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

RESOLUÇÃO DE CORREFERÊNCIAS EM LÍNGUA PORTUGUESA: PESSOA, LOCAL E ORGANIZAÇÃO

EVANDRO BRASIL FONSECA

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

Orientadora: Prof. Renata Vieira

Porto Alegre
2014

Dados Internacionais de Catalogação na Publicação (CIP)

F676r Fonseca, Evandro Brasil
Resolução de correferências em língua portuguesa : pessoa, local e organização / Evandro Brasil Fonseca. – Porto Alegre, 2014.
78 p.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Profª. Drª. Renata Vieira.

1. Informática. 2. Processamento da Linguagem Natural. 3. Linguística Computacional. 4. Aprendizagem de Máquina. I. Vieira, Renata. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Resolução de Correferência em Língua Portuguesa: Pessoal, Local e Organização" apresentada por Evandro Brasil Fonseca como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, aprovada em 21/03/2014 pela Comissão Examinadora:

Profa. Dra. Renata Vieira – Orientadora — PPGCC/PUCRS

Prof. Dr. Rafael Heitor Bordini — PPGCC/PUCRS

Profa. Dra. Valéria Delisandra Feltrim — UEM

Homologada em 22/05/2014, conforme Ata No. 008 pela Comissão Coordenadora.

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

DEDICATÓRIA

Dedico este trabalho:

A minha esposa, Natália, amor da minha vida e fonte de inspiração.

Aos meus pais, Elemar e Ilda, por me ensinarem a lutar pelos meus sonhos.

“Escolha um trabalho que você ame e não terás que trabalhar um único dia em sua vida.”

Sigmund Freud.

AGRADECIMENTOS

Primeiramente agradeço a minha orientadora, professora Renata Vieira, pelos conselhos, ensinamentos e atenção recebida.

Agradeço a meus pais, Elemar Fonseca e Ilda Brasil, pela educação e apoio recebido ao longo de meu caminho.

A Natália, minha noiva, por entender e apoiar minha opção pela vida acadêmica.

A Caroline Fonseca e Cesar Schütz, por me acolherem e dar apoio para que eu pudesse iniciar o mestrado.

A minha irmã, Carolina Fonseca, pelo apoio e incentivo.

Meus colegas do Grupo de Pesquisa em PLN da PUCRS, cujo apoio foi essencial para a conclusão das disciplinas do mestrado. Em especial, a Aline Vanin, pelas infindáveis correções e pelo acompanhamento durante a produção deste trabalho.

A Faculdade de Informática da PUCRS, excelência em ensino e pesquisa, a quem devo minha formação acadêmica.

Por fim, meu agradecimento ao CNPq pelo auxílio financeiro que possibilitou a realização deste trabalho.

RESOLUÇÃO DE CORREFERÊNCIAS EM LÍNGUA PORTUGUESA: PESSOA, LOCAL E ORGANIZAÇÃO

RESUMO

Resolução de correferências é um processo que consiste em identificar as diversas formas que uma mesma entidade nomeada pode assumir em um determinado texto. Em outras palavras, esse processo consiste em identificar determinados termos e expressões que remetem a uma mesma entidade. A resolução automática de correferência textual está inserida num contexto muito importante na área de Processamento da Linguagem Natural (PLN), pois vários sistemas necessitam dessa tarefa, como, por exemplo, a extração de relação entre entidades nomeadas. O nível de processamento linguístico depende do conhecimento de mundo, e isso ainda é um desafio para a área. A necessidade crescente por ferramentas de PLN e a escassez de recursos livres para a língua portuguesa motivaram trabalhar com essa língua nesta dissertação de mestrado. O presente trabalho teve por objetivo desenvolver uma ferramenta *open source* para a resolução de correferências em língua portuguesa, tendo como foco as categorias de entidades nomeadas Pessoa, Local e Organização. Optou-se por essas três categorias por essas serem as mais relevantes para a maioria das tarefas de PLN, pelo fato de tratarem entidades mais específicas e de interesse comum. Além disso, são as categorias mais exploradas em trabalhos voltados à resolução de correferência. Escolheu-se trabalhar apenas com recursos *open source* pelo fato de a maioria dos trabalhos para a língua portuguesa utilizar recursos proprietários. Isso acaba limitando a disponibilidade da ferramenta e, consequentemente, o seu uso. A metodologia utilizada é baseada em aprendizado de máquina supervisionado. Para tal, o uso de *features* que auxiliem na correta classificação de pares de sintagmas como correferentes ou não-correferentes é fundamental para, posteriormente, agrupá-los, gerando cadeias de correferência. Embora ainda existam muitos desafios a serem resolvidos, os resultados do sistema descrito nesta dissertação são animadores, quando comparados indiretamente, por meio de uma mesma métrica, ao atual estado da arte.

Palavras-Chave: Resolução de Correferência, Processamento da Linguagem Natural, Entidades Nomeadas, Aprendizado de Máquina.

COREFERENCE RESOLUTION IN PORTUGUESE: PERSON, LOCATION AND ORGANIZATION

ABSTRACT

Coreference resolution is a process that consists in identifying the several forms that a specific named entity may assume on certain text. In other words, this process consists in identifying certain terms and expressions that refer certain named entity. The automatic textual coreference resolution is in a very important context in the Natural Language Processing (NLP) area, because several systems need its tasks, such as the relation extraction between named entities. The linguistic processing level depends on the knowledge about the world, and this is a challenge for this area, mainly for the Portuguese language. The growing necessity of NLP tools and the lack of open source resources for Portuguese have inspired the research on this language, and they became the focus of this dissertation. The present work aims at building an open source tool for the Coreference resolution in Portuguese, focusing on the Person, Location and Organization domains. These three categories were chosen given their relevance for most NLP tasks, because they represent more specifically entities of common interest. Furthermore, they are the most explored categories in the related works. The choice for working only with open source resources is because most of related works for Portuguese uses private software, which limits his availability and his usability. The methodology is based on supervised machine learning. For this task, the use of features that help on the correct classification of noun phrase pairs as coreferent or non-coreferent are essential for grouping them later, thus building coreference chains. Although there are still many challenges to be overcome, the results of the system described in this dissertation are encouraging when compared indirectly, by using the same metric, to the current state of the art.

Key-words: *coreference resolution; Natural Language Processing; named entities; machine learning.*

LISTA DE FIGURAS

Figura 1: Saída CoGrOO 35
Figura 2: Anotações Repentino 36
Figura 3: Saída NERP-CRF 37
Figura 4: Tokens Corpus Summ-it 39
Figura 5: Sintagmas Nominais 39
Figura 6: Informações sintático-semânticas 40
Figura 7: Distribuição das categorias do segundo HAREM [FRE10] 41
Figura 8: Dataset de treinamento. 44
Figura 9: Arquitetura supervisionada proposta nesta dissertação. 45
Figura 10: Aprendizado não supervisionado, retirado de [SIL11] 46
Figura 11: Árvore de decisão gerada pelo algoritmo SimpleCart 52
Figura 12: Arquitetura CORP 54
Figura 13: Saída CORP 55
Figura 14: Anotação de correferência no corpus do HAREM [HAR08] 58
Figura 15: Resultados dos quatro experimentos realizados no CORP. 64
Figura 16: Análise de erros exemplo 1. 65
Figura 17: Possíveis caminhos da árvore para o exemplo existente na FIGURA 16 66
Figura 18: Análise de erros exemplo 2. 67
Figura 19: Caminho correto para a classificação do par de sintagma da FIGURA 17. 68
Figura 20: Cadeia gerada incorretamente (arquivo 2ght33.txt [HAR08]) 69
Figura 21: Cadeia gerada utilizando os filtros do quarto experimento 70

