descritas, o CoGrOO possui outras duas funcionalidades que até o momento poucos recursos disponíveis para o português possuem. O CoGrOO, da mesma forma que o OGMA [42] e o PALAVRAS [7], possibilita a anotação de sintagmas nominais, com o diferencial de ser portável a qualquer sistema operacional, pelo fato de ter sido desenvolvido em JAVA. Além disso, o CoGrOO faz também a anotação morfológica de palavras. Na Figura 4.1, é possível visualizarmos um exemplo de anotação para a sentença “ França é o único país que se recusa a aceitar a determinação européia.”, retirada do texto “CIENCIA_2000_6381 ” do corpus Summ-it[12].

Na Figura 4.1 temos:

- Tokens, contendo as palavras de entrada em sua forma original, sua forma canônica (lemma), sua classe gramatical e análise morfológica.
- Chunks, contendo os sintagmas.
- Shallow Structure, contendo a análise sintática da sentença.

Na Tabela 4.1, temos a definição das *tags* utilizadas pelo CoGrOO nas tarefas de *Part-of-speech tagging*, *Chunking* e *Shallow parsing*. Estas são as mesmas utilizadas por nosso modelo (Capítulo 5).

Figura 4.1 – Exemplo de anotação fornecida pelo CoGrOO

```
Sentence: A França é o único país que se recusa a aceitar a determinação européia.
  Tokens:
    A           [o]           art     F=S
    França      [França]      prop    F=S
    é           [ser]         v-fin   PR=3S=IND
    o           [o]           art     M=S
    único       [único]       adj     M=S
    país        [país]        n       M=S
    que         [que]         pron-indp M=S
    se          [se]          pron-pers M=3S=ACC
    recusa      [recusar]     v-fin   PR=3S=IND
    a           [a]           prp     -
    aceitar     [aceitar]     v-inf   -
    a           [o]           art     F=S
    determinação [determinação, determinar] n       F=S
    européia    []            adj     F=S
    .           [.]           .       -
  Chunks:
          [NP: A França ] [VP: é ] [NP: o único país ] [NP: que ]
          [NP: se ] [VP: recusa ] [PP: a ] [VP: aceitar ] [NP: a determinação européia ]

  Shallow Structure:
                    [SUBJ: A França ] [P: é ] [SC: o único país ] [ACC: que ] [SUBJ: se ]
                    [P: recusa ] [PIV: a ] [P: aceitar ] [ACC: a determinação européia ]
```

Tabela 4.1 – Tags utilizadas pelo CoGrOO

POS	n	Substantivo
	prop	Nome próprio
	art	Artigo
	pron-pers	Pronome pessoal
	pron-det	Pronome determinante
	pron-indp	Pronome independente
	adj	Adjetivo
	n-adj	Substantivo ou adjetivo
	adv	Advérbio
	v-fin	Verbo finito
	v-inf	Verbo no infinitivo
	v-pcp	Verbo no particípio
	v-ger	Verbo no gerúndio
	num	Numeral
	prp	preposição
	conj-s	Conjunção subordinada
	conj-c	Conjunção coordenativa
	intj	Interjeição

Chunks	NP	Sintagma Nominal
	VP	Sintagma Verbal
	ADVP	Sintagma Adverbial
	PP	Sintagma preposicional
Shallow Structure	SUBJ	Sujeito
	ACC	Objeto direto
	DAT	Objeto indireto com preposição
	PIV	Objeto preposicional
	SC	Predicado nominal
	OC	Predicativo do objeto
	P	Predicado
	AS	Objeto Adverbial
	ADVL	Objeto Adverbial
	APP	Identificação de aposto

4.2 Corpus Anotado

4.2.1 Summ-it++

Concebido a partir do corpus Summ-it[12], o Summ-it++[1] consiste em uma nova versão do Summ-it portada para o formato SemEval [58] e enriquecida com duas novas camadas de anotação semântica: Relação entre entidades nomeadas [13]; e Categorias de Entidades Nomeadas [15]. O Summ-it++, assim como o Summ-it, possui 5033 menções, 3022 *links*, 560 cadeias de correferência. Adicionalmente, possui 1086 entidades nomeadas classificadas e 37 descritores de relação entre essas entidades. Na Tabela 4.2, podemos visualizar como são dispostas as informações do corpus.

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

Tabela 4.2 – Esquema de anotação Summ-it++.

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
1	A	o	art	F=S	_	_	_	_
2	opinião	opinião	n	F=S	0	_	_	_
3	é	ser	v-fin	PR=3S=IND	_	_	_	_
4	de	de	prp	_	_	_	_	_
5	o	o	art	M=S	_	_	_	(2
6	agrônomo	agrônomo	n	M=S	0	_	_	_
7	Miguel_Guerra	_	prop	M=S	0	PES	(9:11)	_
8	,		_	_	_	_	_	_
9	de	de	prp	_	_	_	_	_
10	a	o	art	F=S	_	_	_	_
11	UFSC	_	prop	F=S	0	ORG	(9:7)	(3)
12	(	(	(	_	_	_	_	_
13	Universidade_de_Santa_Catarina	_	prop	F=S	0	ORG	_	(3) | 2)
14	)	)	)	_	_	_	_	_
15	.	.	.	_	_	_	_	_
1	Guerra	_	prop	M=S	0	PES	_	(2)
2	participou	participar	v-fin	PS=3S=IND	_	_	_	_
...								

Conforme podemos visualizar na Tabela 4.2, cada coluna de anotação representa:

ID: ID de cada palavra na ordem em que elas aparecem na sentença;

Token: palavra ou multi-palavra;

Lemma: lemma;

POS: *Part-of-speech tagging* de cada palavra;

Feat: *features* (gênero e número) de cada palavra;

Head: denota se a palavra é um núcleo de sintagma nominal (caso sim, o campo recebe o valor ’0’);

NE: representa a categoria semântica das entidades nomeadas;

Rel: representa o descritor que expressa a relação entre um par de entidades nomeadas. Quando essa relação existe, ambas entidades nomeadas envolvidas recebem o ID das palavras que compõem o descritor de relação, seguido do id do token o qual dada entidade se relaciona.

Coref: cada sintagma nominal inicia por “(” seguido do id da cadeia de correferência. Note que “) ” apenas ocorre no último token do sintagma nominal. Basicamente, menções correferentes recebem o mesmo ID.

4.2.2 Corref-PT

É um corpus contendo anotação semi-automática[2] de correferência para o Português. Este é proveniente de um esforço coletivo, durante o IBEREVAL-2017 (Evaluation of Human Language Technologies for Iberian languages)[3]. Nele, propomos a tarefa "*Collective Elaboration of a Coreference Annotated Corpus for Portuguese Texts*" [20], cujo objetivo foi a elaboração coletiva de um corpus anotado com correferência para o Português. Para isso, cada equipe de participantes apresentou um conjunto de textos de seu próprio interesse para realizar a anotação. Sete equipes, com um total de vinte e um anotadores, falantes nativos do Português, variando entre estudantes e professores da área de linguística computacional participaram da tarefa. O corpus resultante é composto por textos jornalísticos [44]; Por textos diversos (livros, revistas, entre outros) [16]; E artigos da Wikipédia[4], selecionados aleatoriamente. Na Tabela 4.3 temos alguns detalhes do corpus, tais como: quantidade de textos, menções, cadeias, entre outras.

Tabela 4.3 – Corref-PT - Estatísticas do corpus

Corpus	**Textos**	**Tokens**	**Menções**	**Menções Correferentes**	**Cadeias de Correferência**	**Maior Cadeia**	**Tam. Médio das Cadeias**
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

O Corref-PT está disponível em quatro formatos distintos: TXT, XML, HTML[5] e SemEval [58]. O recurso é gratuito e pode ser obtido na página do grupo de Processamento da Linguagem Natural da PUCRS[6]

4.3 CorrefVisual

O CorrefVisual[7] [61] é uma ferramenta a qual tem sido aprimorada com o propósito de simplificar a tarefa de anotação manual de correferência. O recurso provê uma interface gráfica que permite ao usuário a visualização de informações que sejam relevantes para a

[2] seus textos foram anotados de forma automática pelo CORP [26] e revisados manualmente utilizando a ferramenta CorrefVisual [61].

[3] http://sepln2017.um.es/ibereval.html

[4] https://pt.wikipedia.org/

[5] Os mesmos formatos disponibilizados pelo CORP, veja na Seção 5.5.

[6] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/corref-pt/

[7] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/correfvisual/

tarefa (como o texto e cadeias de correferência a serem editadas/manipuladas) de forma simplificada.

4.4 Recursos Semânticos

4.4.1 Onto.PT

Construído de forma automática por meio de dicionários e *thesaurus* da língua portuguesa, o Onto.PT [49] é uma ontologia, disponível para o português. Similar ao Wordnet [46], o Onto.PT possui uma estrutura baseada em *synsets*[8] e relações semânticas conectando esses *synsets*, como: hiperonímia, hiponímia, sinonímia, meronímia, entre outras. Na Tabela 4.5, podemos visualizar a quantidade de relações existentes na ontologia. O Onto.PT conta também com uma API[9] que, tendo como entrada um par de palavras, retorna todas as relações entre elas, conforme podemos visualizar na Tabela 4.4.

Tabela 4.4 – Onto.PT: Exemplos de relações semânticas para um dado par de palavras

Par	**Relação**
estudo, pesquisa	sinonimoDe
abelha, inseto	hiponimoDe
animal, cachorro	hiperonimoDe

[8]Grupos de sinônimos que possuem o sentido das palavras que contêm os possíveis significados para uma dada palavra em sua forma lexical ex: carro, automóvel.

[9]http://github.com/rikarudo/OntPORT

Tabela 4.5 – Quantidade de instâncias existentes no Onto.PT

Relação	Argumentos	Quantidade
Sinônimo_De	substantivo, substantivo	84.015
	verbo, verbo	37.068
	adjetivo, adjetivo	45.149
	advérbio, advérbio	2.626
Hiperônimo_De Hipônimo_De	substantivo, substantivo	91.466
Parte_De	substantivo, substantivo	3.809
	substantivo, adjetivo	5.627
Membro_De	substantivo, substantivo	6.369
	substantivo, adjetivo	114
	adjetivo, substantivo	948
Contido_Em	substantivo, substantivo	364
	substantivo, adjetivo	280
Material_De	substantivo, substantivo	873
Causa_De	substantivo, substantivo	1.411
	substantivo, adjetivo	30
	adjetivo, substantivo	706
	substantivo, verbo	78
	verbo, substantivo	10.144
Produtor_De	substantivo,substantivo	1.721
	substantivo, adjetivo	77
	adjetivo, substantivo	505
Propósito_De	substantivo,substantivo	7.100
	substantivo, adjetivo	85
	verbo, substantivo	8.713
	verbo, adjetivo	373
Tem_Qualidade	substantivo, substantivo	998
	substantivo, adjetivo	1.258
Tem_Estado	substantivo, substantivo	345
	substantivo, adjetivo	216
Propriedade_De	adjetivo, substantivo	10.617
	adjetivo, verbo	27.431
Antônimo_De	substantivo, substantivo	17.172
	verbo, verbo	49.422
	adjetivo, adjetivo	25.321
	advérbio, advérbio	683
Lugar_De	substantivo, substantivo	1.393
Maneira_De	advérbio, substantivo	2.166
	advérbio, adjetivo	1.800
Maneira_Sem	advérbio, substantivo	249
	advérbio, verbo	16
Total		448.738

4.4.2 Repentino

O Repentino[60], REPositório para reconhecimento de Entidades Nomeadas, é um recurso público que contém, em média, 490 mil exemplos de entidades nomeadas. Ou seja, trata-se de uma grande lista contendo diversos nomes próprios, como de pessoas, locais, substâncias químicas, organizações, entre outros. Os exemplos de entidades, armazenados no Repentino, encontram-se divididos por várias categorias, cada uma das quais contendo diversas subcategorias, numa estrutura em árvore, garantindo assim uma razoável organização desses exemplos. Na Figura 4.2 podemos visualizar como está disposta a organização das entidades de categoria “EN_SER” (seres vivos) e subcategoria “HUM” (humanos). Outra vantagem e também uma grande motivação de utilizarmos o Repentino em nosso trabalho se dá pelo fato deste recurso possuir anotação de Entidades Mencionadas. Isto é, dentro do Repositório, também encontramos nomes comuns de entidades tais como: “asma, úlcera, homem, vendaval, artrite, entre outros".

Em nosso trabalho reestruturamos as categorias e sub-categorias semânticas do repentino, de forma a considerar estas de forma mais genérica. Essa lista de categorias é descrita detalhadamente na Seção 4.4.4

Figura 4.2 – Exemplo de codificação adotada pelo Repentino

```
<EN_SER subcat="HUM">Abílio Albino As Silva Nunes</EN_SER>
<EN_SER subcat="HUM">Abdul</EN_SER>
<EN_SER subcat="HUM">Abel De Pinho Soares</EN_SER>
<EN_SER subcat="HUM">Abel Feldmann Da Câmara Carreiro</EN_SER>
<EN_SER subcat="HUM">Abel Fernando Queiros Figueiredo</EN_SER>
<EN_SER subcat="HUM">Abraham Lincoln</EN_SER>
<EN_SER subcat="HUM">Achille Talon</EN_SER>
<EN_SER subcat="HUM">Adalberto Alves</EN_SER>
<EN_SER subcat="HUM">Adalberto Nuno da Silva Leite de Freitas</EN_SER>
<EN_SER subcat="HUM">Adalberto Nuno de Silva Leite de Freitas</EN_SER>
<EN_SER subcat="HUM">Adélio Amaro</EN_SER>
<EN_SER subcat="HUM">Adília Lopes</EN_SER>
<EN_SER subcat="HUM">Adelaide Rosa Coelho Teles Madureira</EN_SER>
<EN_SER subcat="HUM">Adelino José Da Silva Oliveira</EN_SER>
<EN_SER subcat="HUM">Adelino Luís Ferreira De Morais E Castro</EN_SER>
<EN_SER subcat="HUM">Adelma Margarida Ferreira de Freitas</EN_SER>
<EN_SER subcat="HUM">Adolf Hitler</EN_SER>
```

4.4.3 Listas Auxiliares

Além do Repentino, de forma a auxiliar na classificação de entidades nomeadas e de substantivos comuns que possam indicar entidades de categorias semânticas específicas, utilizamos três listas, criadas a partir do conteúdo de sites web, como o Wikipedia[10]. Estas possuem nomes próprios e comuns que podem indicar as categorias 'Pessoa', 'Local' e 'Organização'. Essas listas foram utilizadas em trabalhos anteriores [21], de forma melhorar a precisão no processo de REN.

Como exemplo, a lista "Pessoa" possui nomes de profissões e nomes de pessoas comumente utilizados, como: 'agrônomo', 'advogado', 'engenheiro', 'Diego', 'João', 'Aline', 'Tiago', etc. A lista "Local" possui alguns nomes como de cidades, por exemplo, e alguns substantivos comuns, como: 'praça', 'praia', 'cidade', 'morro', 'travessa', 'rua', 'bairro', 'avenida', 'rio', 'lagoa', etc. A lista "Organização" utiliza nomes próprios de empresas mais conhecidas e de substantivos comuns, como: 'instituto', 'agência', 'empresa', 'organização', 'ONG', 'partido', 'comércio', etc.

É importante deixarmos claro que o papel do Repentino e das listas de entidades é auxiliar na anotação semântica das cadeias de correferência e não na resolução de correferência em sí. Basicamente usamos a anotação de correferência para prever possíveis categorias de entidades e atribuir uma única categoria semântica para todos sintagmas de uma cadeia. Como por exemplo:

- o agrônomo Miguel Guerra, Miguel Guerra, M.G., o agrônomo.

Dificilmente recursos para o reconhecimento de entidades nomeadas irão prever que "M.G. é uma 'Pessoa', e não um 'Local'. Ao identificarmos que 'M.G' é uma menção referencial à 'Miguel Guerra' é possível atribuirmos a categoria 'Pessoa' a 'M.G.'.

4.4.4 Categorias de Entidades Consideradas

Em nosso trabalho, por acreditarmos possuir uma lista mais completa, utilizamos rótulos de categorias baseadas no Repentino [60]. Contudo, não fazemos uso de subcategorias, apenas uma categoria principal é atribuída. Essas são descritas abaixo:

1. **Pessoa** – Representa nomes comuns, próprios ou profissões, que remetem à pessoas. Tais como: "Aline, Marcos, Barack Obama, menino, garoto, advogado, juíz, agrônomo, presidente...";

[10] http://pt.wikipedia.org

2. **Organização - Local** – inclui todas as organizações e locais, tais como nomes de empresas, de cidades, estados países. Nomes comuns também são levados em consideração, tais como: “praça, avenida, rua...”. Optamos por unir essas duas categorias pelo fato de ser muito comum nos referirmos uma organização como um local, por exemplo:

   - “João é aluno da PUCRS”
   - “João está indo na PUCRS”

   Note que PUCRS (Pontifícia Universidade Católica do Rio Grande do Sul) é uma entidade do tipo Organização, mas também pode ser do tipo Local. Isso vai depender de seu contexto.

3. **Eventos** – inclui-se todo e qualquer evento, cujo início ou duração estejam claramente definidos. Dentro desse contexto temos: “eventos esportivos, reuniões que envolvam qualquer atividade social ou cultural, como feiras e exposições; acontecimentos históricos, acontecimentos científicos (conferências, Simpósios);

4. **Comunicação** – incluem-se apenas entidades que são produtos relacionados com arte, mídia ou comunicação. Tais como filmes, livros, músicas, jogos digitais, publicações (como jornais e revistas); programas de tv, radio e teatro;

5. **Produtos** – inclui todo o tipo de produtos: comerciais, financeiros, farmacêuticos, industriais; tais como: ferramentas, eletrônicos, eletrodomésticos, produtos identificados por marcas (OMO, Aspirina-C, Clorofina...).

6. **Documentos** – incluem-se documentos em geral, tais como: leis, decretos, tratados, pactos, normas e planos.

7. **Abstração** – inclui entidades abstratas tais como disciplinas, ciências, processos (fotossíntese, pseudomorfose, sulfatação, osmose), teorias, doenças, estados condicionais, símbolos religiosos (crucifixo, pentagrama, Selo de Salomão...), crimes, índices de taxas (PIB, NASDAQ).

8. **Natureza** – incluem-se animais e vegetais assim como fenômenos naturais: ciclones, tufões, micro-organismos, elementos que constituem organismos vivos (músculos, células, ossos...).