LISTA DE TABELAS

Tabela 1: *Features* utilizadas por [SOO01] 32
Tabela 2: Descrição das *features* 47
Tabela 3: Precisão dos classificadores [FON13b]. 50
Tabela 4: Multilayer Perceptron. 51
Tabela 5: LBR. 51
Tabela 6: Random Forest. 51
Tabela 7: SimpleCart 51
Tabela 8: Vetor de características CORP 55
Tabela 9: Tipos de pares considerados no experimento 2. 62
Tabela 10: Tipos de pares considerados no experimento 3. 63
Tabela 11: Tipos de pares considerados no experimento 4. 63
Tabela 12: Avaliação CORP 64
Tabela 13: Resultados não comparativos, baseados na métrica MUC[VIL95]. 65

LISTA DE SIGLAS

AM- Aprendizado de Máquina

ARFF - *Attribute-Relation File Format*

CONLL - *Conference on Computational Natural Language Learning*

EN - Entidade nomeada

HAREM - Avaliação de Sistemas de Reconhecimento de Entidades Nomeadas

MUC – Message Understand Conference

PLN - Processamento da Linguagem Natural

SN - Sintagma nominal

XML - *Extensible Markup Language*

WSD - Word Sense Disambiguation

SUMÁRIO

1. INTRODUÇÃO 21
1.1. MOTIVAÇÃO 21
1.2. OBJETIVOS 22
1.3. METODOLOGIA 22
1.4. ESCOPO E RESULTADOS OBTIDOS 23
2. FUNDAMENTAÇÃO TEÓRICA 25
2.1. REFERENTES 25
2.1.1 Entidades nomeadas 25
2.1.2. Sintagmas 26
2.1.3 Tipos de referentes: 26
2.1.4 Correferência e anáfora: 27
3. TRABALHOS RELACIONADOS 31
4. RECURSOS UTILIZADOS 35
4.1. COGROO 35
4.2. REPENTINO 36
4.3. NERP-CRF 37
4.4. LISTAS AUXILIARES 37
4.5. WEKA 38
4.6. CORPUS SUMM-IT 38
4.7. CORPUS DO HAREM 40
5. MODELO DE CLASSIFICAÇÃO 43
5.1. APRENDIZADO DE MÁQUINA 43
5.1.1 Abordagens supervisionadas 43
5.1.2 Abordagens não supervisionadas 46
5.2. SELEÇÃO DE *FEATURES* 46
5.2.1 *Features* na etapa de construção do modelo de classificação 48
5.2.2 *Features* na etapa de construção do sistema de resolução de correferências 49
5.3. CONSTRUÇÃO E SELEÇÃO DO MODELO DE APRENDIZADO 49
6. CORP 53
6.1. ARQUITETURA DO SISTEMA 53
7. AVALIAÇÃO DO CORP 57
7.1. CORPUS DE AVALIAÇÃO 57
7.2. MÉTRICA UTILIZADA 58
7.3. VALIDAÇÃO E TESTES 62
7.4. ANÁLISE DE ERROS 65
8. CONSIDERAÇÕES FINAIS 71
8.1. PUBLICAÇÕES 71
8.2. CONTRIBUIÇÕES DESTE ESTUDO 72
8.3. DESAFIOS 72
8.4. LIMITAÇÕES 73
8.5. TRABALHOS FUTUROS 73
REFERÊNCIAS 75

1. INTRODUÇÃO

1.1. Motivação

Com o crescimento da tecnologia, o acesso à informação ficou muito mais fácil e ágil. Toda essa facilidade, tanto de inserção de informação na web quanto de acesso, culminou em uma grande sobrecarga de conteúdo, requerendo ferramentas que tratem todo este "acúmulo de informações" de forma mais eficiente. Dado esse cenário, as ferramentas computacionais existentes precisam aprimorar-se e adequar-se às novas necessidades. Uma dessas necessidades é a compreensão da linguagem natural por sistemas computacionais. Esse tipo de compreensão em nível computacional não é uma tarefa fácil. Prover conhecimento autônomo a uma máquina, de forma que essa consiga "compreender" o significado semântico existente em determinado conteúdo textual é um dos desafios da área de PLN.

Uma das tarefas de PLN que depende desse tipo de compreensão textual é a resolução de correferências. Resolução de correferências é um processo que consiste em identificar as diversas formas que uma mesma entidade nomeada pode assumir em um determinado texto. Em outras palavras, esse processo consiste em identificar determinados termos e expressões que remetem a uma mesma entidade. Na sentença: "Natália passou no vestibular. A estudante está muito feliz com a notícia", podemos afirmar que "A estudante" é uma correferência de "Natália". Após a etapa de identificação, é possível agrupar essas entidades, formando, assim, cadeias de correferência. A resolução de correferências é uma tarefa relevante e também um grande desafio para a área de linguística computacional. E, tratando-se da língua portuguesa, esse desafio é ainda maior. Isto é, a quantidade de recursos para a língua portuguesa na área de PLN é bem limitada, se comparada com a quantidade de recursos que temos disponíveis para o inglês.

Existem muitos trabalhos voltados a tal assunto, porém grande parte dessas produções científicas está voltada à língua inglesa. Anualmente, competições como a CoNLL [CON11] são realizadas, visando motivar o desenvolvimento de sistemas que resolvam cadeias de correferência, focando principalmente no inglês. A CoNLL disponibiliza um corpus chamado Ontonotes [PRA11] com as anotações de correferência (*gold mentions*). Esse corpus serve para medir a precisão dos sistemas desenvolvidos pelos candidatos.

No âmbito da língua portuguesa, temos o HAREM [FRE10][HAR08]. Assim como a CoNLL, o HAREM é uma atividade de avaliação conjunta com o intuito de incentivar as pesquisas na área de

Processamento da Linguagem Natural, mas com foco na língua portuguesa. Em 2008, uma tarefa relacionada ao reconhecimento de relações entre entidades nomeadas foi proposta pela primeira vez. O HAREM disponibiliza um corpus com as anotações de correferência com o mesmo propósito da CoNLL: avaliar os sistemas desenvolvidos. O grande contraste entre esses dois corpora está no tamanho deles. O corpus do Ontonotes possui 1,3 milhões de anotações, divididas em várias camadas, como: camada sintática, camada de proposições, entidades nomeadas, correferência e word sense disambiguation (WSD). Já o corpus do Harem [FRE10] possui pouco mais de 290 mil anotações. Isso nos dá uma ideia da diferença em relação à quantidade de recursos para as duas línguas. A carência de recursos para a língua portuguesa é um dos desafios e também uma forte motivação deste trabalho. Outro ponto importante é que em uma busca por trabalhos relacionados à resolução de correferências dentro do domínio mencionado, não foram encontrados muitos trabalhos que visem esses domínios específicos – e, quando visam, em sua maioria utilizam recursos proprietários. Neste trabalho optou-se pela utilização de apenas recursos *open source*.

1.2. Objetivos

O presente trabalho tem por objetivo desenvolver um modelo e um recurso que possibilite resolver correferências para a língua portuguesa a partir de documentos de texto livres de anotação, tendo como foco três categorias específicas de entidades nomeadas: Pessoa, Local e Organização. Optou-se por essas três categorias por essas serem as mais relevantes para a maioria das tarefas de PLN, pelo fato de tratarem entidades mais específicas e de interesse comum. Além disso, são as categorias mais exploradas em trabalhos voltados à resolução de correferência.

1.3. Metodologia

Para a concepção deste trabalho, inicialmente foi realizado uma revisão da literatura, buscando por trabalhos voltados à resolução de correferências que trouxessem diferentes abordagens. Essas abordagens podem ser classificadas em: aprendizado supervisionado, aprendizado não supervisionado e baseada em regras determinísticas. Após o estudo dessas diferentes abordagens, optou-se por utilizar aprendizado de máquina supervisionado. Optamos por essa metodologia pelo fato de ser bem consolidada e trazer bons resultados, como podemos ver nos trabalhos de [SOO01] e [FER12]. A arquitetura proposta nesta dissertação divide-se em duas partes: **construção do classificador** e **resolução de correferências**. A etapa de construção do classificador consiste em, por meio de um corpus de treinamento, criar um dataset contendo

amostras de dados positivas e negativas. Para a construção do dataset de treinamento utilizamos o corpus Summ-it [COL07]. O algoritmo de classificação escolhido para a utilização no modelo proposto foi o SimpleCart. Optamos pela utilização do SimpleCart, devido a seus bons resultados e independência de outros recursos externos. Como poderemos ver no capítulo 5 desta dissertação.

Para a etapa de resolução de correferências, assume-se que o modelo não terá as informações providas de um corpus e sim apenas um texto puro dado como entrada. Obviamente o modelo necessita extrair essas informações de algum lugar. Para isso utilizamos a Api CoGrOO [SIL13] e alguns outros recursos, como podemos ver no capítulo 6 desta dissertação.

1.4. Escopo e Resultados Obtidos

Assim como a maioria dos trabalhos existentes na literatura, o modelo proposto também possui suas limitações, como tratar apenas nomes próprios do tipo Pessoa, Local e Organização. Embora não seja possível uma comparação direta com o atual estado da arte, devido aos diferentes escopos e idiomas, o modelo proposto obteve bons resultados, chegando a 77,97% de precisão e 59,38% de cobertura.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo trazemos a fundamentação teórica deste trabalho, apresentando conceitos-base fundamentais para o entendimento deste trabalho, como: O que são referentes, entidades nomeadas, sintagmas, tipos de referentes, correferência e anáfora?

2.1. Referentes

Como o próprio nome sugere, "referente" é a forma como nos referirmos a determinada entidade/sujeito. Em um texto, essas referências podem aparecer como uma entidade nomeada específica ou dentro de um sintagma nominal. Temos também correferência, que consiste na co-ocorrência dessas menções, referindo-se à mesma entidade/sujeito. Esta seção descreve cada um desses conceitos, bem como os tipos de referentes existentes.

2.1.1 Entidades nomeadas

São elementos utilizados para se fazer referência a objetos ou entidades de determinado discurso ou domínio [VIE01]. Os domínios podem ser nomes de pessoas, empresas, lugares, termos de alguma área específica, como genes, proteínas, entre outros. Por meio dos exemplos abaixo, podemos identificar diversas entidades nomeadas (ENs), como Banco Nacional de Desenvolvimento Econômico e Social (2.1.1.1), Apple (2.1.1.2), bandas musicais (2.1.1.3). Qualquer termo pode ser considerado uma entidade. Isso irá depender do foco em questão.

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

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

(2.1.1.3) "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

Segundo [ABR05], um sintagma é uma palavra ou um conjunto de palavras que constitui uma unidade significativa dentro da sentença. Os sintagmas desempenham diferentes funções na sentença e combinam-se em torno de um núcleo. Esse núcleo pode ser um nome ou pronome (sintagma nominal), uma preposição (sintagma preposicional), um adjetivo (sintagma adjetival) e advérbio (sintagma adverbial). Dada a definição de sintagma, bem como as suas categorizações, os sintagmas nominais são as expressões linguísticas utilizadas para referenciar entidades em um discurso. No caso de um sintagma nominal, o núcleo pode configurar-se em nome comum, próprio ou um pronome. Os pronomes podem apresentar-se, basicamente, nas formas de pronome pessoal, demonstrativo, indefinido e possessivo.

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.

(2.1.3.1) "Microsoft informou que irá resolver o bug que reinicia Windows Phone em dezembro."

Nesse caso, 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 que também serão foco deste trabalho, como Pessoa e Local,(2.1.3.2) e (2.1.3.3), respectivamente.

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

(2.1.3.3) "Roger Waters faz seu segundo show em São Paulo."

Nesse exemplo, 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 (2.1.3.4), (2.1.3.5) e (2.1.3.6).

(2.1.3.4) "Policiais invadiram a casa, porém os bandidos já haviam fugido...".

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

(2.1.3.6) "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 (2.1.3.7) e (2.1.3.8)

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

(2.1.3.8) "A viagem foi ótima, porém o tempo podia estar melhor."

2.1.4 Correferência e anáfora:

Para o entendimento do 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 [VIE08]. Há casos de anáfora em que o termo anafórico e o antecedente são correferentes, isto é, remetem a uma mesma entidade (como os Exemplos (2.1.4.1) e (2.1.4.2)

ilustram), mas há também casos de anáfora sem correferência (2.1.4.3) .

(2.1.4.1) "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.

(2.1.4.2) "A sala de aula está degradada. As carteiras estão todas riscadas."

Note que a interpretação referencial do sintagma nominal "as carteiras" depende da sua relação anafórica com o sintagma nominal "a sala de aula".

(2.1.4.3) "O João faz 18 anos no dia 2 de Julho de 2001. No dia seguinte, parte para uma grande viagem pela Europa."

Já nesse exemplo, 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.

Correferência: é um fenômeno que ocorre quando duas ou mais menções no texto referem-se a uma mesma entidade. O conjunto de menções a uma mesma entidade no texto é denominado de cadeia de correferência. Podemos ter dois tipos de correferência, identidade e aposto [PRA11]. O tipo identidade é usado para correferência anafórica, isto é, as ligações entre menções pronominais, nominais e de entidades de referentes específicos.

(2.1.4.4) "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.

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

(2.1.4.6) "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".

(2.1.4.7) "A sala de aula está degradada. As carteiras estão todas riscadas."

Resgatando o exemplo (2.1.4.2), a interpretação referencial do sintagma nominal "as carteiras" depende da sua relação anafórica com o sintagma nominal "a sala de aula". Esse também é um exemplo de aposto, pois a frase "as carteiras estão todas riscadas" nos explica a expressão anterior "a sala de aula está degradada". Nesse caso, "carteiras" é uma correferência do tipo aposto de "sala de aula".

(2.1.4.8) (retirado do texto CIENCIA_2000_6389.txt do corpus Summ-it) [COL07] "A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia. A opinião é do agrônomo 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 cansativo. A dificuldade dessa tarefa pode ser explicada pela dependência da compreensão do contexto, que está relacionada a questões linguísticas e a habilidades cognitivas humanas complexas, de difícil reprodução por sistemas computacionais. 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 podem retomar uma referência ou ativar um novo referente. A anáfora pressupõe um par ordenado (antecedente, anáfora) e a correferência remete à ideia de conjunto [COR10].

3. TRABALHOS RELACIONADOS

Neste capítulo apresentamos os principais trabalhos encontrados na literatura, que relatam métodos de resolução de correferência tanto para o português como para outros idiomas. Na literatura referente à resolução de correferência, encontramos alguns trabalhos que são puramente baseados em regras, outros utilizam uma abordagem mais dinâmica, baseada em aprendizado de máquina. Na CoNLL 2011 [CON11] *(Conference on Computational Natural Language Learning)*, [LEE11] apresentaram seu sistema puramente baseado em regras para a resolução de correferências na língua inglesa. Contrariando o significado da palavra "*learning*", [LEE11] mostraram a eficiência de seu sistema, ficando em primeiro lugar. O sistema, *"Stanford's Multi-Pass Sieve Coreference Resolution System"*, puramente determinístico, atingiu uma eficiência de 57.79%. Essa eficiência foi medida pela média entre três métricas de desempenho (MUC, B-CUBED e CEAFe), descritas em [PRA11]

Em 2012, na CoNLL, [FER12] apresentaram a seguinte estratégia: um sistema de aprendizado de máquina baseado em um algoritmo perceptron. Sua proposta baseou-se em duas principais técnicas de modelagem: *latent coreference trees* e *entropy guided feature induction*. O sistema possui alguns passos básicos, como:

(a) ***Mention detection*****:** para cada documento de texto foi gerada uma lista contendo as menções candidatas, usando a estratégia de [SAN11]. A ideia básica foi usar todos os sintagmas nominais e, adicionalmente, entidades nomeadas. Verbos não foram incluídos como menções.