9. **Outros Seres** – incluem-se todos os seres reais, ficcionais ou mitológicos, assim como os mitos, exceto pessoas e profissões (os quais são classificados como Pessoa). Exemplos: qualquer ser (real ou ficcional) que não seja humano, vivo ou morto; toda e qualquer entidade mitológica. Ex: Pégaso, Minotauro, Ícaro, Adamastor, Afrodite, Cupido, etc; grupos de pessoas (reais ou ficcionais) que partilhem a mesma identidade geográfica, política, étnica ou ideológica, embora não pertençam a uma organização estruturada, tais como: Incas, Budistas, Dadaístas, Nudistas, Marcianos, Atlantes, Visigodos, etc.

10. **Substâncias** – incluem-se elementos e substâncias como: Paracetamol, H2O, Anelina, penicilina, ácido ascórbico, acetilsalicilato de lisina, boldenona, hematoxilina, lecitina de soja, lidocaína, Mebendazol, nandrolona, Oxibutinina, álcool, glicose, etc.

11. **Outros** – Nesta categoria incluem-se exemplos que não foram encaixados em nenhuma das categorias utilizadas.

4.5 CoNll Scorer

Desenvolvido com o intuito de atender as necessidades da CoNLL shared task [53], [52], o CoNLL Scorer[11] [51] consiste em uma API cujo objetivo é avaliar modelos de resolução de correferência. Seu objetivo principal é prover uma forma automatizada e justa de avaliar tais modelos. Isso porque, como podemos ver em [51], cada métrica favorece uma característica específica entre os *links* de menções. Dados os fatos, o recurso utiliza a média entre as três principais métricas, para determinar uma pontuação única.

Basicamente, tendo como entrada dois arquivos (ambos necessitam estar no formato SemEval [58], um formato muito conhecido e utilizado pela maioria dos corpora): o primeiro, contendo as anotações que são o padrão de referência, e o segundo contendo as anotações, providas automaticamente pelo modelo a ser avaliado, o CoNLL Scorer calcula uma pontuação. Além disso, o recurso fornece também os resultados de todas as métricas conhecidas (MUC, B^3, Ceaf e BLANC) [72, 3, 41, 57], descritas no Capítulo anterior. Nesta tese utilizamos o CoNLL scorer, em conjunto com os corpora Summ-it++ e Corref-PT para avaliação de nossa abordagem.

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

4.6 Considerações do Capítulo

Neste capítulo apresentamos os principais recursos utilizados para avançar nos estudos desta tese. Inicialmente utilizamos o corpus Summ-it, como estudo de caso, de forma a analisar padrões que pudessem indicar uma relação de correferência entre duas ou mais menções. A partir dessa análise, realizamos estudos em trabalhos relacionados, como o modelo de Lee et al. [39](Capítulo 3).

Para anotação de sentenças, *part-of-speech tagging*, *lemma*, *chunking*, gênero e número, utilizamos o parser Cogroo, um recurso muito utilizado por ferramentas de escritório de código aberto, como Open-Office[12] e LibreOffice[13]. De forma a introduzir conhecimento semântico em nossa proposta, utilizamos três principais recursos: Repentino e listas de nomes comuns e próprios, para o reconhecimento de entidades nomeadas e o Onto.PT, uma ontologia concebida para o Português, construída automaticamente, a partir de uma série de outros recursos, como *thesaurus*, dicionários, enciclopédias e corpora.

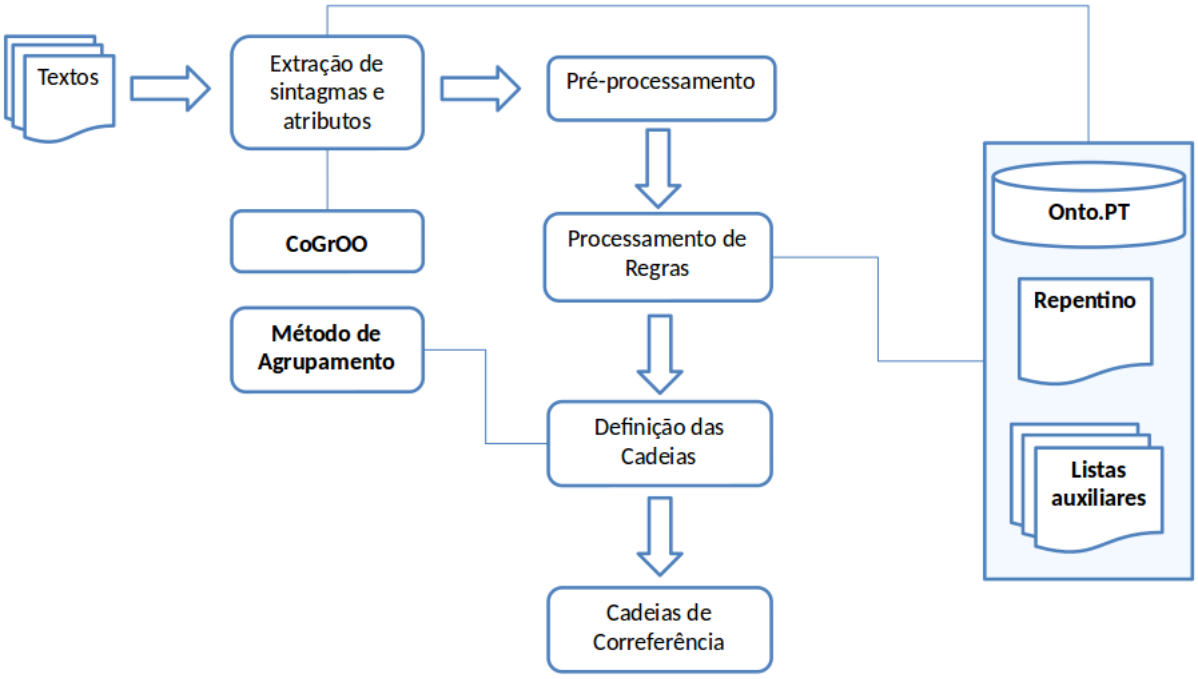
De forma a avaliar nosso modelo proposto, assim como a conferência CoNLL[52], propomos o uso do CoNLL Scorer, um recurso que, por meio de dois arquivos no formato SemEval, permite avaliar o desempenho de modelos de resolução de Correferência. Além disso, mostramos outros recursos, concebidos no decorrer desta tese de doutoramento, como Summ-it++ [2], Corref-PT [20] e CorrefVisual [61]. No capítulo a seguir descrevemos em detalhes nosso processo para resolução automática de correferência em textos da língua Portuguesa.

[12]https://www.openoffice.org/pt-br/
[13]https://pt-br.libreoffice.org/

5. MODELO PROPOSTO

Neste capítulo, descrevemos o processo para resolução automática de correferência na língua Portuguesa, proposto no âmbito desta tese de doutorado. Para isso, consideramos o termo “menção” como uma expressão textual associada a entidade(s) ou evento(s) do mundo real. Em um discurso, menções que referem a uma mesma entidade são chamadas menções correferentes e formam um conjunto de menções, definido como cadeia de correferência [50]. Uma visão geral do processo proposto é ilustrada na Figura 5.1

Figura 5.1 – Modelo proposto



Inicialmente, realizamos a extração de sintagmas nominais e seus respectivos atributos, por meio do parser CoGrOO [63]; seguido de um pré-processamento, o qual removemos sintagmas que: iniciem com entidades numéricas como percentual, dinheiro, cardinais e quantificadores (9%, $10, 000, Dez, Mil, 100 metros). Apesar de existir correferência numérica, essa possui baixa ocorrência e requer um tratamento distinto dos demais sintagmas. Portanto, neste primeiro momento, optamos por não tratá-las. Após as etapas de extração de sintagmas e pré-processamento aplicamos nosso conjunto de regras (Seções 5.4.1 e 5.4.2). Como recursos semânticos, utilizamos a ontologia Onto.PT[33] (descrita na Seção 4.4.1), para relações de hiponímia e sinonímia; e, Repentino, em conjunto de listas auxiliares para prover as categorias de entidades nomeadas (Seções 4.4.2 e 4.4.3). Conforme dito na Seção 4.4.3, nossa abordagem proposta não faz uso das categorias de entidades nomeadas para melhorar o processo de resolução de correferência. Ao contrário disso, utilizamos a informação de cadeias de correferência para inferir categorias a sintagmas não classificados semanticamente no processo de REN. Em trabalhos anteriores [21]

realizamos estudos envolvendo diferentes combinações de uso da *feature* Categoria Semântica. Basicamente, limitamos os pares candidatos a correferência a apenas pares com categorias de entidades idênticas. Como resultado disso, houve uma perda significativa em abrangência do modelo. Isso porque nem sempre os recursos de REN irão atribuir corretamente uma categoria de entidade nomeada, fazendo com que o modelo perca pares importantes. Além disso, quando lidamos com substantivos comuns, a probabilidade de uma categoria não ser aferida corretamente é ainda maior, dado que os modelos de REN são voltados a nomes próprios.

Em arquiteturas baseadas em regras, cada etapa consiste em aplicar determinado filtro/regra, objetivando agrupar duas menções m_x e m_y, caso suas restrições sejam satisfeitas. Em nossa abordagem não realizamos esses agrupamentos de imediato. Utilizamos uma estrutura semelhante a grafos para armazenar o resultado de cada regra processada, assim como as possíveis ligações de correferência. Tendo essas informações processadas, utilizamos um novo método de agrupamento de menções(Seção 5.3.2) que objetiva identificar quando uma menção é anafórica ou nova no discurso. Por meio de nosso algoritmo proposto foi possível melhorar a precisão de nosso modelo em 10% sem abrir mão de sua abrangência, conforme veremos em experimentos realizados.

5.1 O Problema no Processo de Agrupamento de Menções

Como bem sabemos a resolução de correferência não é uma tarefa trivial e envolve dois principais níveis de processamento: i) Detecção de menções, o qual consiste em determinar quais *tokens* pertencem a quais sintagmas nominais e ii) Classificação, que consiste em classificar um dado par de menções como correferentes ou não. Contudo, gerar cadeias de correferência envolve um processo mais elaborado, o qual devemos nos basear em pares de menções referenciais e menções previamente agrupadas. Este processo é chamado de processo de agrupamento (*Clustering*), o qual consiste em criar partições, que contenham menções, que referem-se às mesmas entidades.

Atualmente, o método mais utilizado (por modelos baseados em regras) para realizar tal agrupamento consiste em aplicar técnicas de pré e pós modificadores em conjunto de regras que podem indicar uma relação referencial, como nos trabalhos de Lee et al., Garcia et al. e Fonseca et al. [39], [31], [25]. Suas abordagens agrupam uma menção m_x a seu antecedente se pelo menos uma regra/*sieve*[1] for satisfeita. Contudo, em alguns casos precisamos decidir se uma menção pertence a uma determinada cadeia ou a outra. Considere o exemplo:

[1] Lee et al. utilizam uma estrutura denominada *Sieve*. Um *Sieve* consiste em uma ou em um conjunto de regras.

a) "*. . . informado por [o governador de São Paulo$_1$], [Geraldo Alckmin$_2$], em seu último discurso. [O governador$_3$] disse que . . . Contudo, [José Ivo Sartóri$_4$] disse que isso não irá mudar . . . [O governador$_5$]. . .*"

Note que temos duas EN (Entidades Nomeadas): "Geraldo Alckmin" e "José Ivo Sartori" (ambos governadores). Contudo existe uma combinação exata entre os sintagmas [O governador$_3$] e [O governador$_5$]. Assim, os modelos de correferência provavelmente irão realizar o agrupamento entre esses dois sintagmas. Isso porque os pré e pós modificadores utilizados pela maioria dos modelos de regras não são suficientes para evitar um agrupamento incorreto. Em outras palavras, o desafio no agrupamento de menções é um problema bem conhecido e consiste em determinar se uma menção m$_x$ pertence a uma cadeia C$_x$ ou C$_y$. Nesse caso, a menção [O governador$_5$] refere-se a [José Ivo Sartóri$_4$], não a [O governador$_3$], o qual é uma correferência de [o governador de São Paulo$_1$] e [Geraldo Alckmin$_2$] . Quando consideramos as relações semânticas esse problema torna-se mais desafiador. Isso ocorre porque o conhecimento semântico adquirido por meio da relações de sinonímia e hiponímia são mais suscetíveis à ambiguidade, como em:

b) "*[a Terra$_1$] é [um astro$_2$] e está a uma distância de 149.600.000 quilômetros do [Sol$_3$] . . . o universo é amplo e, quando consideramos o tamanho do [Sol$_4$], isto é, [a estrela$_5$]. . .*"

No exemplo (b), existe uma relação de hiponímia entre [a Terra$_1$] e [o astro $_2$] e; uma relação de sinonímia entre [o astro$_2$] e [a estrela$_4$]; contudo, é possível notarmos que no trecho de texto existem duas cadeias de correferência: C$_1$={ [a Terra$_1$], [o astro$_2$]} e C$_2$={[Sol$_3$], [Sol$_4$], [a estrela$_5$]}. Considerando os métodos de agrupamento atuais (como o proposto em [39]), a saída será uma única cadeia de correferência, contendo todas as menções. Em abordagens semânticas, casos como este são frequentes e reduzem consideravelmente a precisão dos modelos. Isso acontece pela probabilidade de obtermos ligações incorretas ser muito maior quando consideramos as relações semânticas. Nesta tese realizamos experimentos envolvendo diferentes abordagens para lidar com o problema no agrupamento de menções e, como resultado, mostramos que métodos mais elaborados podem prover melhorias significativas a tal tarefa.

5.2 Arquitetura

Em nossa abordagem propomos uma estrutura semelhante à teoria dos grafos, assim como em [43]. Para isso definimos três objetos: 'Nodo', '*Feature*' e 'Aresta'. Podemos considerar que cada menção representa um Nodo (nó em um grafo); o objeto *Feature*, carrega suas informações morfo-sintáticas; e, o objeto Aresta consiste em representar suas

possíveis ligações entre outros objetos Nodo. Nas tabelas 5.1, 5.2 e 5.3 são exibidas as características de cada objeto.

Na Tabela 5.1, representando “Arestas” ou possíveis ligações que podem existir entre duas ou mais menções (Nodo), temos as regras e seus valores verdade. Em modelos de correferência, que fazem uso de regras linguísticas, uma relação de correferência existe quando pelo menos uma regra linguística apresenta-se verdadeira. Contudo, introduzindo relações semânticas como sinonímia e hiponímia a chance de obtermos falsas ligações é maior. Dessa forma, em nossa abordagem, não basta uma regra retornar um valor verdade, para que essa ligação ocorra, é necessário explorar a representação do discurso, de forma a decidir se dada menção é um referente ou se é nova no discurso. Retomando o exemplo da cadeia “a França”(Figura 5.2), podemos ver como dada cadeia pode ser representada (Figura 5.3).

Figura 5.2 – Cadeia de correferência “a França”

Após o anúncio de o sequenciamento de o genoma , em a semana passada , **[a França [5]]** resiste como **[único país de a União Européia a não permitir o patenteamento de genes [5]]**. A UE adota , desde junho de 1998 , diretiva favorável a o patenteamento de genes . O texto , redigido por o Parlamento_Europeu , Comissão_Européia_e_Conselho_de_Ministros , utiliza o princípio de que o genoma não é patenteável , mas a sequência de um gene pode ser . em o entanto , há restrições . O patenteamento só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento de o gene é detalhado . **[A França [5]]** é **[o único país que se recusa a aceitar a determinação européia [5]]**. A ministra de a Justiça de **[o país [5]]** , Elisabeth_Guigou , disse que a norma é incompatível com as leis francesas de bioética . em o início de o mês , o CCNE (Comitê_Consultivo_Nacional_de_Ética) , órgão que orienta o governo francês sobre aspectos éticos de a biotecnologia , reforçou a posição de a ministra , alegando que o conhecimento de a sequência de um gene não pode ser assimilado como produto patenteado e , portanto , não é patenteável . Bem comum de a humanidade , (o sequenciamento de genes) não pode ser limitado por patentes que pretendem , em nome de o direito de propriedade industrial , proteger a exclusividade de esse conhecimento , diz parecer de o CCNE . O assunto deve ser debatido durante a presidência francesa de a UE , em o segundo semestre .

Figura 5.3 – Representação da Cadeia “a França”

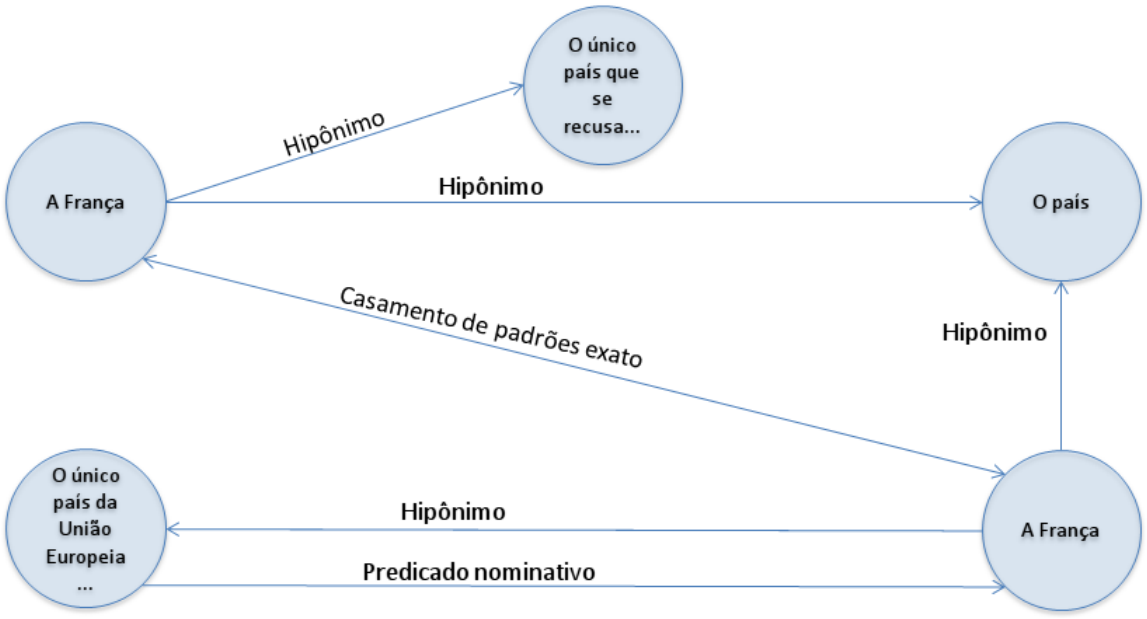


Tabela 5.1 – Representação do objeto Aresta

Variável	Tipo de Dado
CasamentoDePadroesExato (R1)	Booleano
CasamentoParcialPeloNucleo (R2)	
ApostoExplicativo (R3)	
ApostoEspecificativo (R4)	
Acronimo (R5)	
PredicadoNominativo (R6)	
PronomeRelativo (R7)	
CasamentoRestritoPeloNucleo (R8)	
CasamentoRestritoPeloNucleo (R9)	
CasamentoEntreNomesProprios (R10)	
CasamentoParcialEntreNomesProprios (R11)	
Hiponimia (R12)	
Sinonimia (R13)	
QuantidadeRegrasVerdade	Inteiro
SnID	

Ainda na Tabela 5.1 podemos notar que, além de regras, temos dois atributos do tipo inteiro. O primeiro (QuantidadeRegrasVerdade), apesar de não ser tão intuitivo, é utilizado por nosso algoritmo de agrupamento, objetivando auxiliar no processo de agrupamento de menções; o segundo, remete ao "ID" da menção a qual dada aresta faz ligação. Na tabela 5.2 são exibidos atributos referentes às palavras que cada menção pode possuir. E, no objeto nodo, composto por, além de seus atributos-base, um vetor de "*Features*" e de "Arestas", representando todas possíveis ligações entre outros nós.