(b) ***Mention Clustering*****:** na subtarefa agrupamento de menções, uma instância de treino (x,y) consiste em um grupo de menções x para um documento e seus grupos de correferências y. A estrutura do algoritmo perceptron aprende para um dado conjunto de treino D={(x,y)} de pares corretos de entrada/saída.

(c) ***Coreference trees*****:** de forma a reduzir o problema de complexidade de predição das menções, [FER12] utilizaram árvores para representar o agrupamento de menções que são correferentes entre si. Uma árvore de correferência é uma árvore cujos nós são dirigidos às menções, e os arcos representam alguma relação entre menções correferentes.

A tarefa da CoNLL em 2012 foi a resolução de correferência em três idiomas: inglês, chinês e árabe. Segundo [FER12], seu sistema baseado em aprendizado pode ser facilmente adaptado

para diferentes línguas. Em seus experimentos, pequenas mudanças foram necessárias para resolver a correferência em três línguas distintas. A necessidade de adaptações no sistema foi devido a: falta de *features* de entrada para alguns idiomas; diferentes grupos de *tags* utilizadas no *Part-of-Speech* (POS) dos corpora; e a inexistência de uma lista estática de pronomes específicos de cada língua. [FER12] foram os ganhadores da competição CoNLL em 2012, resolvendo correferências em múltiplos idiomas, conforme a tarefa proposta. A precisão do sistema, segundo as métricas utilizadas por [PRA11] , foi de 58.49% para o chinês, 54.22% para o árabe e 63.37% para o inglês, obtendo um escore global de 58.69%.

Outro trabalho realizado para a língua inglesa, mas que teve uma contribuição significativa independentemente do idioma é o de [SOO01]. [SOO01] foram o precursores na tarefa de resolução de correferências utilizando o aprendizado de máquina. Seu trabalho trata a tarefa de resolução de correferências utilizando todos os domínios de textos. A proposta de [SOO01] uniu os conceitos de aprendizado de máquina e processamento de corpus, técnica utilizada por muitos sistemas atuais. Em seu sistema, os autores utilizaram doze *features*, conforme a Tabela 1.

***Feature*:**	**Descrição:**
Str_Match	Se os sintagmas são iguais.
Alias	Se um sintagma é sigla do outro.
I_Pronoun	Se o antecedente é pronome.
J_Pronoun	Se a anáfora é pronome.
Def_NP	Se a anáfora começa pelo artigo the.
Dem_NP	Se a anáfora começa por this, that, these ou those.
Number	Se ambos os sintagmas são numerais.
Gender	Se os sintagmas possuem o mesmo gênero.
Proper_Name	Se os termos são nomes próprios.
Appositive	Se a anáfora é aposto do antecedente.
Dist	Número de frases que separam os termos.
SemClass	Se os sintagmas possuem a mesma categoria semântica. Essa Informação é extraída da WordNet [MIL95].

Tabela 1: *Features* utilizadas por [SOO01]

Para quase todas as *features* são considerados apenas os valores “true” ou “false”, com exceção das *features* “SemClass” e “Dist”: para a *feature* “SemClass”, além de “true” e “false”, é

considerado o valor “unknown”, caso o sistema não identifique a categoria de algum dos sintagmas. A *feature* “Dist” retorna apenas um valor numérico, sendo este a distância entre os dois sintagmas. Devido ao fato de o trabalho de [SOO01] ser o primeiro a abordar aprendizado de máquina para a resolução de correferências utilizando um conjunto de *features*, ele é comumente considerado o *baseline* na área de PLN para a resolução de correferência. Uma das contribuições do trabalho de [SOO01] está no experimento guiado pelo autor: seu principal experimento objetivou verificar a cobertura, precisão e medida-F de cada *feature* separadamente. Como resultado, os autores constataram que apenas as *features* “STR_Match”, “Alias” e “Appositive” tiveram um retorno significativo. De acordo com os autores, esse resultado demonstra que as *features* são importantes para a tarefa de resolução de correferências, pois isoladamente apresentam retorno na classificação dos pares de sintagmas.

Para o português, [SIL11] propôs, em sua dissertação de mestrado, um sistema de resolução de correferência, utilizando um algoritmo de aprendizado não supervisionado. Seu sistema é dividido basicamente em duas fases: identificação das menções (sintagmas nominais) e características e identificação das cadeias de correferência. A primeira fase, a de identificação das menções (SNs) e de suas características, tem como entrada um conjunto de textos. Foram utilizados textos jornalísticos que tratam de um mesmo assunto. Esses textos foram previamente agrupados, já que o sistema em si não possui uma etapa de agrupamento para identificar os sintagmas e extrair os atributos. [SIL11] utilizou o analisador sintático PALAVRAS [BIC00], um reconhecedor de entidades nomeadas, Rembrandt [CAR08], e o tesauro TeP2.0 [MAZ08]. A segunda etapa, a de identificação das cadeias de correferência, recebe como entrada a saída da etapa anterior. Com essa informação, realiza o agrupamento das menções em cadeias. A fase inicia com a utilização de um método não supervisionado de aprendizado de máquina para um primeiro agrupamento. Após esse agrupamento, são aplicadas regras heurísticas com o propósito de melhorar a qualidade das cadeias geradas. Foram utilizadas um conjunto de ferramentas disponíveis para o português, se a disponibilidade de ferramentas fosse como é hoje para o inglês, eles argumentam que ainda poderiam melhorar seus resultados. Os resultados da avaliação de seu sistema mostraram-se promissores: 58.11% utilizando a medida MUC, e 60.07% utilizando B-CUBED. Apesar de não ser possível uma comparação direta com os sistemas de [LEE11] e de [FER12], devido às diferenças de corpora, de línguas e dos tipos de entidades tratadas, o trabalho proposto por [SIL11] teve uma significativa contribuição por tratar desse domínio para o português.

[COR10] propõe a resolução de correferências com foco nas categorias de entidades

nomeadas. Segundo o autor, trabalhos voltados à língua inglesa obtiveram bons resultados utilizando categorias de entidades específicas. Baseando-se nessas premissas, [COR10] partiu da hipótese de que o uso de categorias específicas de entidades nomeadas tem um impacto positivo na tarefa de resolução de correferência, já que cada categoria apresenta características distintas e bem definidas. Como a categorização delimita o domínio, torna-se mais viável o uso de informação semântica como instrumento de apoio no processo de resolução de correferência. O sistema de [COR10] baseou-se em aprendizado de máquina, categorização de entidades nomeadas, como Pessoa, Organização, Local, Obra, Coisa e Outro, provenientes do corpus do HAREM [FRE10], do analisador sintático PALAVRAS [BIC00] e de recursos do corpus Summ-it [COL07]. [COR10] compara duas versões do sistema, sendo elas: Baseline e Recorcaten (REsolução de CORreferência por CATegorias de ENs). A primeira versão teve como objetivo gerar os pares de sintagmas sem considerar as categorias de ENs. Já a segunda gera os pares considerando essas categorias de entidades. Como contribuição, por meio de experimentos com as duas versões, [COR10] mostrou que o uso de categorias de entidades proporcionou uma melhora no percentual de acerto ao definir se um par é anafórico ou não. Mostrou também a importância do conhecimento de mundo para essa linha de pesquisa, dado o fato de que algumas categorias, como as de Acontecimento e de Organização, não apresentaram um retorno satisfatório na classificação dos pares correferentes. Isso porque existem certas dificuldades no processo de desambiguação de palavras (WSD), ressaltando, assim, a importância de bases com sinônimos, como a da Wordnet [MIL95], para complementar e apoiar a resolução de correferência. As limitações dessa produção deram-se no tamanho do corpus utilizado nos experimentos. Conforme mostrado anteriormente, ainda existem poucos recursos para o português. Outra consideração, segundo o autor, foram os problemas de anotação do analisador sintático utilizado.

Como já mencionamos, existem muitos trabalhos dentro do contexto de resolução de correferências, porém esses, em sua maioria, são protótipos, isto é, modelos que resolvem correferências apenas para um corpus específico, visando calcular sua precisão e abrangência. O trabalho contido na presente dissertação objetiva ir mais além. Isto é, o objetivo é desenvolver um modelo de resolução de correferências que não se limite à construção de um classificador, mas que obtenha como produto final uma ferramenta *open source*, que receba como entrada um texto puro e obtenha como saída um arquivo XML, contendo anotação de correferências desse texto. Essa proposta é semelhante à de [SIL11], com o diferencial de que neste trabalho são utilizados apenas recursos livres.

4. RECURSOS UTILIZADOS

Resolução de correferência é uma tarefa complexa e depende da saída de diversos recursos de PLN. Este capítulo relata todos os recursos utilizados na elaboração deste trabalho. Como veremos mais adiante, o sistema de resolução de correferências descrito nesta dissertação foi construído em duas etapas, sendo elas: treinamento e construção do modelo de classificação e construção do sistema de resolução de correferências.

4.1. CoGrOO

CoGrOO [SIL13] 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 estatísticas de Processamento de Linguagens Naturais 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 é uma ferramenta que, da mesma forma que o OGMA [MAI08], 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. Como veremos adiante, esses recursos proporcionados pela ferramenta foram indispensáveis para a construção do sistema de resolução de correferências. Na Figura 1 podemos visualizar um exemplo de anotação fornecida pela ferramenta para a sentença: “O tecno de Detroit é arquitetura.”.

```
Sentence: O tecno de Detroit é arquitetura.
  Tokens:
    O          [o]          art    M=S
    tecno      []           n      M=S
    de         [de]         prp    -
    Detroit    [Detroit]    prop   M=S
    é          [ser]        v-fin  PR=3S=IND
    arquitetura [arquitetura] n      F=S
    .          [.]          .      -
  Chunks: [NP: O tecno ] [PP: de ] [NP: Detroit ] [VP: é ] [NP: arquitetura ]
  Shallow Structure: [SUBJ: O tecno de Detroit ] [P: é ] [SC: arquitetura ]

Enter the sentence or 'q' to quit:
```

Figura 1: Saída CoGrOO

Como podemos visualizar, na Figura 1 a ferramenta fornece como saída:

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

4.2. Repentino

O Repentino [REP05], 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 2, podemos visualizar como está disposta a organização das entidades de categoria “EN_SER” (seres vivos) e subcategoria “HUM” (humanos).

```
<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>
```

Figura 2: Anotações Repentino

4.3. NERP-CRF

Desenvolvido por [AMA13], o NERP-CRF é um Sistema de Reconhecimento de Entidades Nomeadas por meio de *Conditional Random Fields* para a Língua Portuguesa. Isto é, o NERP-CRF é, assim como um identificador de entidades nomeadas, um classificador. Dado um texto, ou um conjunto de textos, o sistema permite processá-lo(s), tendo como saída todas as entidades nomeadas presentes nele, incluindo suas categorias, como: Pessoa, Local, Organização, Acontecimento, entre outras. As anotações do NERP-CRF seguem um padrão bastante parecido com o corpus do HAREM [FRE10]. Na Figura 3, podemos visualizar o padrão de saída da ferramenta, proveniente da anotação de um texto do corpus Summ-it [COL07].

```
<!DOCTYPE colHAREM>
<colHAREM versao="NERP_CRF 02 Apr 2013">
<DOC DOCID="CIENCIA_2000_6381">
Após o anúncio do sequenciamento do genoma , na semana passada , a <EM
ID="CIENCIA_2000_6381" CATEG="LOCAL">França </EM>resiste como único país da <EM
ID="CIENCIA_2000_6381" CATEG="LOCAL">União Européia </EM>a não permitir patenteamento de
genes .
A <EM ID="CIENCIA_2000_6381" CATEG="ORGANIZACAO">UE </EM>adota , desde junho de 1998 ,
diretiva favorável ao patenteamento de genes.O texto , redigido pelo <EM
ID="CIENCIA_2000_6381" CATEG="ORGANIZACAO">Parlamento Europeu </EM>, <EM
ID="CIENCIA_2000_6381" CATEG="PESSOA">Comissão Européia </EM>e Conselho de Ministros ,
utiliza o princípio de que "o genoma não é patenteável , mas a sequência de um gene pode
ser" .No entanto , há restrições .
O patenteamento só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o
funcionamento do gene é detalhado.A
França é o único país que se recusa a aceitar a determinação européia .
A ministra da Justiça do país , <EM ID="CIENCIA_2000_6381" CATEG="PESSOA">Elisabeth Guigou
```

Figura 3: Saída NERP-CRF

4.4. Listas Auxiliares

Antes do treinamento do modelo de classificação, foi notado que ambos os recursos utilizados com o propósito de etiquetagem de entidades (Repentino e NERP-CRF) possuíam limitações dentro do contexto de busca utilizado. No caso do Repentino, o problema são as ambiguidades. Por exemplo, ao buscarmos pela entidade 'Amazônia', o Repentino pode conter tanto a categoria 'Local' quanto 'Organização' – respectivamente, 'Amazônia' e 'Banco da Amazônia'. Já o sistema NERP-CRF possui uma taxa de acerto de 83.99% – isto é, não classifica todas as entidades corretamente.

Pensando nessas duas limitações, a ideia foi procurar pelas entidades presentes nos SNs em ambos os recursos, alinhando seus resultados. Dessa forma, a saída torna-se mais confiável e precisa. Sempre que ambos os recursos retornarem o mesmo resultado, é assumido que a categoria da entidade nomeada foi aferida corretamente, porém os resultados desses sistemas

nem sempre concordam entre si. Para isso, foram criadas três listas, uma para cada categoria de entidade: Pessoa, Local e Organização. Quando os recursos retornam resultados diferentes, essas três listas são percorridas, visando etiquetar corretamente a categoria da entidade em questão. Essas listas possuem nomes comuns e próprios, que ajudam a identificar o tipo de entidade nomeada. 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. Para a construção do modelo de classificação, essas listas auxiliares são utilizadas quando inexiste a concordância entre as etiquetas do Repentino e do NERP-CRF. Já para o sistema de resolução de correferência, como veremos mais adiante, essas listas são utilizadas em conjunto apenas com Repentino. As listas auxiliares foram construídas por meio de conteúdo proveniente da WIKIPEDIA [WIK13], uma enciclopédia livre, disponível em diversos idiomas.

4.5. Weka

O Weka [BOU13] é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Possui recursos de pré-processamento, classificação, agrupamento, visualização, entre outros. Sua implementação se dá em linguagem Java, que tem como principal característica ser portável. Por isso, o Weka pode ser executado nas mais variadas plataformas, aproveitando os benefícios de uma linguagem orientada a objetos. O Weka possui também uma Api, que pode ser executada diretamente por linha de comando. A ferramenta Weka foi indispensável neste trabalho, já que essa possibilitou a criação e validação do modelo de treinamento utilizado [FON13b].