Tabela 5.2 – Representação do objeto *Feature*

Variável	**Tipo de Dado**	**Descrição**
Token	String	token / palavra
POS		Informações referentes à classe gramatical do token.
Morfo		Informações morfológicas do token, como gênero e número.
Lema		Forma canônica do token, token lematizado
TokenId	Inteiro	Id único para o token.
Sentença		Sentença em que o token ocorre.

Tabela 5.3 – Representação do objeto Nodo

Variável	Tipo	Descrição
Sintagma	String	Sintagma nominal.
Nucleo		Núcleo do Sintagma.
Lema		Lema do Núcleo.
CategoriaSemântica		Categoria semântica do sintagma nominal.
Words	Vetor de *Features*	Vetor de objetos *Feature*, contendo informações de POS, análise morfológica, tokens, lema, entre outras. (Tabela 5.2)
PaiDe	Vetor de Inteiros	Possui o SnID de possíveis sintagmas-filho. (adjuntos adnominais)
Ligações	Vetor de Arestas	Vetor de objetos Aresta, contendo todas as possíveis ligações deste com outros sintagmas e o valor de cada regra (verdadeiro ou falso, Tabela 5.1).
FilhoDe	Inteiro	Caso o sintagma seja um adjunto, este campo carrega informação de seu sintagma pai.
TokenInicial		Id do primeiro token do sintagma.
TokenFinal		Id do último token do sintagma.
SnID		Id único do sintagma.
Sentença		Sentença em que o sintagma ocorre.

5.3 Métodos de Agrupamento

Nesta seção apresentamos dois métodos de agrupamento de menções considerados em nossos estudos. Inicialmente apresentamos o Baseline, método de agrupamento de menções utilizado pelos modelos de regras atuais e, posteriormente, apresentamos nosso método de agrupamento proposto. Veremos que este pode ter como base, diferentes critérios de agrupamento.

5.3.1 Baseline:

Comumente utilizado por abordagens baseadas em regras, consiste em agrupar uma menção a seus antecedentes quando pelo menos uma regra/*sieve* apresenta valor verdade.

5.3.2 Método Proposto

Nosso método proposto (Algoritmo 5.1) basicamente recebe como entrada uma lista ordenada de menções "M" e devolve uma lista de Cadeias contendo essas menções devidamente agrupadas, de acordo com o critério selecionado. Nosso método é baseado no trabalho de Heim [34] e consiste em explorar a representação do discurso[2] (Pragmática[3]). Para isso, assumimos que qualquer menção é nova no discurso se não possuir ligação de correferência com uma ou mais menções antecedentes. Essas ligações são consideradas utilizando nosso conjunto de regras proposto. Assim, sempre que uma menção não possui uma relação referencial (nenhuma regra é satisfeita), uma nova cadeia é gerada. Basicamente utilizamos uma lista de menções M (esta lista está ordenada na ordem em que as menções ocorrem no texto), contendo todas as menções de um documento de entrada. Note que cada menção pode ter uma ligação de correferência entre uma ou mais cadeias 'C'. Dessa forma, armazenamos o Id dessas cadeias em um vetor 'S' (apenas se M_0 possui alguma relação de correferência com C_i (se alguma regra retorna o valor verdade)). O próximo passo é responsável por agrupar uma menção atual M_0 a uma cadeia existente C_k ou criar uma nova cadeia de correferência, usando M_0[4] Isso depende do critério de agrupamento utilizado (*CritérioDeAgrupamento* – ver critérios de agrupamento na Seção 5.3.3). Realizamos experimentos considerando cinco critérios. Esses experimentos são descritos no Capítulo 6.

5.3.3 Critérios de Agrupamento

Nesta seção descrevemos os cinco critérios de agrupamento propostos, considerados em nossos experimentos. Veremos que cada critério explora características distintas,

[2]Consideramos como representação do discurso a forma como as ideias são construídas em textos de linguagem natural, considerando sua construção linguística e seu contexto de uso, bem como suas formas de expressões comuns.

[3]Ramo da linguística que analisa o uso concreto da linguagem pelos falantes da língua em seus variados contextos.

[4]Note que, para cada iteração, M_0 muda.

Algorithm 5.1 Algoritmo de Agrupamento

```
1: enquanto (tamanho de M > 0) faça
2:     int j ← 0;
3:     int[ ] S;
4:     para cada i ∈ C faça
5:         se M_0 tem relação com C_i então
6:             S_j ← C_i
7:             j ← j + +
8:         fim se
9:     fim para
10:    se j > 0 então
11:        int k ← CritérioDeAgrupamento(M_0, S, C)
12:        C_k ← M_0
13:    senão
14:        C ← criaNovaCadeia(M_0)
15:    fim se
16:    M ← Remove(M, 0)
17: fim enquanto
```

como peso de regras, quantidade de regras com valor verdade, quantidade de menções referenciais, entre outras.

Cadeia mais Próxima:

Consiste em, para cada menção M_0, explorar o conjunto C (conjunto de cadeias), objetivando buscar a menor distância de M_0 e a cadeia em questão. Essa distância é baseada na posição dos sintagmas já pertencentes à cadeia em questão.

Peso por Regra:

Para cada menção, explora o conjunto C, objetivando encontrar o maior peso (em nível de cadeias). Nosso modelo proposto possui treze regras. Para cara regra satisfeita, soma-se 1 à pontuação. Assim, se, por exemplo, em C_x, existem duas menções correferentes com M_0 (M_a e M_b). M_a e M_b possuem respectivamente três e duas regras com valor verdade, o peso da cadeia será cinco.

Peso por Menção:

Seleciona C_x, que possuir a maior quantidade de menções correferentes, relacionadas à M_0. Cada menção correferente acrescenta 1 ao peso de C_x.

Peso por Regras e Menções:

Este critério considera a junção de dois pesos. Este peso é obtido por meio da soma de dois métodos previamente mencionados (Peso por Regra e Peso por Menção).

Peso por F-Score:

Este peso é calculado usando a soma dos pesos de todas as regras com valores verdade entre M_0 e as menções de C_x. Para determinar os pesos de cada regra, utilizamos os valores obtidos em experimentos individuais com cada regra [23]. Cada peso é obtido por meio da métrica CoNLL/100. Assim, supondo que exista um valor verdade para a regra Sinonímia entre M_0 e M_a, e um valor verdade para a regra "Casamento de Padrões Exato" entre M_0 e M_b; seu peso será Peso(C_x)=(0,177+0,333). Esse critério seleciona a cadeia C_x em que Peso(C_x) obtiver o maior peso. Cada regra contabiliza uma "probabilidade" de M_0 pertencer à C_x. Os pesos de cada regra são exibidos na Tabela 5.4.

Tabela 5.4 – F-Score, peso das regras

Regra	Peso
Casamento de Padrões Exato	0,333
Casamento Parcial pelo Núcleo	0,398
Aposto Explicativo	0,146
Aposto Especificativo	0,019
Acrônimo	0,018
Predicado Nominativo	0,005
Pronome Relativo	0,004
Casamento Restrito pelo Núcleo1	0,466
Casamento Restrito pelo Núcleo2	0,469
Casamento entre Nomes Próprios	0,147
Casamento Parcial entre Nomes Próprios	0,155
Hiponímia	0,056
Sinonímia	0,177

5.4 Regras Linguísticas

5.4.1 Regras Básicas

Nossas regras formam um conjunto facilmente encontrado em trabalhos realizados para o Inglês [39], [54], [64]. Contudo, nosso trabalho possui características que o difere dos demais trabalhos existentes, tendo como diferencial o idioma para o qual é voltado e

sua combinação específica de regras. Além disso, poucos trabalhos, mesmo para o Inglês, abordam o uso de regras semânticas, como Hiponímia e Sinonímia, para a resolução de correferência. Outro diferencial de nossa abordagem reside na metodologia de agrupamento de menções. Muitas de nossas regras foram adaptadas da literatura, considerando o padrão linguístico do Português e as limitações dos recursos disponíveis para o nosso idioma.

Casamento de Padrões Exato: **(Regra 1)**

Considera como correferentes duas menções, cujos sintagmas nominais sejam exatamente iguais, incluindo seus modificadores e determinantes.

c)[o Brasil], [o Brasil]

d)[a Amazônia], [a Amazônia]

Esta regra não considera pronomes e, para considerar duas menções como referenciais, essas não podem pertencer a uma construção de aposto especificativo (regra 4); caso elas pertençam, seus sintagmas ligeiramente anteriores devem ser iguais. Com essa restrição evitamos um possível[5] agrupamento de menções como:

e)[[o telescópio] [**Gemini**]], [[o projeto] [**Gemini**]]

Note que os sintagmas “Gemini” são exatamente iguais, no entanto são sub-sintagmas (adjuntos) de “o telescópio” e “o projeto”. Em poucas palavras, após o processo de chunking[6], temos os seguintes sintagmas nominais: [o telescópio], [Gemini],[o projeto] e [Gemini]. Logo, mesmo esses sintagmas nominais possuindo um casamento exato não necessariamente significa que existe uma uma relação de correferência, dado que estes são adjuntos adnominais.

Casamento Parcial pelo Núcleo: **(Regra 2)**

Considera como correferentes duas menções, cujo casamento obtido por meio do truncamento de seus sintagmas seja igual num mesmo contexto. O truncamento das menções é realizado levando em consideração seus núcleos, como nos exemplos abaixo:

f)[o piloto americano], [o piloto]

g)[o ministro da justiça], [o ministro]

[5]Tratamos como “possível agrupamento”, pois não necessariamente, menções cujas regras retornem um valor verdade serão agrupadas, dado que este agrupamento é realizado por nosso algoritmo proposto.

[6]Nem sempre o CoGrOO efetua a separação dos adjuntos adnominais. No entanto, para ambos os casos esta restrição é válida e previne ligações incorretas, aumentando a precisão do modelo

Assim como na regra Casamento de Padrões Exatos, pronomes e menções que estejam em uma construção de Aposto Especificativo não são consideradas por esta regra.

Aposto Explicativo: (Regra 3)

Considera duas menções como referenciais, caso essas estejam em uma construção de aposto [9], [6]. Essa regra consiste em buscar por marcações padrões que ajudam a identificar o aposto, como parênteses e menções entre vírgulas.

h) [A Embrapa] ([Empresa Brasileira de Pesquisa Agropecuária])

i) [A ministra da justiça do país], [Elisabete Guigou], ...

Aposto Especificativo: (Regra 4)

Consiste em verificar se duas menções vizinhas, m_i e m_{i+1}, estão em uma construção de aposto especificativo[7] [9], [6]. Basicamente, se satisfazem as seguintes restrições:

- A menção m_{i+1} é um nome próprio;
- A menção m_i é um substantivo comum;
- A menção m_i deve possuir um artigo definido;
- A menção m_{i+1} não pode possuir um determinante;
- As menções m_i e m_{i+1} devem estar na mesma sentença e serem adjacentes no texto (não pode haver outras palavras entre elas).
- Caso o determinante de m_i esteja no plural, considera como correferentes todas as menções subsequentes que:
  - sejam nomes próprios;
  - estejam na mesma sentença;
  - estejam separadas por vírgula (ou “e” após as vírgulas).

j) [o arqueólogo português], [Francisco Alves]

k) [o galeão], [Nossa Senhora dos Mártires]

l) [os brasileiros], [Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi].

Acrônimo: (Regra 5)

[7] Diferente de [39], aplicamos esta regra a todos os sintagmas nominais, não apenas à categoria pessoa.

Considera duas menções como correferentes, quando m_i é sigla de m_j.

m) [Organização das Nações Unidas], [a ONU]

Predicado Nominativo: (Regra 6)

Tem como objetivo identificar predicados nominativos e retornar um valor verdade para suas menções referenciais. Para isso, buscamos por uma sequência que possua um verbo de ligação seguido de um determinante/artigo, como, por exemplo, (é um, é uma, foi o, foram os...); encontrada a sequência (verbo de ligação + determinante), essas menções adjacentes podem ser consideradas como correferentes, como em:

n) [A França] **é** [**o** único país que se recusa a aceitar a determinação europeia]

Nessa regra, consideramos apenas o verbo "ser", conjugado no passado, presente e futuro do singular e do plural. Outros verbos de ligação não foram considerados, pois geralmente confundem-se por adjetivos, como por exemplo:

– Cláudia **anda** nervosa.
– Diana **continua** feliz.
– Nicole **ficou** triste.
– João **está** feliz.

Pronome Relativo: (Regra 7)

Busca por menções que possuam/sejam pronomes relativos. Identificado um pronome relativo m_{i+1}, este é considerado um referente da menção anterior adjacente m_i:

o) [Wilkinson Microwave Anisotropy Probe], [cujos] primeiros dados.

Casamento Restrito pelo Núcleo: (Regras 8 e 9)

Consiste em considerar como correferentes, duas menções, caso seus núcleos sejam iguais. Esse casamento, ao considerar apenas o núcleo dos sintagmas, muitas vezes pode causar um agrupamento incorreto, já que não considera que possam existir modificadores incompatíveis, como, por exemplo: Universidade de São Paulo e Universidade de Brasília. Note que os núcleos desses sintagmas são iguais, no entanto referem-se a entidades distintas. Para evitar esse tipo de agrupamento incorreto, esta regra implementa algumas cláusulas restritivas, que devem ser combinadas de modo a produzirem uma possível ligação de correferência mais precisa.

- **Casamento entre Núcleos:** O núcleo da menção atual m_j precisa ser o mesmo do antecedente m_i.

p) [Universidade Federal de São Paulo] ... [a Universidade] ...

- **Palavra Modificadora**

Todas as palavras (substantivos comuns, próprios, verbos, adjetivos e advérbios) de dada menção m_j, não consideradas como *stopwords*, são incluídas em uma lista e comparadas com a menção antecedente m_i. Dessa forma, é possível verificar se existe alguma palavra que modifica o núcleo do antecedente. Essa cláusula explora a propriedade de discurso que nos diz que é incomum introduzirmos novas informações em novas menções a uma mesma entidade. Basicamente, menções subsequentes a uma mesma entidade possuem a tendência de serem menos explicativas.

q) [A menina que caiu e se machucou], [A menina que está feliz]

Note que as palavras “está” e “feliz”, existentes na menção atual, não são *stopwords*, então verificamos se essas duas palavras modificam o antecedente. Como o antecedente não possui as palavras “está e feliz”, elas naturalmente o modificarão. Portanto, as menções são consideradas não-referenciais.

r) [A estrada de Minas Gerais que ficará pronta], [A estrada que talvez esteja pronta]

As menções contidas no exemplo acima também seriam consideradas não-referenciais, dado que o advérbio “talvez” e o verbo “esteja” (contidos em “A estrada que talvez esteja pronta”) modificam o antecedente.

- **Modificadores Compatíveis**

Os modificadores de uma menção m_j atual são todos incluídos na lista de modificadores do candidato antecedente m_i. Essa cláusula é semelhante à “Palavra Modificadora”, com o diferencial de que considera apenas modificadores que são substantivos e adjetivos. Em outras palavras, essa regra verifica se os modificadores do tipo adjetivos e substantivos, quando existem na menção, são iguais aos da menção anterior. Note que essa heurística retornaria o mesmo resultado que a regra “Palavra Modificadora” para o exemplo “q”, porém teria um resultado diferente para o exemplo “r”. Ou seja, o fato de haver um modificador – advérbio (talvez) e um verbo (esteja), por exemplo – não afeta o fato de serem correferentes, altera apenas o sentido do enunciado. Logo, a clásula “Modificadores Compatíveis” retornaria um valor verdade para as menções do exemplo “r”, pois as palavras da menção atual, m_j, (A estrada que talvez esteja pronta), consideradas não stopwords são: “Estrada” e “pronta”, palavras que não modificariam o antecedente.

- **Encapsulamento de Menções**

Nesta cláusula, para duas menções serem correferentes, uma menção não pode ser parte constituinte da outra. De forma a reconhecer este tipo de dependência, utilizamos o reconhecimento de preposições, como: 'de' (e suas variações 'do', 'da', 'dos', 'das') e 'em' (e suas variações 'no', 'na', 'nos' e 'nas'). No exemplo "s", [o menino] não pode fazer referência a [o pijama listrado] justamente porque a regra faz com que a preposição torne-se parte indispensável para haver correferência. Desse modo, a preposição 'de' torna o sintagma [o pijama listrado] expressão adjunta de [o menino].

s) [O menino de pijama listrado], [o pijama listrado].

É importante mencionar que a Regra "Casamento Restrito pelo Núcleo" consiste de duas etapas. A primeira (Regra 8) considera as cláusulas (Casamento entre Núcleos $\wedge$ Palavra Modificadora $\wedge$ Encapsulamento de Menções). A segunda (Regra 9) busca menções em que (Casamento entre Núcleos $\wedge$ Modificadores Compatíveis $\wedge$ Encapsulamento de Menções) sejam satisfeitas. Essas duas variações foram propostas por [39] e mostraram uma melhoria de 0,9% na medida-f, quando utilizadas linearmente.

Casamento entre Nomes Próprios: (Regra 10)

Considera como correferentes duas menções caso as seguintes condições sejam satisfeitas:

- ambas as menções devem conter nomes próprios;
- os nomes próprios precisam ser iguais lexicalmente;
- as duas menções não devem estar encapsuladas, ou seja, devem respeitar a cláusula "Encapsulamento de Menções".

t) [Califórnia],[a região sul da Califórnia].

No exemplo acima, temos a violação da terceira condição. Note que ambos os sintagmas nominais possuem o mesmo nome próprio, mas violam a cláusula "Encapsulamento de Menções", de modo semelhante ao exemplo "s". Neste caso, [Califórnia] e [da Califórnia] não podem ser correferentes pelo fato de a segunda menção estar ligada a uma preposição, tornando-a adjunto adverbial de lugar. Portanto, há uma especificação, em que não se está referindo a toda a Califórnia, mas somente à região sul desse estado.

Casamento Parcial entre Nomes Próprios: (Regra 11)

Semelhante à regra “Casamento entre Nomes Próprios”, mas permite que o núcleo da menção atual m_j combine com qualquer palavra existente na menção anterior m_i. Como em: [o agrônomo da UFSC, Miguel Guerra] e [Guerra]. Contudo, para duas menções serem consideradas referenciais, algumas cláusulas devem ser respeitadas:

- ambas as menções devem conter nomes próprios;
- pelo menos uma palavra de m_j deve ser igual à m_i;
- o agrupamento deve respeitar a cláusula “Palavra Modificadora”.

5.4.2 Regras Sintático-Semânticas

Hiponímia: **(Regra 12)**

Considera como correferentes duas menções (m_i e m_j) se os lemas, provenientes dos núcleos de m_i e m_j são hipônimos. Para encontrar tais relações, utilizamos o Onto.PT. Esta regra ajuda no reconhecimento de menções como as do exemplo abaixo:

u) Já se perguntou como as abelhas fabricam mel? Os insetos saem em busca de...

Para evitar o agrupamento incorreto de menções (exemplo “v”), foram combinadas técnicas de pré, pós modificadores e a verificação de determinantes. Em “v”, se extrairmos o lema do núcleo das menções e efetuarmos uma busca pela existência de relações semânticas entre “quebra-cabeça” e “problema”, veremos que “quebra-cabeça” possui uma relação de hiponímia com “problema”, mas note que as menções “o quebra-cabeça genético” e ”um problema ambiental” não são correferentes. Para evitar tais casos, adicionamos a cláusula “Palavra Modificadora” e restrição de ordem entre os pronomes. Isto é, assumimos que, em um texto bem estruturado, é incomum partirmos de algo definido e, na sequência, nos referirmos a essa mesma menção de forma indefinida (ex: [a aeronave], [um avião]). Dessa forma, além de o termo “ambiental” tornar-se um modificador, é possível também verificarmos que seus determinantes não são compatíveis. Assim, o agrupamento das menções não é realizado. É importante mencionar que nas regras de Hiponímia e Sinonímia os núcleos não são considerados palavras modificadoras.

v) Foi o tempo em que decifrar o genoma ... o **quebra-cabeça** genético... Isso é um **problema** ambiental...

Nesse sentido, para duas menções serem consideradas referenciais, quatro condições precisam ser satisfeitas:

- o lema do núcleo das menções m_i e m_j necessita possuir uma relação de hiponímia;
- se m_i carregar um pronome definido, m_j não pode carregar um pronome indefinido;
- m_i e m_j precisam concordar em número[8] (singular/plural);
- não podem haver palavras que modifiquem as menções (cláusula Palavra Modificadora).

Nós consideramos apenas a relação de hiponímia entre um referente e seu antecedente (não utilizamos hiperonímia), dado que no Português é mais comum introduzirmos uma entidade de forma mais específica e, em suas próximas menções, utilizarmos termos mais gerais para referir à mesma entidade, conforme o exemplo "u". Além disso, testes realizados com a regra Hiperonímia foram realizados, no entanto, a regra gerou muitas ligações incorretas[9] entre as menções.

Sinonímia: **(Regra 13)**

Semelhante à regra Hiponímia, a regra Sinonímia considera duas menções como referenciais, quando há uma relação de sinonímia entre elas, respeitando as seguintes restrições:

- o lema do núcleo das menções m_i e m_j necessitam possuir uma relação de sinonímia;
- se m_i carregar um pronome definido, m_j não pode carregar um pronome indefinido;
- m_i e m_j precisam concordar em número;
- não podem haver palavras que modifiquem as menções;
- cada nova menção a ser agrupada a dada cadeia de correferência, por esta regra, necessita possuir uma relação de sinonímia com todas as menções da cadeia em que pretende ser agrupada (esta verificação é realizada durante o processo de agrupamento, realizado por nosso método de agrupamento proposto). Respeitando esta restrição, evitamos agrupar menções como em:

w) A Terra é um astro do sistema solar. Esse planeta orbita a uma distância de 149.600.000 km do Sol.

Note também, que com as restrições utilizadas, a regra Sinonímia não considera referenciais os sintagmas [a Terra] e [um astro do sistema solar], apenas os sintagmas [A Terra] e [planeta]. Contudo, o sintagma [um astro do sistema solar] é reconhecido pela regra 6 (Predicado Nominativo).

[8] mesmo existindo casos em que esse tipo de relação possa existir, como em : [os ossos] e [o fóssil], por meio dessa restrição foi possível aumentar a precisão de nosso modelo em 2,93% (métrica MUC).

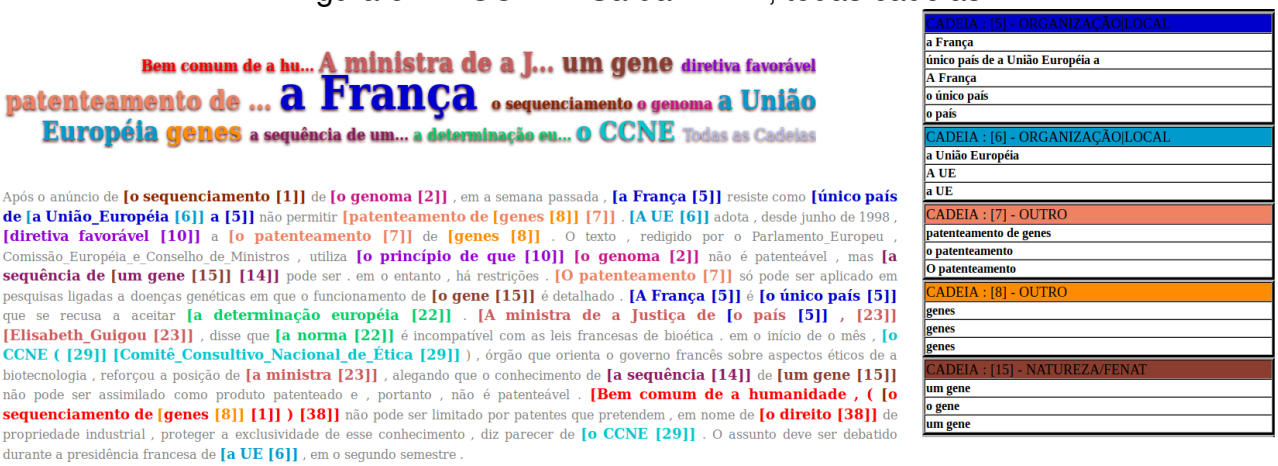
[9] ex: (o espaço – hora), (o espaço – uma bola), (a Terra – uma bola)

5.5 CORP

Proveniente do estudo de caso de nosso modelo proposto, o CORP é um recurso para a resolução de correferência em Português, o qual consideramos como parte das contribuições desta tese. Inicialmente, o CORP foi concebido com o intuito de validar nossos estudos e experimentos. Contudo, tornou-se um recurso funcional e tem auxiliado em diversas tarefas de PLN, inclusive na concepção de outros recursos anotados, como o Corref-PT (Seção 4.2.2).

Referente aos formatos de entrada e saída, o CORP recebe como entrada arquivos “txt”, livres de quaisquer anotação, e produz dois tipos de arquivos: o primeiro em HTML, para facilitar a visualização da informação; e, o segundo, em XML, o qual promove a facilidade de processamento posterior da informação gerada. Nas Figuras 5.4 e 5.5 apresentamos as visualizações que são baseadas em html. A *tag cloud* exibe a primeira menção de cada cadeia, destacando (pelo tamanho da fonte) as cadeias com maior número de menções. Como, por exemplo, ‘genes’ e ‘a França’ (Figura 5.5) A lista de sintagmas constitui uma forma mais objetiva de exibição dos termos, apresentando-os em forma de lista. Nessa lista, cada cabeçalho possui a coloração e id específico de dada cadeia. A exibição das cadeias no texto se dá de acordo com a seleção realizada na *tag cloud*. Nesse caso, menções com mesma colocação e id pertencem a uma mesma cadeia de correferência. É possível exibir apenas uma cadeia selecionada, ou todas as cadeias de uma só vez. O exemplo da Figura 5.5 encontra-se disponível para interação, em nosso WebDemo[10]. Contudo, o CORP está disponível também na versão Desktop[11]. Em sua versão mais recente, o CORP utiliza como base todas as regras propostas nesta tese, em conjunto com nosso método de agrupamento, o qual utiliza o critério “Peso por Regra”

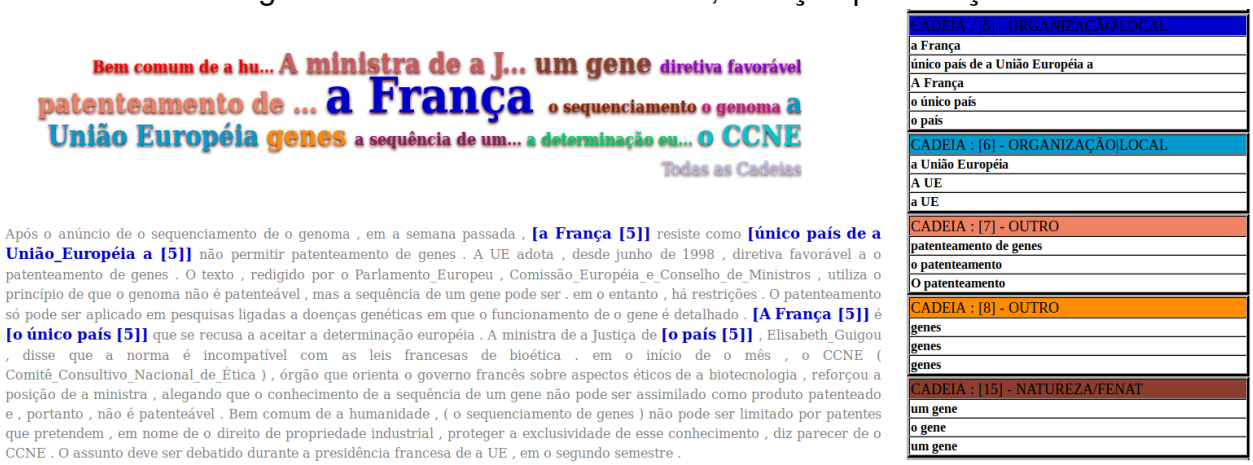
Figura 5.4 – CORP - Saída HTML, todas cadeias



[10] http://ontolp.inf.pucrs.br/corref/CIENCIA_2000_6381.txt.html

[11] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/corp-coreference-resolution-for-portuguese/

Figura 5.5 – CORP - Saída HTML, exibição por seleção



As informações contidas na saída XML são constituídas por: texto original submetido ao recurso, lista de sentenças identificadas e enumeradas, lista de tokens com suas marcações de POS, lista de cadeias de correferência e lista de menções únicas (menções que não possuem outros referentes no texto). Na Figura 5.6 é possível visualizarmos a estrutura de uma cadeia de correferência anotada pela ferramenta. Cada cadeia possui uma mesma estrutura, sendo ela constituída por uma lista de sintagmas, e cada sintagma, constituído por uma lista de tokens (word_id). É possível notarmos também, que o CORP provê a categoria semântica de cada sintagma, informações atualmente providas pelo Repentino e listas de entidades. Contudo, de forma a simplificar essas categorias semânticas, não utilizamos as subcategorias de entidades contidas no Repentino. A lista completa das classes que consideramos e o que cada classe representa pode ser vista na Seção 4.4.4.

Figura 5.6 – CORP - Saída XML

```
▼<Cadeia_5>
  ▼<sn id="4" tokens="15...16" nucleo="França" sintagma="a França" Categoria="ORGANIZAÇÃO|LOCAL" sentenca="1">
     <word_15 token="a" lemma="o" pos="art" features="F=S"/>
     <word_16 token="França" lemma="França" pos="prop" features="F=S"/>
   </sn>
  ▼<sn id="5" tokens="19...22" nucleo="país" sintagma="único país de a União Européia a" Categoria="ORGANIZAÇÃO|LOCAL"
   sentenca="1">
     <word_19 token="único" lemma="único" pos="adj" features="M=S"/>
     <word_20 token="país" lemma="país" pos="n" features="M=S"/>
     <word_21 token="de" lemma="de" pos="prp" features="-"/>
     <word_22 token="a" lemma="o" pos="art" features="F=S"/>
     <word_23 token="União_Européia" lemma="" pos="prop" features="F=S"/>
     <word_24 token="a" lemma="a" pos="prp" features="-"/>
   </sn>
  ▼<sn id="27" tokens="107...108" nucleo="França" sintagma="A França" Categoria="ORGANIZAÇÃO|LOCAL" sentenca="6">
     <word_107 token="A" lemma="o" pos="art" features="F=S"/>
     <word_108 token="França" lemma="França" pos="prop" features="F=S"/>
   </sn>
  ▼<sn id="28" tokens="110...112" nucleo="país" sintagma="o único país" Categoria="ORGANIZAÇÃO|LOCAL" sentenca="6">
     <word_110 token="o" lemma="o" pos="art" features="M=S"/>
     <word_111 token="único" lemma="único" pos="adj" features="M=S"/>
     <word_112 token="país" lemma="país" pos="n" features="M=S"/>
   </sn>
  ▼<sn id="34" tokens="128...129" nucleo="país" sintagma="o país" Categoria="ORGANIZAÇÃO|LOCAL" sentenca="7">
     <word_128 token="o" lemma="o" pos="art" features="M=S"/>
     <word_129 token="país" lemma="país" pos="n" features="M=S"/>
   </sn>
 </Cadeia_5>
```

5.6 Considerações do Capítulo

Neste capítulo foram apresentados detalhes referentes à nossa abordagem proposta, considerada a principal contribuição desta tese – um modelo para a resolução de correferência em Português, baseado em regras linguísticas e conhecimento semântico, concebido por meio de três recursos: Onto.PT, Repentino e Listas de entidades nomeadas. Além de propormos um modelo para a resolução de correferência em Português, apresentamos um novo método de agrupamento de menções que visa reduzir agrupamentos incorretos, considerando uma série de atributos. Mostramos que o atual método de agrupamento de menções utilizado na literatura possui deficiências, principalmente quando envolvemos conhecimento semântico, dado que a margem à ambiguidade é muito maior. Outra contribuição que julgamos significativa, é o CORP, um recurso funcional para a tarefa de resolução de correferência em Português que pode auxiliar em diversas tarefas de PLN.

No Capítulo 6 mostramos os experimentos conduzidos durante esta tese. Veremos que nosso método de agrupamento é superior ao atual estado da arte. Este aumentou a precisão de nosso modelo, sem grandes perdas em sua abrangência.

6. EXPERIMENTOS E RESULTADOS

Neste capítulo, apresentamos experimentos envolvendo três corpora, com anotação de correferência, disponíveis para o Português: Summ-it++, Corref-PT e o corpus de Garcia et al.[32]. Basicamente realizamos dois tipos de avaliação, a primeira, usando o Summ-it++, um corpus relativamente pequeno[1], porém contendo cadeias de correferência totalmente revisadas; e, a segunda, usando o Corref-PT, um corpus contendo aproximadamente oito[2] vezes mais cadeias que o Summ-it++, porém anotado semi-automaticamente.

Inicialmente, de forma a validar nosso conjunto de regras proposto (Seção 6), utilizamos o corpus Summ-it++. Lee et al. [39] realizaram esses mesmos experimentos, com o mesmo propósito: avaliar o comportamento e desempenho de suas regras propostas. Contudo, os autores utilizaram um conjunto de regras para o Inglês e o corpus Ontonotes [53]. Outra utilização do corpus Summ-it++ se deu na avaliação de nosso método de agrupamento (Seção 6.3). Nesses experimentos, avaliamos nosso método de agrupamento proposto, considerando 5 critérios de agrupamento distintos e comparamos seus resultados com o atual método utilizado por modelos de regras, o qual chamamos de “Baseline”.

Após os experimentos realizados com o corpus Sum-it++, realizamos experimentos com o corpus Corref-PT, com o objetivo de analisar o impacto que a semântica, em conjunto com nosso método de agrupamento, pode prover. O motivo de escolhermos o corpus Corref-PT para esses experimentos se dá pela sua quantidade de menções que possuem relações semânticas. No total, para o corpus Summ-it++, contabilizamos 775 menções que possuem relação de Sinonímia/Hiponímia[3] entre si. Já para o corpus Corref-PT foram encontradas 6219 menções. Dessa forma acreditamos que, mesmo o corpus Corref-PT não possuindo a mesma qualidade de anotação que o Summ-it++, este possui uma quantidade de amostras semânticas oito vezes maior, o que pode ser a melhor opção para tal experimento. Nos experimentos descritos usamos seis métricas amplamente utilizadas pela literatura (descritas em 2.5). Cada uma delas objetiva avaliar um aspecto específico no modelo e calcular seu desempenho. Como último experimento, realizamos uma análise comparativa entre o nosso modelo e o de Garcia et al. envolvendo dois textos de seu corpus. Após os experimentos descritos, realizamos uma análise detalhada dos principais erros encontrados em nosso modelo proposto. Nessa análise utilizamos 5 textos provenientes do corpus Summ-it++ e Corref-PT.

[1] 50 textos e 560 cadeias de correferência

[2] 182 textos e 3898 cadeias de correferência

[3] a quantidade de menções foi contabilizada tendo como base as relações existentes no Onto.PT

6.1 Avaliação das Regras Propostas

Conforme mencionado na introdução deste capítulo, nesta seção efetuamos dois tipos de avaliação com o corpus Summ-it++: na primeira (Tabela 6.1)[4], avaliamos os ganhos que cada regra pode prover ao modelo, de forma independente; na segunda (Tabela 6.2), avaliamos os ganhos que cada regra agrega ao modelo, de forma cumulativa. Para conceber esses experimentos utilizamos o método de agrupamento mais utilizado na literatura, o Baseline. O motivo de escolhermos o método Baseline para esses primeiros experimentos se dá pelo fato de que nosso método de agrupamento proposto considera sempre um conjunto de regras e atributos para construir as cadeias de correferência. Já o método Baseline é mais simples, bastando apenas que uma regra seja satisfeita para realizar o agrupamento de menções.

Nos experimentos descritos notamos que tanto no corpus Summ-it (corpus o qual originou o Summ-it++) quanto no corpus Summ-it++ o aposto e sua menção referente formam apenas uma menção. Dessa forma, sintagmas que aparecem na forma de aposto são considerados como uma única menção, como em: “o Instituto Nacional de Pesquisas Espaciais (INPE)...”. No corpus de referência temos apenas um sintagma [o Instituto Nacional de Pesquisas Espaciais (INPE)]. Já nosso modelo identifica como duas menções e as agrupa, formando uma cadeia: [o Instituto Nacional de Pesquisas Espaciais], [Inpe]. Dessa forma, em nossa avaliação, consideramos como acerto a criação dessa ligação nesses casos.