4.6. Corpus Summ-it

O Summ-it [COL07] é um corpus composto por cinquenta textos jornalísticos do caderno de Ciências da Folha de São Paulo, retirados do corpus PLN-BR [BRU08]. Cada documento corresponde a um arquivo de texto (ASCII) com tamanho entre 1 kbyte e 4 kbytes (de 127 a 654 palavras). Os textos foram anotados com informação sintática, de correferência e de estrutura retórica. O Summ-it também conta com sumários construídos de forma manual e automática. Com o intuito de melhorar a visualização das informações extraídas do analisador, o arquivo gerado foi

dividido em três outros arquivos: um arquivo com as informações dos *tokens*, composto pelo *token* e seu respectivo ID (Figura 4); outro com as informações dos sintagmas (Figura 5), isto é, qual o ID do *token* inicial e final do sintagma; e outro com as informações sintático-semânticas associadas ao ID do *token* (Figura 6). Os arquivos estão em formato XML. O corpus Summ-it teve um importante papel para o treinamento e validação do modelo de classificação.

```
<word id="word_1">Astrônomos</word>
<word id="word_2">brasileiros</word>
<word id="word_3">esperam</word>
<word id="word_4">fotografar</word>
<word id="word_5">os</word>
<word id="word_6">primeiros</word>
<word id="word_7">planetas</word>
<word id="word_8">fora_de</word>
<word id="word_9">o</word>
<word id="word_10">Sistema_Solar</word>
```

Figura 4: Tokens Corpus Summ-it

```
<markable id="markable_1" span="word_1..word_2"
np_n="yes" np_form="bare-np"/>
<markable id="markable_76" span="word_5..word_10"
status="new" np_n="yes" np_form="def-np"/>
<markable id="markable_3" span="word_9..word_10"
member="set_4" status="new" np_n="yes"
np_form="def-pn"/>
<markable id="markable_77" span="word_12..word_23"
status="new" np_n="yes" np_form="def-np"/>
<markable id="markable_78" span="word_15..word_23"
member="set_11" status="new" np_n="yes"
```

Figura 5: Sintagmas Nominais

```
− <word id="word_1">
  − <n canon="astrônomo" gender="M" number="P">
      <secondary_n tag="Hprof"/>
    </n>
  </word>
− <word id="word_2">
    <adj canon="brasileiro" gender="M" number="P"/>
  </word>
− <word id="word_3">
  − <v canon="esperar">
      <fin tense="PR" person="3P" mode="IND"/>
    </v>
  </word>
```

Figura 6: Informações sintático-semânticas

Como podemos visualizar nas Figuras 4, 5 e 6, as informações do corpus Summ-it estão divididas em fragmentos. Para capturar e estruturar toda essa informação de maneira correta, foi desenvolvido um programa em Java. Este programa, basicamente capturou e estruturou as informações, de maneira que essas pudessem ser manipuladas.

4.7. Corpus do HAREM

Como já mencionado no primeiro capítulo desta dissertação, o HAREM [FRE10] é um corpus anotado proveniente de uma avaliação conjunta, com o intuito de incentivar as pesquisas na área de Processamento da Linguagem Natural. O corpus é constituído por 129 textos, contendo ao todo 290 mil anotações de ENs, bem como suas categorias e relações de correferência. Na Figura 7 podemos visualizar como estão dispostas essas categorias no corpus.

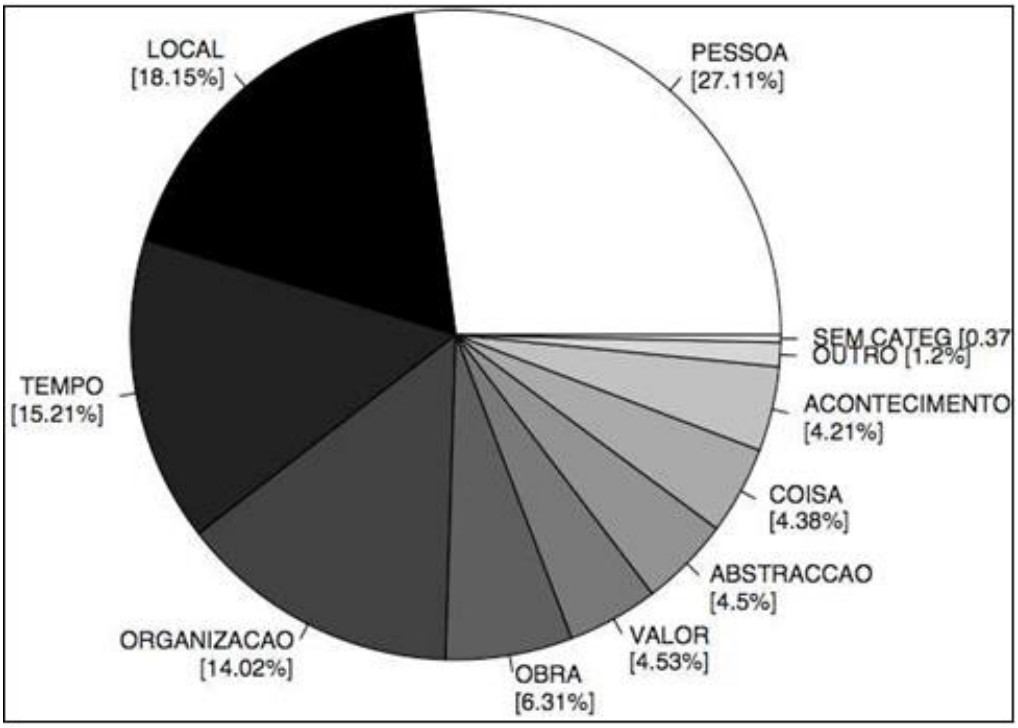

Figura 7: Distribuição das categorias do segundo HAREM [FRE10]

Na Figura 7 podemos visualizar que as categorias mais frequentes no corpus são: PESSOA, com 27.11%; LOCAL, com 18.18%; TEMPO, com 15.21% e ORGANIZAÇÃO, com 14.02%. Como o objetivo deste trabalho é resolver correferências paras as categorias PESSOA, LOCAL e ORGANIZAÇÃO, o corpus do HAREM tornou-se o ideal para a validação deste sistema.

5. MODELO DE CLASSIFICAÇÃO

Para a concepção do modelo de classificação foi utilizado aprendizado de máquina. Este capítulo explica o que são algoritmos de aprendizado de máquina e seus diferentes tipos. Além disso, relata de forma detalhada, como foi realizada a construção do modelo de classificação e descreve as seguintes etapas: seleção e implementação das *features*; testes; geração e escolha do modelo.

5.1. Aprendizado de Máquina

No atual estado da arte, existem dois tipos de algoritmos de aprendizado de máquina comumente utilizados para a resolução de correferência: os algoritmos supervisionados e os não supervisionados. Nas subseções 5.1.1 e 5.1.2, são descritos os dois métodos, dando mais ênfase para a abordagem supervisionada, método utilizado para a concepção do sistema de resolução de correferências descrito nesta dissertação. Na seção 5.2, é descrito como o conjunto de *features* foi escolhido e elaborado.