Tabela 6.1 – Regras individuais

	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CoNLL
	P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
CPadrExt. (R1)	66,4	22,8	34,0	68,0	19,1	29,8	64,5	26,5	37,6	50,5	28,1	36,1	83,2	64,5	68,4	33,3
CParcNcl. (R2)	61,9	30,7	41,1	63,3	25,8	36,7	58,9	34,6	43,6	47,3	37,0	41,5	80,6	59,9	62,1	39,8
ApstoExpl. (R3)	74,8	5,9	10,9	78,7	6,9	12,6	80,4	8,6	15,5	70,2	11,8	20,2	92,4	92,4	92,4	14,6
ApstoEsp. (R4)	11,1	0,4	0,7	22,3	0,7	1,4	32,6	1,4	2,8	26,9	1,8	3,5	57,5	57,3	57,3	1,9
Acronimo (R5)	58,8	0,7	1,4	65,5	0,7	1,5	75,9	1,1	2,2	66,7	1,2	2,5	65,1	63,9	63,6	1,8
PredNom. (R6)	18,2	0,1	0,3	34,1	0,1	0,3	50,0	0,5	1,1	26,5	0,4	0,9	47,7	48,2	44,4	0,5
PronRel. (R7)	0,0	0,0	0,0	11,8	0,1	0,3	21,0	0,4	0,8	17,7	0,5	1,0	47,2	46,9	46,4	0,4
CRestPNcl_1 (R8)	61,2	39,4	48,0	60,6	34,2	43,7	61,1	43,4	50,7	52,3	44,5	48,1	76,8	59,7	61,9	46,6
CRestPNcl_2 (R9)	61,1	39,8	48,2	60,5	34,6	44,0	61,3	43,8	51,1	52,4	44,9	48,4	76,7	59,7	61,9	46,9
CEntNomeProp. (R10)	70,2	7,8	14,0	73,0	6,7	12,3	78,6	10,1	17,9	62,4	10,4	17,8	85,9	85,9	85,9	14,7
CParcNomeProp. (R11)	66,7	8,1	14,4	69,7	7,3	13,3	77,4	10,6	18,7	64,3	11,0	18,8	81,7	85,2	83,3	15,5
Hiponimia (R12)	6,0	1,2	2,1	15,9	3,1	5,2	23,5	5,5	8,9	21,0	6,1	9,4	52,5	51,4	45,0	5,6
Sinonimia (R13)	28,5	13,7	18,5	24,3	12,8	16,8	34,1	16,1	21,9	28,5	12,9	17,8	57,5	53,6	50,0	17,7

Analisando a Tabela 6.1, podemos notar que as regras que lidam com o casamento de padrões entre palavras obtiveram precisões acima de 60%, tendo como destaque as regras 8 e 9 (Casamento Restrito pelo Núcleo), cujos resultados ultrapassaram 46% de score para a métrica CoNLL. Podemos notar também que a regra 3 (Aposto Explicativo)

[4]Nas Tabelas 6.1, 6.2, 6.5 e 6.6 ‘P’,‘A’ e ‘F’ representam respectivamente: Precisão, Abrangência e Medida-F.

possui uma alta precisão, no entanto ocorre com pouca frequência no corpus utilizado para teste. Referente às regras semânticas Hiponímia e Sinonímia (12 e 13), notamos que sinonímia apresenta melhores resultados do que hiponímia. Apesar de individualmente não apresentarem os melhores resultados, quando utilizadas em conjunto com outras regras, podemos ver ganhos na abrangência. Por meio de nossas regras semânticas, foi possível identificar ligações como:

- [fungos], [pequenos cogumelos];
- [cientistas], [pesquisadores];
- [a França], [o país].

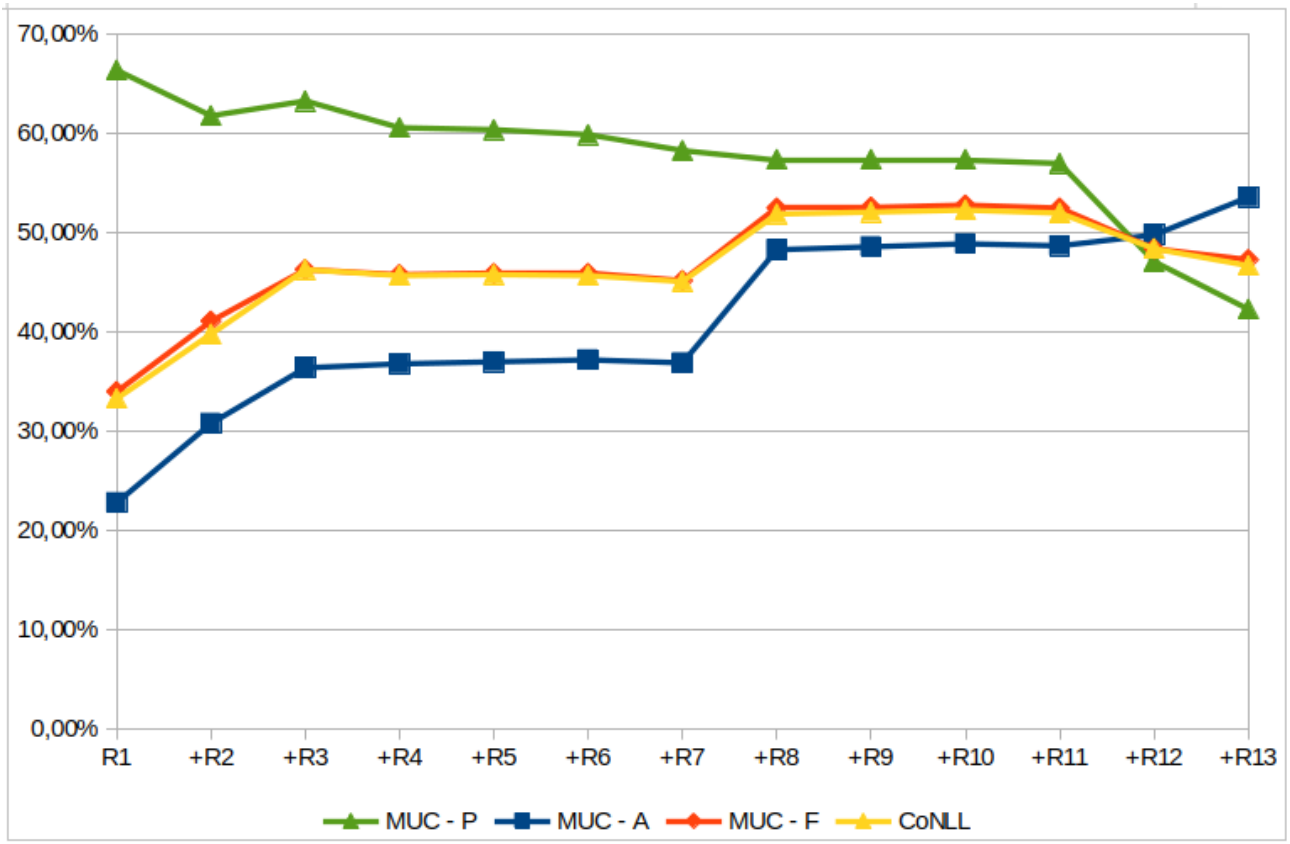
Tabela 6.2 – Regras cumulativas

	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CoNLL
	P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
CPadrExt. (R1)	66,4	22,8	34,0	68,0	19,1	29,8	64,5	26,5	37,6	50,5	28,1	36,1	83,2	64,5	68,4	33,3
+CParcNcl. (R2)	61,8	30,8	41,1	63,1	25,9	36,7	58,8	34,7	43,6	47,2	37,1	41,5	80,2	59,8	62,0	39,8
+ApstoExpl. (R3)	63,3	36,4	46,3	64,8	32,8	43,6	61,2	41,5	49,5	51,7	46,5	49,0	81,5	60,4	63,2	46,3
+ApstoEsp. (R4)	60,6	36,8	45,8	61,9	33,3	43,3	58,9	42,0	49,0	49,6	46,6	48,1	80,2	59,4	61,7	45,7
+Acronimo (R5)	60,4	37,0	45,9	61,7	33,5	43,4	58,7	42,2	49,1	49,6	46,8	48,1	79,9	59,3	61,6	45,8
+PredNom. (R6)	59,9	37,2	45,9	61,1	33,6	43,4	58,2	42,4	49,1	49,1	46,9	48,0	79,6	59,0	61,1	45,7
+PronRel. (R7)	58,3	36,9	45,2	59,7	33,5	42,9	56,8	42,2	48,4	47,7	46,5	47,1	78,9	58,4	59,9	45,1
+CRestPNcl_1 (R8)	57,4	48,3	52,5	56,2	44,6	49,7	57,8	53,2	55,4	51,5	55,9	53,6	75,0	57,7	59,0	51,9
+CRestPNcl_2 (R9)	57,4	48,6	52,6	56,2	44,8	49,8	57,9	53,4	55,6	51,6	56,2	53,8	75,0	57,7	59,0	52,1
+CEntNomeProp. (R10)	57,4	48,9	52,8	56,2	45,1	50,0	57,9	53,8	55,8	51,8	56,5	54,0	75,0	57,7	58,9	52,3
+CParcNomeProp. (R11)	57,0	48,7	52,5	55,4	45,1	49,7	57,9	53,5	55,6	52,0	55,7	53,8	74,1	57,8	59,1	52,0
+Hiponimia (R12)	47,1	49,8	48,4	44,6	46,9	45,7	49,9	53,3	51,6	48,9	53,4	51,1	65,2	55,7	55,5	48,4
+Sinonimia (R13)	42,3	53,6	47,3	38,7	50,8	43,9	45,2	55,6	49,9	45,6	52,8	48,9	62,9	54,6	53,3	46,7

Por meio da tabela 6.2 e Figura 6.1, podemos inferir que a cada nova regra adicionada o modelo perde precisão, mas ganha em abrangência, aumentando, na maioria dos casos, sua medida-F. Adicionalmente, quando acrescentamos semântica ao modelo, há uma redução na medida-F. Isso pode ser facilmente visualizado na Figura 6.1. Na figura, veja que há um grande salto em precisão (negativo) e abrangência (positivo), quando adicionamos as regras semânticas. Veja também que, embora a métrica CoNLL seja composta pela média de três métricas distintas, a medida-F da métrica MUC, para este cenário, representou bem os resultados, sendo estes muito próximos a cada iteração(adição de regra).

Em experimentos realizados, descritos na Seção 6.4 veremos que com nosso novo método de agrupamento proposto, foi possível melhorar a precisão e a medida-f de nosso modelo, com uma perda mínima (quando comparada aos ganhos) de abrangência.

Figura 6.1 – Avaliação cumulativa considerando as métricas MUC e CoNLL



6.2 Avaliação do Método de Agrupamento Proposto

Nesta seção realizamos experimentos de forma a verificar a precisão, abrangência e medida-F de nosso Método de agrupamento. Em nossos experimentos avaliamos cinco variações de nosso método proposto, (cada uma delas usando um critério de agrupamento distinto) e comparamos com o atual estado da arte[5] (Baseline). Na Tabela 6.3, é possível visualizarmos os resultados para cada método/critério.

Tabela 6.3 – Avaliação dos Critérios de Agrupamento de Menções (Corpus Summ-it++)

Método	Critério	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CONLL
		P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
Baseline	-	40,6	55,0	46,7	35,9	52,4	42,6	43,0	54,3	48,0	47,1	50,2	48,6	60,4	54,9	54,1	46,0
Nosso	Cadeia mais Próxima	42,3	48,0	45,0	42,5	46,6	44,4	46,1	55,0	50,2	43,1	56,7	49,0	67,7	53,9	51,3	46,1
	Peso por Regra	45,1	52,1	48,3	43,8	49,5	46,5	48,0	57,0	52,1	45,7	57,4	50,9	68,2	54,7	53,0	48,6
	Peso por Menção	44,3	51,1	47,5	42,7	48,6	45,5	47,6	56,0	51,5	45,3	55,7	50,0	66,7	54,5	52,7	47,7
	Peso (Regras + Menções)	45,0	52,0	48,2	43,6	49,3	46,3	48,0	56,8	52,0	45,7	56,9	50,7	68,0	54,7	53,0	48,4
	Peso por F-Score	45,0	52,0	48,2	43,9	49,5	46,5	52,3	57,3	52,3	45,3	57,8	50,8	68,6	54,6	52,8	48,5

De acordo com os resultados exibidos, todos os cinco critérios propostos apresentaram precisões superiores ao atual método utilizado (Baseline), exceto para a métrica $CEAF_e$. Contudo, para este caso, todos os critérios obtiveram melhor abrangência. Anali-

[5]No que diz respeito a modelos baseados em regras e resolução de correferência [39, 31, 25, 23]

sando especificamente a métrica CoNLL, podemos dizer que todos os critérios propostos superaram o atual “Baseline”. Em especial, o critério “Peso por Regra” apresentou o melhor resultado, ultrapassando o método “Baseline” em 2,6%. Na figura 6.2 podemos ver um gráfico comparativo envolvendo as medidas-F de cada critério/método.

É possível vermos que o critério “Peso por Regra” apresentou os melhores resultados, seguido pelo critério “Peso por F-Score”. A razão do método “Baseline” ser melhor que o critério “Peso por Regra” na medida-F da métrica BLANC se dá pelo fato desta métrica ser calculada usando a média da medida-F entre as ligações(*links*) de correferência e de não-correferência. Na Tabela 6.4, é possível visualizar que o critério “Peso por Regra” perde apenas em abrangência nas ligações de correferência. Isto já era esperado, dado que nosso método e critérios de agrupamento propõem uma geração de ligações/*links* mais restrita que o método Baseline.

Figura 6.2 – Análise comparativa entre Medidas-F

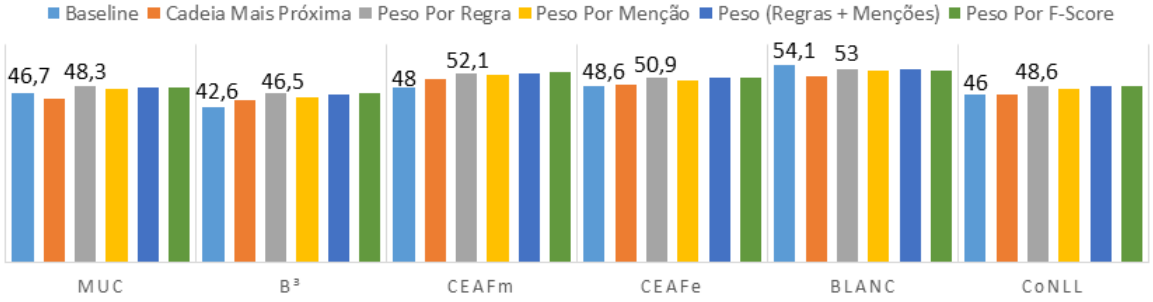


Tabela 6.4 – BLANC – Ligações de correferência e não-correferência entre nosso método proposto, usando o critério Peso por Regra, e o método Baseline

Método	Ligações de Correferência			Ligações de Não-Correferência		
	P	A	F	P	A	F
Baseline	45,4	17,6	25,3	75,3	92,3	82,9
Nosso – Peso por Regra	59,5	12,1	20,2	76,8	97,2	85,8

6.3 Análise Comparativa entre o Método de Agrupamento Proposto e Baseline

Nesta seção apresentamos uma análise comparativa, referente ao comportamento de nosso método de agrupamento, utilizando o critério “Peso por Regra” e o método Baseline. No exemplo abaixo temos um fragmento de texto em que os sintagmas considerados na análise estão em destaque:

- “*...biotecnologias apropriadas ao desenvolvimento d[o país$_1$]. Guerra citou a micropropagação de vegetais (produção de mudas em laboratório feita para evitar doenças e selecionar [vegetais saudáveis$_2$]) como exemplo de biotecnologia de baixo custo.*

Com ela, aumentou-se a produção de moranguinho, n[o sul$_3$] d[o país$_4$], de 3,2 kg para 60 kg por hectare. Para o agrônomo, [o Brasil$_5$] deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de [plantas$_6$] com a capacidade de captar certos elementos presentes n[a terra$_7$]. O presidente d[a Embrapa$_8$] ([Empresa Brasileira de Pesquisa Agropecuária$_9$]), Alberto Portugal, salientou que [a empresa$_{10}$] busca soluções para os problemas da agricultura nacional..."

No fragmento de texto podemos observar duas cadeias de correferência específicas:

Referência (*gold*):

- C1= *{[o país$_1$], [o país$_4$], [o Brasil$_5$]}*

- C2=*{[a Embrapa$_8$], [Empresa Brasileira de Pesquisa Agropecuária$_9$], [a empresa$_{10}$]}*

Cadeias obtidas pelo método Baseline:

- C1=*{[o país$_1$], [vegetais saudáveis$_2$], [o sul$_3$], [o país$_4$], [o Brasil$_5$], [plantas$_6$], [a terra$_7$], [a Embrapa$_8$], [Empresa Brasileira de Pesquisa Agropecuária$_9$], [a empresa$_{10}$]}*

- C2=*{Não houveram elementos}*

Cadeias obtidas pelo método Peso por Regras:

- C1=*{[o país$_1$], [o país$_4$], [a terra$_7$]}*

- C2=*{[a Embrapa$_8$], [Empresa Brasileira de Pesquisa Agropecuária$_9$], [a empresa$_{10}$]}*

Analisando os resultados do método “Baseline”, podemos ver que a cadeia ‘C2’ não foi identificada. Isso porque existe, pelo menos uma regra com valor verdade, que liga ‘C1’ e ‘C2’. Como exemplo disso, em nossa base de dados semântica [33] existem triplas, indicando as seguintes relações: [terra, país, sinonímia], [planta, vegetal, sinonímia], [planta, empresa, hiponímia]. Note que temos um problema de ambiguidade nesse caso, e o método “Baseline” não considera o contexto como nosso método proposto. Relações semânticas como essas podem introduzir muitas falsas ligações. Este exemplo serve para mostrar a importância da representação do discurso. Sobre o nosso método, é possível notar que o mesmo não agrupou o sintagma [o Brasil$_5$] e agrupou [a terra$_7$] usando duas ligações de sinonímia.

6.4 Análise Comparativa Envolvendo Semântica e Métodos de Agrupamento

De forma a validar os ganhos que a semântica e nosso algoritmo de agrupamento podem prover quando utilizados em conjunto, realizamos experimentos adicionais, envolvendo o Corpus Corref-PT e diferentes combinações entre os métodos de agrupamento e semântica. No primeiro, envolvemos o atual método proposto pelo atual estado da arte (Baseline), sem considerarmos o uso de semântica; em nosso segundo experimento consideramos o método Baseline e conhecimento semântico; no terceiro, utilizamos nosso método de agrupamento proposto sem o uso de semântica; no quarto experimento, envolvemos o uso de nosso método de agrupamento e do conhecimento semântico juntos (Tabela 6.5).

Tabela 6.5 – Experimentos envolvendo semântica e métodos de agrupamento

Método	Semântica	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CoNLL
		P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
Baseline	X	55,4	49,5	52,3	49,3	42,4	45,6	52,2	48,7	50,4	46,9	48,8	47,8	65,4	54,2	51,9	48,6
	$\surd$	44,2	52,2	47,9	35,8	45,8	40,2	41,5	46,5	43,9	46,1	43,9	44,9	59,3	55,2	54,0	44,3
Nosso (Peso por Regra)	X	64,2	47,8	54,8	61,2	40,5	48,7	59,9	49,0	53,9	50,2	51,0	50,6	79,8	56,1	55,5	51,4
	$\surd$	54,9	50,2	52,5	51,8	43,6	47,3	52,9	51,6	52,3	46,2	52,8	49,3	73,3	53,8	50,2	49,7

Na tabela 6.5 “X” e “$\surd$” representam respectivamente “sem[6]” e “com” o uso de semântica. Podemos visualizar de forma geral que, quando envolvemos o uso de conhecimento semântico, a abrangência dos modelos sofre um aumento significativo. Contudo, ocorre uma perda substancial em precisão. Dentro desse contexto, considerando o método Baseline, ocorreu um aumento de 2,7% em abrangência, acompanhado de uma perda de 11,2% em precisão (métrica MUC). Em nosso método proposto esse comportamento repetiu-se. Contudo podemos notar que, mesmo nosso método obtendo uma abrangência inferior em 2,2%, nossa precisão foi superior em 10,7%, obtendo uma medida-F superior em 4,6%. Podemos notar também que sempre que há a introdução de conhecimento semântico ocorre uma perda significativa na precisão, independente do caso. Se compararmos as medidas-F dos dois métodos de agrupamento sem semântica, podemos ver que estas são maiores que os resultados envolvendo semântica. Contudo, mesmo com essa perda, acreditamos que o uso da semântica pode trazer ganhos à tarefa. Dado que torna possível identificar casos que vão além da combinação parcial/total entre palavras e de técnicas de justaposição.

[6]Nos Experimentos “sem” semântica utilizamos todas regras propostas exceto Sinonímia e Hiponímia. Já nos experimentos “com” semântica utilizamos todas as 13 regras propostas.

6.5 Resultados Não Comparativos entre o Modelo Proposto e os Principais Trabalhos Relacionados

Na Tabela 6.6, temos os resultados dos principais trabalhos encontrados na literatura, avaliados utilizando as métricas da conferência CoNLL. Esses são resultados não comparativos, dado que são provenientes de idiomas e/ou escopos distintos. Embora que não possamos comparar nosso modelo aos demais, na Tabela 6.6 podemos comparar as avaliações realizadas com os corpora Summ-it++ e Corref-PT. Para essas, é possível notar que a diferença entre os resultados é de apenas 1,1% para a métrica CoNLL, o que nos faz refletir positivamente a respeito da qualidade de anotação do corpus Corref-PT, dado que os resultados de sua avaliação ficaram muito próximos aos da avaliação com o Summ-it++. Nessa comparação, os resultados foram obtidos usando todas as regras de nosso modelo, em conjunto com nosso método de agrupamento proposto, tendo como base o critério de agrupamento "Peso por Regra". Resultados omissos na tabela remetem à ausência de informação nos respectivos trabalhos.

Tabela 6.6 – Resultados não comparativos de nosso e dos principais modelos da literatura

		Detecção de Menções			MUC			B^3			$Ceaf_e$			CoNLL
Modelo	Idioma	P	A	F	P	A	F	P	A	F	P	A	F	F
Martschat et al., 2015	IN	–	–	–	76,8	68,1	72,2	66,1	54,2	59,6	59,5	52,3	55,7	62,5
Fernandes et al., 2014	IN	24,0	90,2	37,9	75,9	65,8	70,5	77,7	65,8	71,2	43,2	55,0	48,4	63,4
	CH	22,4	84,1	35,4	71,5	59,2	64,8	80,5	67,2	73,2	45,2	57,5	50,6	62,9
	AR	12,3	81,1	21,3	49,7	43,6	46,5	72,2	62,7	67,1	46,1	52,5	49,1	54,2
Lee et al., 2013	IN	66,8	75,1	70,7	60,9	59,6	60,3	73,3	68,6	70,9	46,2	47,5	46,9	59,4
Garcia et al., 2014	ES	–	–	–	94,1	84,1	88,8	84,8	62,9	72,2	71,0	83,4	76,7	79,2
	GL	–	–	–	94,6	89,0	91,7	88,4	72,9	79,9	76,6	87,6	81,7	84,4
	PT	–	–	–	92,7	82,7	87,4	84,5	65,8	74,0	67,9	84,4	75,2	78,9
Nosso (Summ-it++)	PT	62,0	63,2	62,6	45,1	52,1	48,3	43,8	49,5	46,5	45,7	57,4	50,9	48,6
Nosso (Corref-PT)	PT	60,4	63,6	62,2	54,9	50,2	52,5	51,8	43,6	47,3	46,2	52,8	49,3	49,7

6.6 Análise Comparativa entre Nosso Modelo e o Modelo de Garcia et al.

Mesmo tendo em mente a diferença de escopo entre o nosso modelo proposto e o de Garcia, realizamos uma avaliação comparativa, envolvendo dois textos, escritos em Português, provindos do corpus de Garcia et al. [32]. Nesse comparativo ambos os modelos utilizaram as menções *gold*[7] previamente anotadas no corpus.

Na Tabela 6.7 é possível notar que nosso modelo foi muito inferior em abrangência; algo que já esperávamos, dado que nosso modelo não trata pronomes pessoais: algo que é foco do modelo e do corpus utilizado por Garcia et al.. Contudo, podemos notar que

[7] No corpus de Garcia et al. apenas menções da categoria Pessoa estão anotadas.

Tabela 6.7 – Análise comparativa de nosso modelo e o de Garcia et al.

Modelo	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CONLL
	P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
Garcia	97,9	96,0	97,0	86,44	81,9	84,1	86,4	86,4	86,4	85,8	95,3	90,3	83,1	83,4	83,1	90,5
Nosso	80,0	16,0	26,7	81,7	7,8	14,2	73,3	18,6	29,7	33,4	18,6	23,9	90,7	83,7	86,0	21,6

nosso modelo obteve valores de precisão relativamente altos para as métricas MUC e B-Cubed (80% e 81,7%, respectivamente) e; para a métrica BLANC, podemos notar que nosso modelo foi superior ao de Garcia et al.. Isso deve-se ao fato da métrica BLANC realizar uma média entre as ligações de correferência e de não correferência. Em poucas palavras, nosso modelo obteve poucas (porém precisas) ligações de correferência. Outro fator a ser considerado é que nosso modelo identificou e anotou corretamente duas cadeias de correferência (uma do tipo “Organização/Local”, contendo 4 menções e outra do tipo “Outro”, contendo duas menções) que não foram identificadas pelo modelo de Garcia et al.; tão pouco encontram-se anotadas em seu corpus de referência, dado que essas cadeias não referem-se à categoria Pessoa.

Acreditamos que para tornar essa comparação mais justa, seria necessário envolvermos outros corpora anotados, como Summ-it++ e Corref-PT. O corpus Summ-it++ possui 5047 menções. Dessas, apenas 226 são categorizadas como pessoa; o corpus Corref-PT possui 33514 menções sendo essas, apenas 3119 do tipo Pessoa. Dito isso, mesmo que o modelo de Garcia et al. reconheça e agrupe corretamente todas essas entidades, seu modelo cobriria apenas 4,5% do corpus Summ-it++ e 9,3% do corpus Corref-PT. Nosso modelo possui abrangências superiores a 50%. Contudo, diferente de nosso modelo, que obtém correferência a partir de textos simples, livres de quaisquer anotação; o modelo de Garcia et al. necessita de uma série de informações previamente anotadas para obtenção das cadeias, tais como: part-of-speech tagging, parsing de dependência, categoria semântica das entidades nomeadas e menções previamente identificadas.

6.7 Análise de Erros

Nesta seção, apresentamos uma análise detalhada dos erros mais frequentes ocorridos durante a obtenção de cadeias de correferência, concebidas de forma automática por nosso modelo proposto. Para efetuar a análise selecionamos cinco textos pertencentes a dois corpora (Summ-it++ e Corref-PT [20]).

Veremos que os tipos mais comuns de erro são devidos ao processo de detecção de menções; à relações semânticas existentes entre duas ou mais menções, sem que estas sejam correferentes no contexto existente; à separação de uma cadeia de correferência em duas e às regras de casamento de padrões:

6.7.1 Texto 1

Após o anúncio de **[o sequenciamento [1]]** de **[o genoma [2]]** , em a semana passada , **[a França [5]]** resiste como **[único país de [a União_Européia [6]] a [5]]** não permitir **[patenteamento de [genes [8]] [7]]** . **[A UE [6]]** adota , desde junho de 1998 , **[diretiva favorável [10]]** a **[o patenteamento [7]]** de **[genes [8]]** . O texto , redigido por o Parlamento_Europeu , Comissão_Européia_e_Conselho_de_Ministros , utiliza **[o princípio de que [10]] [o genoma [2]]** não é patenteável , mas **[a sequência de [um gene [15]] [14]]** pode ser . em o entanto , há restrições . **[O patenteamento [7]]** só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento de **[o gene [15]]** é detalhado . **[A França [5]]** é **[o único país [5]]** que se recusa a aceitar **[a determinação européia [22]]** . **[A ministra de a Justiça de [o país [5]] , [23]] [Elisabeth_Guigou [23]]** , disse que **[a norma [22]]** é incompatível com as leis francesas de bioética . em o início de o mês , **[o CCNE ([29]] [Comitê_Consultivo_Nacional_de_Ética [29]]**) , órgão que orienta o governo francês sobre aspectos éticos de a biotecnologia , reforçou a posição de **[a ministra [23]]** , alegando que o conhecimento de **[a sequência [14]]** de **[um gene [15]]** não pode ser assimilado como produto patenteado e , portanto , não é patenteável . **[Bem comum de a humanidade , ([o sequenciamento de [genes [8]][1]]) [38]]** não pode ser limitado por patentes que pretendem , em nome de **[o direito [38]]** de propriedade industrial , proteger a exclusividade de esse conhecimento , diz parecer de **[o CCNE [29]]** . O assunto deve ser debatido durante a presidência francesa de **[a UE [6]]** , em o segundo semestre .

Cadeias Extraídas:

1.) [o sequenciamento], [o sequenciamento de genes];
2.) [o genoma], [o genoma];
5.) [a França], [único país de a União Européia a], [A França], [o único país], [o país];
6.) [a União Européia], [A UE], [a UE];
7.) [patenteamento de genes], [o patenteamento], [O patenteamento];
8.) [genes], [genes], [genes];
10.) [diretiva favorável], [o princípio de que];
14.) [a sequência de um gene], [a sequência];
15.) [um gene], [o gene],[um gene];
22.) [a determinação européia], [a norma];
23.) [A ministra de a Justiça de o país], [Elisabeth Guigou], [a ministra];
29.) [o CCNE], [Comitê Consultivo Nacional de Ética], [o CCNE];
38.) [Bem comum de a humanidade , (o sequenciamento de genes)], [o direito].

Análise:

Analisando cadeias do texto 1, podemos notar que alguns dos erros encontrados foram decorrentes do truncamento de alguns sintagmas nominais, como na cadeia 5, em que os sintagmas [o único país que se recusa a aceitar a determinação europeia] e [único país de a União Europeia a não permitir patenteamento de genes] não foram reconhecidos por completo. Contudo, podemos notar que foram corretamente agrupados. Na cadeia 7, podemos ver que o sintagma [O assunto] não foi agrupado. Contudo, acreditamos ser algo aceitável, dado que o termo "assunto" é bastante abrangente e, em nossa base semântica utilizada, não está relacionado com as outras menções da cadeia em questão. Podemos notar também que na cadeia 7 ocorre o mesmo problema de truncamento de menções, com o sintagma [o patenteamento], em que o correto seria [o patenteamento de genes]. Encaramos este problema como uma limitação, dado que a correta identificação das menções depende do processo de *chunking*, realizado pelo parser CoGrOO.

Outro tipo de erro encontrado foi a separação de uma cadeia de correferência em duas (cadeias 8 e 15). Para esse caso, como os termos "gene" e "genes" não possuem mesma grafia, o único meio de agrupá-los é por meio de seu lema. Contudo nossas regras semânticas[8] utilizam restrição de número (singular/plural) (veja a Seção 5.4.2). Dessa forma, como a menção [um gene] não concorda em número com as menções da cadeia 8, nosso algoritmo de agrupamento assume que ela é nova no discurso e cria uma nova cadeia de correferência. Na cadeia 10, nosso modelo agrupou incorretamente os sintagmas [diretiva favorável] e [o principio de que]. Isso porque os termos "diretiva" e "principio" apresentam uma relação de hiponímia em nossa base semântica (Onto.PT).

Referente as cadeias 22 e 38: para a cadeia 22, nosso modelo não agrupou nem reconheceu corretamente o sintagma [o princípio de que o genoma não é patenteável, mas a sequência de um gene pode ser]. Este foi agrupado incorretamente na cadeia 10. Contudo, os sintagmas [a determinação europeia] e [a norma] foram agrupados corretamente. E, por fim, na cadeia 38 houve um agrupamento incorreto pelo fato de a regra Hiponímia apresentar um valor verdade na ontologia Onto.PT, para os termos "bem" e "direito".

[8] apenas as regras de Sinonímia e Hiponímia utilizam o lema do núcleo.

6.7.2 Texto 2

O ministro [Roberto_Rodrigues [1]] ([Agricultura [1]]) anunciou ontem [o nascimento de [a bezerra [3]] [Vitoriosa [3]] [2]] . [O animal [4]] é [um clone [4]] gerado a_partir_de um clone a [vaca [5]] Vitória , que havia sido clonada em 2001 . Para [Rodrigues [1]] , a cria coloca a genética de o país em destaque em o cenário mundial . [Clayton_Campanhola , diretor-presidente de [a Embrapa [15]] , [13]] afirma que [o método [16]] ajudará em a multiplicação de animais de elevado valor genético ou em a reprodução de os ameaçados de extinção . Se há um animal de boa qualidade genética , a gente consegue manter isso em um filho (clonado) de o animal , mesmo que esteja velho . É a reprodução de a qualidade. . Segundo [Campanhola [13]] , [a técnica [16]] pode ser aplicada imediatamente em a produção de carne e de leite . [A técnica [16]] existe , pode ser utilizada e já foi testada . Agora é [uma questão de [34]] aplicar e de divulgar melhor [esse conhecimento. [34]] . [Vitoriosa [3]] é [o resultado de um experimento [3]] realizado por [a Embrapa [15]] ([Empresa_Brasileira_de_Pesquisa_Agropecuária [15]]) . Ela surgiu a_partir_de células isoladas de um pedaço de pele retirado de a orelha de [a vaca [5]] [Vitória [5]] , que foi [o primeiro clone bovino de a América_Latina , nascida [41]] em 2001 . [O clone de o clone [41]] coloca o Brasil em a vanguarda científica de esse assunto , como já está em [o seqüenciamento [47]] ([soletração [47]]) de genoma , afirmou [Rodrigues [1]] . em esse experimento , foram produzidos 35 embriões em seguida transferidos para 17 receptoras , as chamadas mães de aluguel . Vitoriosa , que tem 15 dias , é a terceira tentativa de o órgão de criar [um clone [4]] a_partir_de outro . em o ano passado , duas cópias de [Vitória [5]] morreram , uma em o oitavo mês de gestação e outra 36 horas depois de [o nascimento [2]] .

Cadeias Extraídas:

1.) [Roberto Rodrigues],[Agricultura], [Rodrigues], [Rodrigues];
3.) [a bezerra], [Vitoriosa], [Vitoriosa], [o resultado de um experimento];
5.) [vaca], [a vaca], [Vitória], [Vitória];
4.) [O animal], [um clone], [um clone];
15.) [a Embrapa], [a Embrapa], [Empresa Brasileira de Pesquisa Agropecuária];
16.) [o método], [a técnica], [A técnica];
2.) [o nascimento de a bezerra Vitoriosa], [o nascimento];
13.) [Clayton Campanhola , diretor-presidente de a Embrapa], [Campanhola];
34.) [uma questão de], [esse conhecimento];
41.) [o primeiro clone bovino de a América Latina , nascida], [O clone de o clone];
47.) [o seqüenciamento], [soletração].

Análise:

Na cadeia 1, podemos ver que o sintagma “Agricultura” foi agrupado incorretamente com os sintagmas [Roberto Rodrigues], [Rodrigues] e [Rodrigues]. Isso ocorre pelo fato de o sintagma “[Agricultura]” estar entre parênteses após o nome “Roberto Rodrigues”. Dessa forma, nosso modelo identifica tal menção como Aposto Explicativo (ver regra 3, Seção 5.4.1). Podemos notar também que as cadeias 3 e 4 foram separadas. Isso ocorreu pelo fato de não existir nenhuma regra que ligasse o sintagma [O animal] às menções da cadeia 3. Assim, nosso algoritmo de agrupamento reconheceu a menção como nova no discurso, formando uma nova cadeia de correferência. Ainda na cadeia 4, podemos notar que o sintagma [um clone gerado a_partir_de um clone] foi truncado. A última menção do sintagma [um clone] (. . . a terceira tentativa de criar um clone. . .) não faz referência a [a bezerra] e a [O animal] (cadeias 3 e 4), haja vista que o artigo indefinido gera uma expressão genérica, em que se pode fazer referência a qualquer clone no mundo real. Na cadeia 34, além da menção ter sido truncada pelo parser, temos um agrupamento incorreto, ocorrido por meio da regra Predicado Nominativo. Basicamente, essa regra busca por padrões como “verbo de ligação + artigo”. Se analisarmos a sentença: *“Agora é uma questão de aplicar e de divulgar melhor esse conhecimento.”*; veremos que temos a sequência de tokens “é uma”. Contudo, como o token “Agora” representa um sintagma adverbial, nosso modelo agrupou as duas menções subsequentes.