5.1.1 Abordagens supervisionadas

As abordagens supervisionadas são dentre as que utilizam aprendizagem de máquina, as mais exploradas na literatura, em parte pela disponibilização dos corpora anotados, como Summ-it[COL07], HAREM[FRE10], Ontonotes[PRA11]. Trabalhos como os de [COR10] e [FER12] utilizam essa abordagem. As abordagens supervisionadas consistem na construção de um classificador que seja capaz de determinar quais são os SNs correferentes de dado conjunto de textos, anotados com informações de correferência. A abordagem supervisionada é dividida em duas etapas, a de treinamento e a de testes. A etapa de treinamento consiste em criar um *dataset, contendo* um conjunto de *features* distribuídas em valores booleanos (incluindo a informação se o par é correferente ou não), conforme podemos visualizar na Figura 8.

```
@relation coreference

@attribute String_Match {true, false} 1
@attribute Alias {true, false} 2
@attribute Genero { true, false } 3
@attribute Numero {true, false} 4
@attribute categoria_igual {true, false} 5
@attribute categoria_diferente {true, false} 6
@attribute Distancia5 {true, false} 7
@attribute Distancia10 {true, false} 8
@attribute Distancia15 {true, false} 9
@attribute correfentes{true, false} 10

@data
1 2 3 4 5 6 7 8 9 10
true , false , true , true , true , false , false , false , false , true
true , false , true , true , false , false , false , false , false , true
false , false , true , false , true , false , false , false , false , true
true , false , true , true , false , false , false , false , false , true
true , false , true , true , false , false , false , false , false , true
false , false , true , false , false , false , true , false , false , false
false , false , true , false , false , false , true , false , false , false
false , false , false , false, false , false , true , false , false , false
false , false , true , true , false , false , true , false , false , false
false , false , false , true , false , false , true , true , false , false
false , false , true , true , false , false , true , true , false , false
false , false , true , true , false , false , true , true , false , false
```

Figura 8: Dataset de treinamento.

Na Figura 8, podemos visualizar um exemplo de dataset (arquivo .arff) utilizado para treinar um modelo. Note que os atributos/*features* são declarados no início do documento, incluindo os valores que cada *feature* pode assumir. A marca "@data" refere-se ao início do conjunto de dados. Cada linha refere-se às *features* de um par de sintagmas. Este *dataset* possui um total de nove features mais uma décima, informando se o par é correferente ou não. Esta etapa de aprendizado é dependente das informações provenientes de um corpus de treinamento, como o Summ-it [COL07], por exemplo.

A etapa de testes consiste em testar o modelo de classificação construído pela primeira etapa, no mesmo *dataset*. Para essa etapa, é como se o classificador ignorasse a "*feature* supervisionada" 'correferentes', tendo como entrada para o modelo já treinado apenas as outras nove *features*. Com base nos atributos, o modelo tenta prever quais pares são correferentes ou não. Ao fim do processo, a saída é comparada com o dataset original, tendo assim uma precisão do presente modelo criado. Também é possível testar o classificador com outro conjunto de dados. Para isso, basta utilizar o modelo aprendido, em outro *dataset,* que seja supervisionado e compatível. Isto é, que possua o mesmo grupo de *features*.

Na Figura 9, podemos visualizar a arquitetura supervisionada referente ao sistema de

resolução de correferências proposto nesta dissertação. Podemos notar que o sistema é dividido em duas etapas: a de construção do classificador, que consiste em construir e validar o modelo de classificação e a fase de resolução de correferências: nessa fase, o sistema deve extrair os pares de sintagmas de um documento de texto puro (livre de anotação) e, por meio da submissão desses pares ao classificador previamente construído, decidir quais desses pares são correferentes.

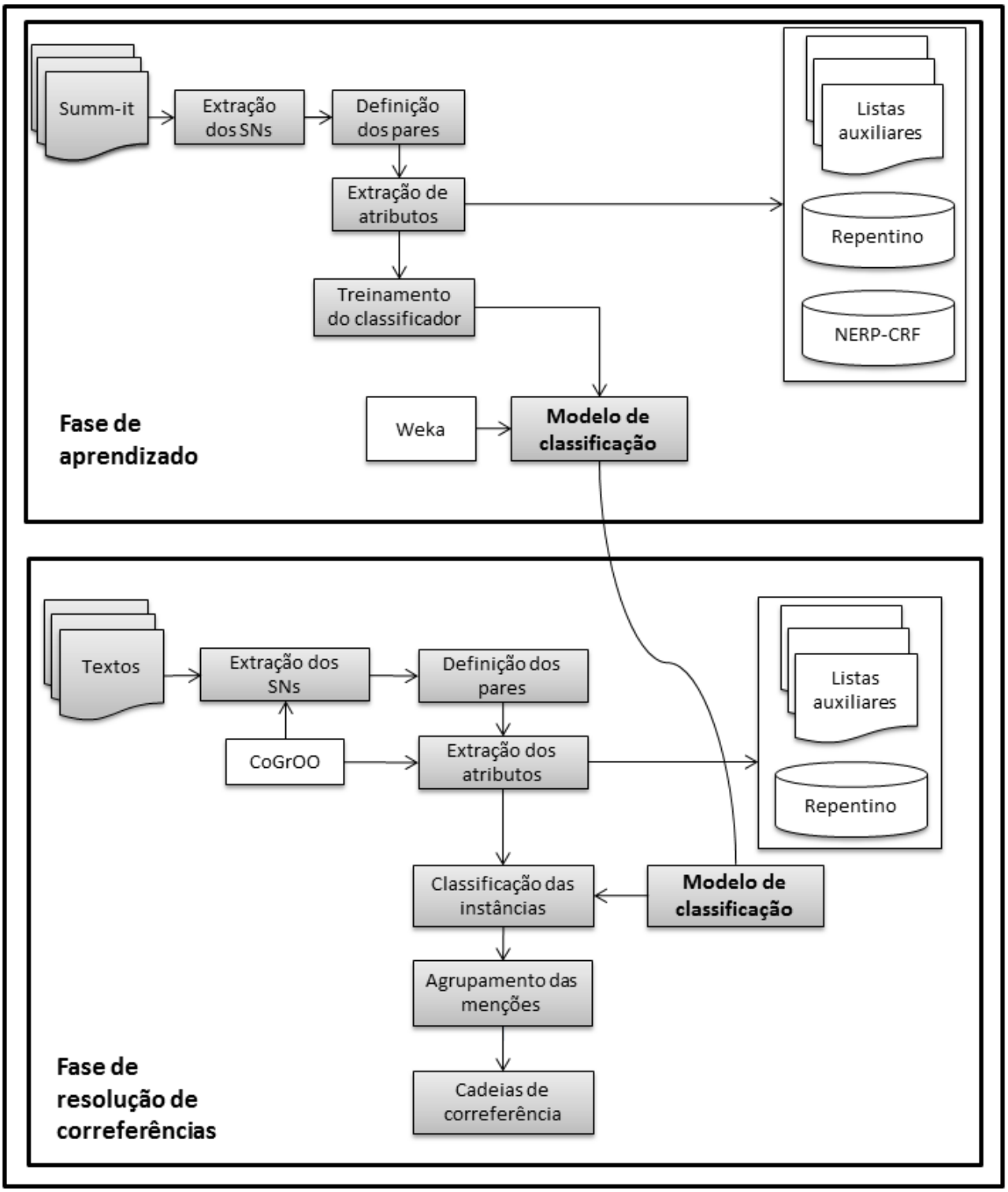

Figura 9: Arquitetura supervisionada proposta nesta dissertação.

5.1.2 Abordagens não supervisionadas

As abordagens não supervisionadas para a resolução de correferência partem do pressuposto que é possível considerar cada cadeia de correferência como uma classe. Dessa forma, a partir de atributos específicos/*features*, é possível agrupar sintagmas e gerar essas cadeias. Os métodos não supervisionados são basicamente mais objetivos do que os supervisionados, não sendo necessário informar qual sintagma é correferente de qual. Na Figura 10, tal como apresentada em [SIL11], podemos visualizar como de dispõe um modelo não supervisionado.