6.7.3 Texto 3

[O ministro de a Defesa , [0]] [Nelson_Jobim [0]] , deve encaminhar o nome de **[a economista [3]] [Solange_Vieira [3]]** para assumir **[uma de as diretorias de [a Agência_Nacional_de_Aviação_Civil [5]] ([4]] [Anac [5]]**) . Ainda não está definida **[a diretoria [4]]** que **[a economista [7]]** vai assumir . . Inicialmente , **[Solange_Vieira , que [3]]** é assessora especial de **[Jobim [0]]** havia sido escolhida para comandar a Secretaria_Nacional_de_Aviação_Civil , a ser criada em a estrutura de **[o ministério [3]]** , segundo a assessoria de imprensa de **[o ministério [3]]** . Mas , diante_da a dificuldade para encontrar pessoas que aceitassem assumir **[uma de as diretorias de a agência reguladora , após [a renúncia [18]] de três diretores [4]]** , **[Jobim [0]]** decidiu indicar **[a economista para [o cargo [3]] [7]]** . . Uma de a três vagas será ocupada por o major-brigadeiro Allemander_Jesus_Pereira_Filho , indicado para exercer o cargo em substituição a Jorge_Luiz_Brito_Velozo , que pediu **[demissão [18]]** em o final de o mês passado . Também renunciaram a **[o cargo [3]]** de diretor de a Anac_Denise_Abreu e Leur_Lomanto .

Cadeias Extraídas:

0.) [O ministro de a Defesa], [Nelson Jobim], [Jobim], [Jobim];

3.) [a economista], [Solange Vieira], [Solange Vieira, que], [o ministério], [o ministério], [o cargo], [o cargo];

4.) [uma de as diretorias de a Agência Nacional de Aviação Civil], [a diretoria], [uma de as diretorias de a agência reguladora , após a renúncia de três diretores];

5.) [a Agência Nacional de Aviação Civil], [Anac];

7.) [a economista], [a economista para o cargo];

18.) [a renúncia], [demissão].

Análise:

Na cadeia 3 podemos notar que o modelo agrupou incorretamente duas ocorrências do sintagma [o cargo] (os quais deveriam pertencer à cadeia 4, pois remetem a [uma de as diretorias]) e duas ocorrências do sintagma [o ministério]. Para estas, o correto seria criar uma nova cadeia com suas ocorrências ([o ministério], o ministério]). Esse agrupamento incorreto ocorreu devido à presença de relações semânticas entre os sintagmas (economista, cargo e ministério) no Onto.PT. Podemos notar que os sintagmas [a economista] e [a economista para o cargo] foram separados em uma nova cadeia (cadeia 7). Isso ocorreu pelo fato de as regras de "casamento de padrões" utilizarem sub-cláusulas restritivas, de forma a não permitir o agrupamento por meio do casamento de padrões quando uma menção está em uma construção de aposto especificativo (ver regras 1, 2 e 4 na Seção 5.4.1); dessa forma, o sintagma [a economista] da cadeia 7 não foi agrupado à cadeia 3. Como estes não foram considerados correferentes, foi gerada uma nova cadeia. Embora a restrição da cláusula de aposto especificativo em conjunto da regra de Casamento de Padrões possa parecer negativa, ela aumentou a precisão de nosso modelo, evitando que adjuntos adnominais sejam agrupados erroneamente devido ao truncamento de sintagmas realizado pelo parser, como veremos na análise do texto a seguir. Podemos notar, ainda, na cadeia 18, que nosso modelo agrupou [a renúncia] e [demissão] incorretamente. Mesmo os termos possuindo relação de Sinonímia, [a renúncia] remete à renúncia de três diretores; já [demissão], remete à demissão de "Jorge Luiz Brito Velozo" apenas.

6.7.4 Texto 4

George_Smoot é figurinha carimbada em o estudo de **[a radiação cósmica [3]]** de **[fundo [4]]** . Foi o principal cientista ligado a **[o satélite [6]] [Cobe [6]]** (**[Cosmic_Background_Explorer [6]]**) , que em 1992 revelou **[flutuações de temperatura [7]]** em **[a radiação [3]]** em a verdade um eco de **[o Big_Bang [11]]** , explosão que teria dado origem a **[o Universo [16]]** . Ela foi descoberta em 1965 , depois de ter sido prevista por o modelo de **[o Big_Bang [11]]** criado por **[o russo-americano [19]] [George_Gamow [19]]** . Mas ninguém conseguira detectar **[flutuações [7]]** até **[o Cobe [6]]** . A_partir_das **[as flutuações [7]]** , é possível estimar a quantidade de **[matéria [4]]** e energia existente em **[o Universo [16]]** e a idade de **[o cosmos [16]]** , entre outras descobertas importantes . Os resultados de **[o Cobe [6]]** foram relevantes , mas o estado de a arte são as imagens de **[o satélite [29]] [WMAP [29]]** (**[Wilkinson_Microwave_Anisotropy_Probe [29]]**) , **[cujos [29]]** primeiros dados saíram em 2003 . . (SN) .

Cadeias Extraídas:

3.) [a radiação cósmica], [a radiação];
4.) [fundo], [matéria];
6.) [o satélite], [Cobe], [Cosmic Background Explorer], [o Cobe], [o Cobe];
7.) [flutuações de temperatura], [flutuações], [as flutuações];
11.) [o Big Bang], [o Big Bang];
16.) [o Universo], [o Universo], [o cosmos];
19.) [o russo-americano], [George Gamow];
29.) [o satélite], [WMAP], [Wilkinson Microwave Anisotropy Probe], [cujos].

Análise:

Analisando as cadeias geradas e o texto, podemos notar que nossa regra de predicado nominativo não identificou [George_Smooth] e [figurinha carimbada]. Isso deu-se pelo fato do sintagma [figurinha carimbada] não possuir um determinante. A regra também não agrupou [o principal cientista ligado a o satélite Cobe (Cosmic_Background_Explorer)] pelo fato de essa estar em uma nova sentença. Na cadeia 3 temos a perda do pronome [Ela]. Esse não é considerado um erro para o escopo do modelo, já que, até o momento, não utilizamos regras de agrupamento de pronomes pessoais. Na cadeia 4 temos um agrupamento incorreto por meio da regra sinonímia entre [fundo] e [matéria]; Na cadeia 16 temos uma

relação de sinonímia entre os sintagmas [o Universo] e [o Cosmos]. Podemos ver que, de forma ampla, esta premissa é verdadeira, mas em textos de domínios da astronomia, esses termos possuem definições específicas, tornando-os entidades distintas portanto, não correferentes. Nas cadeias 6 e 29 (agrupadas corretamente), temos a situação em que a cláusula restritiva de não permitir o agrupamento de sintagmas por meio do casamento de padrões melhora a precisão de nosso modelo. Note que sem essa restrição a menção [o satélite], pertencente a cadeia 29, seria agrupada com a cadeia 6.

6.7.5 Texto 5

O estado de **[São_Paulo [1]]** voltou a sofrer com **[os ataques [2]]** contra postos policiais , agências bancárias e ônibus , em a madrugada de esta segunda-feira . **[As ações [2]]** são atribuídas a a facção criminosa Primeiro_Comando de **[a Capital [9]]** (**[PCC [9]]**) , que já comandou outros ataques em duas ocasiões . Os ataques de esta madrugada , até agora , não deixaram mortos ou feridos . **[Uma bomba de fabricação caseira [15]]** explodiu em frente a o prédio de o Ministério_Público_Estadual e lojas vizinhas também foram atingidas por estilhaços . **[A rua [22]]** está interditada para a perícia e , a os poucos , os comerciantes são autorizados a entrar em seus estabelecimentos . A Secretaria_da_Fazenda também foi atingida por **[uma bomba [15]]** . Duas bases de **[a Guarda_Civil_Metropolitana [29]]** (**[GCM [29]]**) , sendo uma em o Capão_Redondo , Zona_Sul de **[São_Paulo [1]]** , foram alvo de os criminosos . Mais de dez agências bancárias , um posto de gasolina e um supermercado foram atacados . Calcula -se em 15 o número de ônibus incendiados , sendo dez em a região de o ABC e quatro em **[a capital [9]]** . Mesmo_assim , o sistema de transporte coletivo de a cidade está normal em esta manhã . **[A rua [22]]** onde fica **[o Departamento_de_Investigações_Sobre_o_Crime_Organizado ([47]] [Deic [47]]**) foi bloqueada e a passagem de veículos está proibida . . .

Cadeias Extraídas:

1.) [São Paulo], [São Paulo];
2.) [os ataques], [As ações];
9.) [a Capital], [PCC], [a capital];
15.) [Uma bomba de fabricação caseira], [uma bomba];
22.) [A rua], [A rua];
29.) [a Guarda Civil Metropolitana], [GCM];
47.) [o Departamento de Investigações Sobre o Crime Organizado], [Deic].

Análise:

Na cadeia podemos notar o agrupamento dos sintagmas [São_Paulo]. Contudo, o primeiro remete para o estado de São Paulo, já o segundo, para a cidade de São Paulo.

Esse tipo de ambiguidade é bem difícil de ser resolvida em nível computacional, pois muitas vezes requer o conhecimento de mundo, como, por exemplo, saber que existe o estado de São Paulo e a Cidade de São Paulo. Na cadeia 9 podemos notar que houve um agrupamento incorreto por meio da regra de Aposto Explicativo entre os sintagmas [a Capital] e [PCC]. Isso porque a regra, sempre que encontra menções entre parenteses, agrupa a menção ligeiramente adjacente. Como o parser separou o sintagma [Primeiro Comando da Capital], a menção foi agrupada incorretamente. Basicamente, o correto seria o modelo gerar duas cadeias: a primeira com ([Primeiro Comando da Capital], [PCC]) e ([a Capital], [São Paulo], [a Capital]). Na cadeia 15, embora as menções satisfaçam a regra de Casamento pelo Núcleo, podemos ver que as menções referem-se a entidades distintas; a primeira refere-se à bomba que explodiu em frente ao prédio do Ministério Público; e a segunda, a uma bomba que atingiu a Secretaria da Fazenda.

6.8 Considerações do Capítulo

Neste capítulo mostramos os experimentos realizados, envolvendo dois principais corpora de correferência para o Português. Esses tiveram como objetivo avaliar o desempenho e a relevância de nossas regras propostas, de forma individual e cumulativa. Avaliamos também nosso algoritmo de agrupamento de menções, envolvendo diferentes critérios de agrupamento e inclusão de conhecimento semântico, mostrando que nosso método de agrupamento supera o “Baseline”, atual método utilizado por trabalhos que lidam com esse tipo de abordagem.

Realizamos também uma análise, mostrando os tipos mais comuns de erros, ocorridos durante a obtenção automática das cadeias de correferência. Salientamos aqui duas principais limitações existentes: o processo de detecção de menções e o tamanho da base semântica utilizada em nosso modelo. Podemos ver também que o uso do conhecimento semântico traz uma margem muito maior à ambiguidade. Contudo, este é um problema existente em níveis lexicais também, como visto no texto 5: (São Paulo - Estado) e (São Paulo - Cidade).

7. CONSIDERAÇÕES FINAIS

Conforme apresentado ao longo deste trabalho, diferentes abordagens computacionais para a resolução de correferência têm sido estudadas e aplicadas. Contudo, estas, em sua maioria, focam-se em métodos de aprendizado de máquina aplicados a outros idiomas. Além disso, poucas dessas abordagens propõem o uso da semântica para enriquecer seus resultados. Embora abordagens baseadas em regras sejam promissoras para idiomas como o Português, em nosso trabalho existiu um grande esforço para conceber o conjunto de regras proposto. Isto é, mesmo algumas regras sendo semelhantes às de outros trabalhos, como as regras de Lee et al. [39], modelo o qual nos baseamos, existem muitos detalhes a serem observados, tais como: comportamento do idioma em questão e efetividade de cada regra: diferente de abordagens baseadas em aprendizado de máquina, em abordagens baseadas em regras não existem algoritmos responsáveis por "filtrar" regras/*features* irrelevantes. Essa etapa de detecção envolve uma série de testes, estudos e adaptações da parte de quem as projeta. É importante destacar que no trabalho de Lee et al. as menções são agrupadas caso haja uma ou mais regras satisfeitas. Nosso processo de agrupamento de menções faz uso da representação do discurso de forma a decidir ponderadamente se a menção atual m_j é uma referência do antecedente m_i ou se é nova no discurso. Outro diferencial, além do idioma de escopo e critério de agrupamento, é que Lee et al. não faz uso de recursos semânticos. Nossa abordagem proposta introduz o uso da semântica por meio do Onto.PT.

Referente às nossas regras básicas, podemos observar que todas mostraram-se significativas e, à medida que foram adicionadas, nosso modelo perdeu um pouco de precisão, mas ganhou em abrangência, aumentando, na maioria dos casos, sua medida-f. No caso de nossas regras semânticas obtivemos um salto em abrangência, contudo, houve uma perda significativa em precisão. Isso nos leva a refletir sobre o motivo de poucos trabalhos que tratam correferência proporem o uso de semântica para auxiliar na tarefa. Desses trabalhos, menos ainda tratam de questões contextuais. Neste trabalho, faz-se uma tentativa de lançar um olhar para o contexto em que as cadeias de correferência se desenvolvem; é relevante notar a relação de co-dependência entre os sintagmas explorados nos textos analisados. Quando analisamos nossos resultados pelo F-Score da CoNLL, à primeira vista estes demonstram que nosso modelo é 1,7% superior quando não usamos semântica. Contudo, se analisarmos os resultados de forma mais detalhada, é possível notar que, mesmo perdendo pontos em precisão, obtivemos ganhos em abrangência: 2,4% para a métrica MUC, 3,1% para a métrica B^3 e 1,8% para $Ceaf_e$. De forma a minimizar a perda de precisão obtida pela introdução de semântica, aplicamos nosso novo método de agrupamento, o qual mostrou-se eficaz. Por meio deste, foi possível obtermos um aumento de 10,7% para a métrica MUC, 16% para B^3 e 0,1% para $Ceaf_e$, melhorando o F-Score em

5,4%. Se analisarmos os resultados dos principais trabalhos relacionados, veremos que a resolução de correferência é uma tarefa desafiadora e existe uma grande dificuldade para obtenção de bons resultados. Dada essa dificuldade, alguns trabalhos optam por escopos reduzidos, como o de Garcia et al. [31], o qual resolve correferência apenas para a categoria 'Pessoa'. Em nossa proposta não tratamos pronomes pessoais: nosso modelo resolve correferência nominal do tipo identidade: nomes comuns e próprios, independente de sua categoria semântica ou domínio. Em outras palavras, nossa abordagem engloba regras genéricas que servem para qualquer tipo de texto escrito em Português. Entretanto, para a obtenção de resultados satisfatórios, estes precisam estar bem estruturados e livres de erros gramaticais, o que pode representar um obstáculo para a tarefa que propomos.

É importante destacarmos que o Português tem sido uma língua carente de recursos, dificultando ainda mais a obtenção de bons resultados. Contudo, felizmente, por meio do esforço de muitos pesquisadores, este cenário vem mudando. Um exemplo disso é a concepção do corpus Corref-PT, um corpus construído por meio de um esforço colaborativo durante o IBEREVAL-2017. Acreditamos também que a metodologia empregada durante esta tarefa colaborativa servirá para conceber novos corpora anotados com correferência, uma vez que a tarefa de anotação se torna menos onerosa quando podemos contar com o auxílio de ferramentas, como o CORP e CorrefVisual. Acreditamos também que os futuros corpora gerados, assim como o Corref-PT, poderão auxiliar em diversas tarefas de PLN e, principalmente, na concepção de novos modelos de correferência para o Português.

Nesta tese foram apresentados estudos do atual estado da arte, ao que remete à tarefa de resolução automática de correferência (Capítulo 3). Esse estudo resultou na proposta de um processo para a resolução de correferência em língua Portuguesa, apresentado no Capítulo 5. Além de nosso processo, propomos também um novo método para o agrupamento de menções, o qual melhorou significativamente a precisão de nosso modelo, como pôde ser visto no Capítulo 6. Ainda no Capítulo 6, em Análise de Erros, foi possível vermos os principais erros cometidos por nosso modelo. Essa análise nos permitiu projetar novas direções de pesquisa, objetivando minimizar essas deficiências. Como pudemos notar, o erro mais comum ocorrido em nosso modelo é devido a ambiguidade de termos, não apenas em nível semântico, mas também em nível léxico-semântico, como em São Paulo (Estado) e São Paulo (Capital). Muitos outros erros são decorrentes do processo de identificação de menções, os quais infelizmente encaramos como uma limitação em nosso modelo, dado que este nível de processamento é realizado pelo parser. Outra limitação existente remete para o tamanho de nossa base semântica utilizada [49], contudo, nossas regras são genéricas. Logo, nada impede a adição de novas bases de conhecimento ao nosso modelo, algo que projetamos como direções futuras para este trabalho.

7.1 Contribuições

A proposta de um processo para resolução de correferência em textos da língua Portuguesa é uma das principais contribuições deste trabalho. Conforme apresentado no Capítulo 3, em geral os sistemas de resolução de correferência são baseados em aprendizado de máquina para a língua Inglesa, pelo fato de ainda existir certa carência de recursos anotados para o Português, ou seja, ausência de recursos ricos em amostras de correferência para conceber modelos para a resolução de correferência que sejam eficientes; ao propormos uma abordagem baseada em regras foi possível contornar esta limitação. Além de ser para o Português, nosso modelo possui outro diferencial que é a utilização de conhecimento semântico, algo que mesmo em trabalhos concebidos para a língua Inglesa é pouco abordado.

Dentre as contribuições desta tese destacam-se também um novo método para a geração de cadeias de correferência; a concepção de uma nova versão do corpus Summ-it, o Summ-it++ [2]; uma ferramenta para anotação automática de correferência, o CORP [26]; uma ferramenta para correção manual das cadeias de correferência, produzidas automaticamente pelo CORP, CorrefVisual [67]; um novo corpus anotado, contendo aproximadamente 4000 cadeias de correferência, Corref-PT [20]. Este possui 8 vezes mais cadeias, quando comparado com os atuais corpora de referência para a tarefa, Summ-it e Summ-it++.

7.2 Trabalhos Futuros

Como trabalhos futuros pretendemos explorar novas bases semânticas, como ConceptNet [65] e BabelNet [47]; e métodos automatizados para concebê-las [74, 37], de forma a usar/criar novas bases e analisar os ganhos que cada uma delas pode prover em nossa abordagem. Objetivamos também, explorar a viabilidade de unificá-las, de forma a enriquecer o conhecimento semântico utilizado. Pretendemos também realizar estudos em técnicas de *Word Sense Disambiguation*, de forma a reduzir o agrupamento incorreto por meio de termos ambíguos. Além disso, visto que nosso método de agrupamento proposto não funciona apenas para o Português, objetivamos avaliá-lo em diferentes modelos e idiomas.

Outra linha de trabalho que pretendemos seguir remete para a resolução de correferência pronominal. Para isso, pretendemos explorar trabalhos como o de Garcia et al. [31], de forma a introduzir correferência pronominal em nosso modelo. Além disso, pretendemos realizar uma nova análise comparativa entre nosso modelo e o de Garcia, que envolva outros corpora, como Summ-it++ e Corref-PT. Pretendemos também explorar técnicas ba-

seadas em análise de sentimento, que podem prover ganhos significativos na resolução de pronomes, como podemos ver no trabalho de Rahman et al. [56].

Devido ao fato de nosso modelo proposto ser o primeiro modelo baseado em regras e conhecimento semântico para o Português, acreditamos que exista uma grande margem à melhorias. Não apenas na parte semântica, mas na parte de desenvolvimento de nossas regras, como um melhor tratamento para a discordância de número entre menções (singular/plural). Além disso, acreditamos que seja possível melhorar os resultados de nosso modelo por meio de uma busca gulosa, envolvendo diferentes combinações de regras. Dito isso, pretendemos seguir explorando a literatura e os padrões linguísticos, de forma a projetar novas regras, que possam trazer melhorias de abrangência e precisão ao nosso modelo. Pretendemos também estudar a viabilidade e os possíveis benefícios de empregar Word Embeddings em nosso modelo.

7.3 Principais Publicações no Contexto Desta Tese

Nesta Seção descrevemos sucintamente nossos principais trabalhos publicados durante esta tese. Estes estão em ordem cronológica:

- [28] Fonseca, E. B.; Vieira, R.; Vanin, A. A. "**Coreference resolution in portuguese: Detecting person, location and organization**". In: Journal of the Brazilian Computational Intelligence Society, 2014, pp. 86–97.

  Neste *journal* exploramos o uso de aprendizado de máquina para tratar a tarefa de resolução de correferência para o Português envolvendo categorias de entidades específicas.

- [24] Fonseca, E. B.; Vieira, R.; Vanin, A. "**Dealing with imbalanced datasets for coreference resolution**". In: Proceedings of The Twenty-Eighth International Flairs Conference, 2015.

  Este artigo aborda o problema de desbalanceamento de classes, residente na tarefa de Resolução de Correferência, envolvendo aprendizado de máquina. Diferentes níveis de balanceamento foram testados com o objetivo de encontrar a melhor proporção para treino de nosso modelo.

- [27] Fonseca, E. B.; Vieira, R.; Vanin, A. "**Improving coreference resolution with semantic knowledge**". In: Proceedings of the 12th International Conference on the Computational Processing of Portuguese, 2016, pp. 213–224.

  Neste artigo utilizamos os níveis de balanceamento de classes propostos em nosso trabalho anterior [24] para conceber um novo modelo. Este foi nosso primeiro modelo baseado em conhecimento semântico. Como resultado, constatamos que bases como

o Onto.PT[33] podem prover ganhos significativos para a tarefa, no âmbito da língua Portuguesa.

- [25] Fonseca, E. B.; Vieira, R.; Vanin, A. "**Adapting an entity centric model for portuguese coreference resolution**". In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation, 2016.

  Este artigo remete para nossa primeira tentativa de gerar um modelo para a resolução de correferência em Português, baseado em regras linguísticas. De forma a avaliar nosso modelo, utilizamos o corpus do Harem[29] e calculamos valores de precisão, abrangência e medida-f para cada categoria de entidade nomeada.

- [26] Fonseca, E. B.; Vieira, R.; Vanin, A. "**Corp: Coreference resolution for portuguese**". In: 12th International Conference on the Computational Processing of Portuguese, Demo Session, 2016.

  Nesta demo apresentamos a primeira versão do CORP, nossa API para a resolução de correferência em Português.

- [2] Antonitsch, A.; Figueira, A.; Amaral, D.; Fonseca, E.; Vieira, R.; Collovini, S. "**Summ-it++: an enriched version of the summ-it corpus**". In: Proceedings of 10th edition of the Language Resources and Evaluation Conference, 2016.

  Neste trabalho apresentamos uma nova versão do corpus Summ-it. Em sua nova versão, acrescentamos duas novas camadas semânticas: a de relação entre entidades nomeadas; e a de categorias de entidades nomeadas. Adicionalmente portamos-o para o formato SemEval. Um formato muito utilizado pela maioria dos corpora.

- [23] Fonseca, E. B.; Sesti, V.; Antonitsch, A.; Vanin, A. A.; Vieira, R. "**Corp - uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências**", Linguamatica, vol. 9–1, 2017, pp. 3–18.

  Neste *journal* damos detalhes de nossa abordagem baseada em regras e conhecimento semântico proposta para resolver correferência em textos do Português. Neste, avaliamos nossas regras propostas de forma individual e cumulativa, calculando valores de precisão, abrangência e medida-f, usando o corpus Summ-it++[2] como referência.

- [20] Fonseca, E.; Sesti, V.; Collovini, S.; Vieira, R.; Leal, A. L.; Quaresma, P. "**Collective elaboration of a coreference annotated corpus for portuguese texts**". In: Proceedings of II workshop on Evaluation of Human Language Technologies for Iberian Languages, 2017, pp. 68–82.

  Neste trabalho propomos uma tarefa colaborativa para a concepção de um novo corpus anotado com correferência, para o Português. O novo corpus concebido, Corref-PT, foi anotado automaticamente por nossa API [26] e corrigido manualmente por 21

participantes, divididos em 7 equipes. As equipes participantes foram constituídas por falantes nativos do português, variando entre professores e alunos da área de linguística computacional.

7.4 Publicações em Áreas Suplementares Desta Tese

- [1] Amaral, D. O.; Fonseca, E. B.; Lopes, L.; Vieira, R. "**Comparative analysis of portuguese named entities recognition tools**". In: Proceedings of Language Resources and Evaluation Conference, 2014, pp. 2554–2558

  Neste trabalho realizamos uma análise comparativa entre os principais recursos reconhecedores de entidades nomeadas para o Português. Para avaliá-los utilizamos o corpus do HAREM.

- [22] Fonseca, E. B.; Chiele, G.; Vieira, R.; Vanin, A. A. "**Reconhecimento de entidades nomeadas para o português usando o opennlp**". In: XI Encontro Nacional de Inteligência Artificial e Computacional, 2015.

  Neste trabalho propomos o uso do OpenNLP e do corpus Amazônia[1] para treinar um modelo reconhecedor de entidades nomeadas para o Português.

- [69] Vieira, R.; do Amaral, D.; Collovini, S.; Fonseca, E.; Freitas, A.; Freitas, L.; Granada, R.; Hilgert, L.; Lopes, L.; Schmidt, D.; et al.. "**Language resources for information extraction and semantic computing-nlp at pucrs**". In: Workshop of International Conference on Computational Processing of the Portuguese Language, 2016.

  Neste artigo apresentamos os principais recursos desenvolvidos pelo grupo de pesquisa em Processamento da Linguagem Natural da PUCRS nos últimos anos.

[1] Disponível em: http://www.linguateca.pt/floresta/ficheiros/gz/amazonia.ad.gz

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Amaral, D. O.; Fonseca, E. B.; Lopes, L.; Vieira, R. “Comparative analysis of portuguese named entities recognition tools”. In: Proceedings of Language Resources and Evaluation Conference, 2014, pp. 2554–2558.

[2] Antonitsch, A.; Figueira, A.; Amaral, D.; Fonseca, E.; Vieira, R.; Collovini, S. “Summ-it++: an enriched version of the summ-it corpus”. In: Proceedings of 10th edition of the Language Resources and Evaluation Conference, 2016.

[3] Bagga, A.; Baldwin, B. “Algorithms for scoring coreference chains”. In: Proceedings of the first 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: Proceedings of the 17th International Conference on Computational Linguistics, 1998, pp. 86–90.

[5] Basso, R. M. “A semântica das relações anafóricas entre eventos”, Tese de Doutorado, Universidade Estadual de Campinas, SP, 2009.

[6] Bechara, E. “Lições de português pela análise sintática”. Fundo de Cultura, 2014, 19 ed..

[7] Bick, E. “The parsing system “palavras": Automatic grammatical analysis of portuguese in a constraint grammar framework.”, Tese de Doutorado, Aarhus University Press, Denmark, 2000.

[8] Bick, E. “A dependency-based approach to anaphora annotation”. In: Proceedings of th 9th International Conference on Computational Processing of the Portuguese Language, 2010.

[9] Cadore, L. A.; Ledur, P. F. “Análise Sintática Aplicada: fundamentos de concordância, regência, crase, colocação, pontuação e significado”. AGE, 2013, 4 ed..

[10] Cardoso, N. “Rembrandt - a named-entity recognition framework”. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012, pp. 1240–1243.

[11] Chang, K.-W.; Samdani, R.; Rozovskaya, A.; Sammons, M.; Roth, D. “Illinois-coref: The ui system in the conll-2012 shared task”. In: Joint Conference on EMNLP and CoNLL-Shared Task, 2012, pp. 113–117.

[12] 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: Proceedings of V Workshop em Tecnologia da Informação e da Linguagem Humana, 2007, pp. 1605–1614.

[13] Collovini, S.; Pugens, L.; Vanin, A. A.; Vieira, R. “Extraction of relation descriptors for portuguese using conditional random fields”. In: Proceedings of the 14th Ibero-American Conference on Advances in Artificial Intelligence, 2014, pp. 108–119.

[14] Coreixas, T. “Resolução de correferência e categorias de entidades nomeadas”, Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, 2010.

[15] do Amaral, D. O. F. “O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa”, Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, 2013.

[16] do Nascimento, M. F. B.; Mendes, A.; Pereira, L. “Providing on-line access to portuguese language resources: Corpora and lexicons.” In: Proceedings of the International Conference on Language Resources and Evaluation , Portugal, 2004.

[17] Durrett, G.; Klein, D. “Easy victories and uphill battles in coreference resolution.” In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1971–1982.

[18] Fernandes, E. R.; dos Santos, C. N.; Milidiú, R. L. “Latent trees for coreference resolution”, *Computational Linguistics*, 2014.

[19] Ferradeira, J. E. d. S. “Resolução de anáfora pronominal”, Dissertação de Mestrado, Universidade Nova de Lisboa, 1993.

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

[21] Fonseca, E. B. “Resolução de correferências em língua portuguesa: pessoa, local e organização”, Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, 2014.

[22] Fonseca, E. B.; Chiele, G.; Vieira, R.; Vanin, A. A. “Reconhecimento de entidades nomeadas para o português usando o opennlp”. In: Proceedings of XI Encontro Nacional de Inteligência Artificial e Computacional, 2015.

[23] Fonseca, E. B.; Sesti, V.; Antonitsch, A.; Vanin, A. A.; Vieira, R. “Corp - uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências”, *Linguamatica*, vol. 9–1, 2017, pp. 3–18.

[24] Fonseca, E. B.; Vieira, R.; Vanin, A. “Dealing with imbalanced datasets for coreference resolution”. In: Proceedings of The Twenty-Eighth International Flairs Conference, 2015.

[25] Fonseca, E. B.; Vieira, R.; Vanin, A. “Adapting an entity centric model for portuguese coreference resolution”. In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation, 2016.

[26] Fonseca, E. B.; Vieira, R.; Vanin, A. “Corp: Coreference resolution for portuguese”. In: Proceedings of 12th International Conference on the Computational Processing of Portuguese, Demo Session, 2016.

[27] Fonseca, E. B.; Vieira, R.; Vanin, A. “Improving coreference resolution with semantic knowledge”. In: Proceedings of the 12th International Conference on the Computational Processing of Portuguese, 2016, pp. 213–224.

[28] Fonseca, E. B.; Vieira, R.; Vanin, A. A. “Coreference resolution in portuguese: Detecting person, location and organization”, *Journal of the Brazilian Computational Intelligence Society*, vol. 12–2, 2014, pp. 86–97.

[29] 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: Proceedings of the International Conference on Language Resources and Evaluation, 2010.

[30] Freitas, C.; Santos, D.; Mota, C.; Oliveira, H. G.; Carvalho, P. “Relation detection between named entities: report of a shared task”. In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions, 2009, pp. 129–137.

[31] Garcia, M.; Gamallo, P. “An entity-centric coreference resolution system for person entities with rich linguistic information”. In: Proceedings of 25th International Conference on Computational Linguistics, 2014, pp. 741–752.

[32] Garcia, M.; Gamallo, P. “Multilingual corpora with coreferential annotation of person entities”. In: Proceedings of the 9th edition of the Language Resources and Evaluation Conference, 2014, pp. 3229–3233.

[33] Gonçalo Oliveira, H. “Onto. pt: Towards the automatic construction of a lexical ontology for portuguese”, Tese de Doutorado, Ph. D. thesis, University of Coimbra, 2012.

[34] Heim, I. “File Change Semantics and the Familiarity Theory of Definiteness”. Wiley-Blackwell, 2008, cap. 9, pp. 223–248.

[35] Hou, Y.; Markert, K.; Strube, M. “A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents”. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 2082–2093.

[36] Jurafsky, D.; Martin, J. H. “Speech and Language Processing”. Pearson Education India, 2014, 2 ed..

[37] Kamel, M.; Trojahn, C.; Ghamnia, A.; Aussenac-Gilles, N.; Fabre, C. “Extracting hypernym relations from wikipedia disambiguation pages: comparing symbolic and machine learning approaches”. In: Proceedings of IWCS 2017-12th International Conference on Computational Semantics, 2017.

[38] Koch, I. G. V.; Travaglia, L. “Texto e coerência”. Cortez, 2012, 13 ed..

[39] 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, 2013, pp. 885–916.

[40] 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: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, 2011.

[41] Luo, X. “On coreference resolution performance metrics”. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2005.

[42] Maia, L. C. G. “Uso de sintagmas nominais na classificação automática de documentos eletrônicos”, Tese de Doutorado, Universidade Federal de Minas Gerais, 2008.

[43] Martschat, S.; Strube, M. “Latent structures for coreference resolution”, *Transactions of the Association for Computational Linguistics*, vol. 3, 2015, pp. 405–418.

[44] Maziero, E. G.; del Rosario Castro Jorge, M. L.; Pardo, T. A. S. “Identifying multidocument relations”. In: Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science, 2010, pp. 60–69.

[45] Maziero, E. G.; Pardo, T. A.; 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: Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, 2008, pp. 390–392.

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

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

[48] Ng, V.; Cardie, C. “Improving machine learning approaches to coreference resolution”. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 104–111.

[49] 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, 2014, pp. 373–393.

[50] Poesio, M.; Stuckardt, R.; Versley, Y. “Anaphora Resolution: Algorithms, Resources, and Applications”. Springer, 2016, 1 ed..

[51] Pradhan, S.; Luo, X.; Recasens, M.; Hovy, E. H.; Ng, V.; Strube, M. “Scoring coreference partitions of predicted mentions: A reference implementation”. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 30–35.

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

[53] Pradhan, S.; Ramshaw, L.; Marcus, M.; Palmer, M.; Weischedel, R.; Xue, N. “Conll-2011 shared task: Modeling unrestricted coreference in ontonotes”. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, 2011, pp. 1–27.

[54] Rahman, A.; Ng, V. “Coreference resolution with world knowledge”. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 814–824.

[55] Rahman, A.; Ng, V. “Narrowing the modeling gap: a cluster-ranking approach to coreference resolution”, *Journal of Artificial Intelligence Research*, 2011, pp. 469–521.

[56] Rahman, A.; Ng, V. “Resolving complex cases of definite pronouns: the winograd schema challenge”. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 777–789.

[57] Recasens, M.; Hovy, E. H. “BLANC: implementing the rand index for coreference evaluation”, *Natural Language Engineering*, vol. 17–4, 2011, pp. 485–510.

[58] 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: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 1–8.

[59] Rocha, M. “A corpus-based study of anaphora in English and Portuguese, Corpus-based and Computational Approaches to Discourse Anaphora”. John Benjamins Publishing Company, 2000, pp. 81–94.

[60] Sarmento, L.; Pinto, A. S.; Cabral, L. “Repentino – a wide-scope gazetteer for entity recognition in portuguese”. In: Proceedings of International Workshop on Computational Processing of the Portuguese Language, 2006, pp. 31–40.

[61] Sesti, V.; Fonseca, E.; Vieira, R. “Correfvisual: Ferramenta para a edição de correferências”. In: Proceedings of V Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana, 2017.

[62] Silva, J. F. d. “Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado”, Dissertação de Mestrado, Universidade de São Paulo, 2011.

[63] Silva, W. D. C. “Aprimorando o corretor gramatical cogroo”, Dissertação de Mestrado, Universidade de São Paulo, 2013.

[64] Soon, W. M.; Ng, H. T.; Lim, C. Y. “A machine learning approach to coreference resolution of noun phrases”, *Computational Linguistics*, vol. 27–4, 2001, pp. 521–544.

[65] Speer, R.; Havasi, C. “Representing general relational knowledge in conceptnet 5”. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012, pp. 3679–3686.

[66] Suchanek, F. M.; Kasneci, G.; Weikum, G. “Yago: a core of semantic knowledge”. In: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 697–706.

[67] Tubino, M. d. O.; Silva, M. M. S. “Visualização, manipulação e refinamento de correferência em língua portuguesa”, Trabalho de conclusão de curso, Pontifícia Universidade Católica do Rio Grande do Sul, 2015.

[68] Versley, Y.; Ponzetto, S. P.; Poesio, M.; Eidelman, V.; Jern, A.; Smith, J.; Yang, X.; Moschitti, A. “Bart: A modular toolkit for coreference resolution”. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session, 2008, pp. 9–12.

[69] Vieira, R.; do Amaral, D.; Collovini, S.; Fonseca, E.; Freitas, A.; Freitas, L.; Granada, R.; Hilgert, L.; Lopes, L.; Schmidt, D.; et al.. “Language resources for information extraction

and semantic computing-nlp at pucrs". In: Workshop of International Conference on Computational Processing of the Portuguese Language, 2016.

[70] Vieira, R.; Gonçalves, P. N.; Souza, J. G. C. d. "Processamento computacional de anáfora e correferência", *Revista de Estudos da Linguagem*, vol. 16–1, 2012.

[71] Vieira, R.; Salmon-Alt, S.; Gasperin, C.; Schang, E.; Othero, G. "Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus", *Anaphora Processing: linguistic, cognitive and computational modeling*, 2005, pp. 385–403.

[72] Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; Hirschman, L. "A model-theoretic coreference scoring scheme". In: Proceedings of the 6th Conference on Message understanding, 1995, pp. 45–52.

[73] Yang, X.; Su, J.; Lang, J.; Tan, C. L.; Liu, T.; Li, S. "An entity-mention model for coreference resolution with inductive logic programming." In: Proceeding of Association for Computational Linguistics, 2008, pp. 843–851.

[74] Zesch, T.; Müller, C.; Gurevych, I. "Extracting lexical semantic knowledge from wikipedia and wiktionary." In: Proceedings of Conference on Language Resources and Evaluation, 2008, pp. 1646–1652.

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