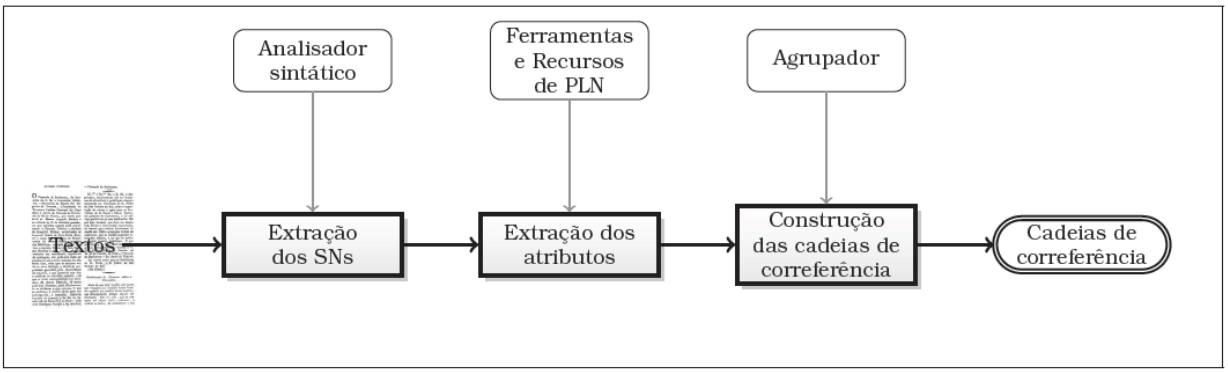

Figura 10: Aprendizado não supervisionado, retirado de [SIL11]

Como podemos observar na Figura 10, uma arquitetura não supervisionada não possui etapa de treinamento. A primeira etapa dessa arquitetura é similar a todo sistema de resolução de correferência, no qual os SNs são extraídos. Na segunda etapa, é realizada a extração de atributos/*features*. A extração de *features* se dá em nível de menções, e não por pares de sintagmas, como no aprendizado supervisionado. Após a extração dos atributos, o sistema tenta agrupá-los de acordo com seus atributos em comum, formando cadeias de correferência.

5.2. Seleção de *Features*

A seleção de *features* para gerar o modelo de classificação deu-se com base no estudo da literatura, principalmente com base nos resultados obtidos nos experimentos de [SOO01]. Soon et al. realizaram um experimento com o propósito de verificar o impacto de determinadas *features* na correta classificação de pares correferentes. Como resultado, os autores constataram que algumas *features*, como String Match, Alias e Aposto, apresentam um retorno significativo na correta classificação dos pares. Além das *features* mais relevantes existentes em outros trabalhos,

por meio de experimentos realizados foi constatada a necessidade da adição de *features que calculassem a distância para dado par de sintagmas*, conforme podemos visualizar na Tabela 2.

P_StringMatch	Se pelo menos uma String do SNx está contida em SNy.(exceto stopwords)
Alias	Se uma das palavras de SN1 é sigla de SN2.
M_Gênero	Se os sintagmas concordam em gênero (masculino/feminino).
M_Número	Se os sintagmas concordam em número (singular/plural).
Categ_semântica_Igual	Se as categorias de entidades (Pessoa, Local ou Organização) são iguais.
Categ_semântica_Diferente	Se as categorias de entidades (Pessoa, Local ou Organização) são diferentes.(Quando o sistema não consegue aferir uma categoria, ambas, Categ_igual e Categ_Dif são *false*).
Distância>5	Se a quantidade de sentenças entre um sintagma e outro é maior que 5, retorna *true*, *false* caso contrário.
Distância>10	Se a quantidade de sentenças entre um sintagma e outro é maior que 10, retorna *true*, *false* caso contrário.
Distância>15	Se a quantidade de sentenças entre um sintagma e outro é maior que 15, retorna *true*, *false* caso contrário.

Tabela 2: Descrição das *features*

A construção do modelo de resolução de correferências é feita em duas etapas: a de treinamento, que objetiva construir e validar um modelo de classificação de pares de sintagmas, e a etapa de construção do sistema. As *features* utilizadas nas duas etapas são as mesmas, no entanto a forma como cada uma é obtida muda de acordo com a etapa.

Na etapa de treinamento a maioria das features foi adquirida a partir do corpus Summ-it. Já na construção do sistema de resolução de correferências, obter esse tipo de característica por meio de um corpus é inviável, dado o fato que a entrada do sistema deverá ser um arquivo de texto puro, livre de qualquer anotação. Nas seções 5.2.1 e 5.2.2, é descrito como se deu a implementação e a obtenção de informações nas duas etapas.

5.2.1 *Features* na etapa de construção do modelo de classificação

A construção dos pares de SNs foi realizada com base nas informações de correferência, contidas no *corpus* Summ-it. Tratando-se das *features*, alguns cuidados foram tomados na implementação, de forma a calibrar a precisão e a abrangência de cada uma delas. No caso da *feature Parcial_String_Match*, foram removidas todas as *stopwords* de cada um dos SNs. Em seguida, é utilizado o algoritmo Jaccard [MAN99], para o cálculo de similaridade, objetivando verificar o quão semelhante uma *string* é da outra. O cálculo de similaridade retorna um número próximo de '1', para *strings* idênticas, ou números próximos de '0', para *strings* completamente diferentes. Por meio de testes, notou-se que existiam algumas palavras que eram bastante semelhantes, mas muitas vezes retornavam o valor *false* por ter uma pequena variação, como: "profissional" e "profissionais". Por meio de experimentos preliminares, foi definido o limiar de '0.8', considerando o matching positivo quando seu retorno for maior ou igual a '0.8'. Essa foi a forma que melhor surtiu resultados, minimizando falsos positivos e aumentando abrangência e precisão. Para as *features* M_Gênero e M_Número, as informações foram extraídas do *corpus* Summ-it.

Para a *feature Alias,* assim como na *Parcial_String_Match*, utilizou-se o cálculo de similaridade. Para cada palavra existente em um SN (excluindo-se as stopwords), se a letra inicial for maiúscula, essa é selecionada. Como resultado, tem-se uma *string* com as iniciais do sintagma com e sem pontos '.', como no exemplo: SN1="Instituto Nacional de Pesquisas Espaciais, Inpe" e SN2="INPE". As siglas geradas pela *feature* serão "I.N.P.E.I." e "INPEI". Após esse passo, as siglas geradas são comparadas com cada palavra do SN2 pelo cálculo de similaridade. Notemos que 'INPEI' não é exatamente igual ao SN2 'INPE', porém o cálculo de similaridade dará um resultado muito próximo de "1", o que indica que as *strings* são bastante semelhantes. Com isso, conclui-se que SN2 é sigla de SN1.

Na *feature Cat_Semântica*, foram utilizados três recursos já mencionados anteriormente: o Repentino, o NERP-CRF e listas auxiliares que visam identificar e aferir uma etiqueta de Local, Pessoa e Organização para os pares de sintagmas. Quando o sistema não consegue aferir uma categoria a determinada entidade, o valor *false* é atribuído a estas *features*: 'Categ_semântica_Igual' e 'Categ_semântica_Diferente'. Essa foi uma das formas encontradas para fazer com que o algoritmo de aprendizado consiga diferenciar as entidades com categorias não reconhecidas.

A implementação da *feature Distância_5, 10* e *15* baseou-se nas premissas de que, quanto mais longe uma sentença está da outra em um texto, menores são as chances de essas serem

correferentes. Essa *feature* utiliza como parâmetro o texto original, livre de marcações e dois SNs. Por meio dos dois SNs, a *feature* forma uma expressão regular, capturando do texto todo o trecho existente entre o SN1 e o SN2. Com isso, são contados o número de sentenças (essa contagem foi realizada contando a pontuação do texto, como '.', '!'e '?'). Como resultado, é aferido um valor *true* ou *false* para cada uma dessas três *features*. Os casos em que existe mais de um sintagma idêntico no texto não são um problema, pois cada sintagma, ao ser extraído, recebe um "Id" único, podendo ser diferenciado dos demais sintagmas iguais em escrita.

5.2.2 *Features* na etapa de construção do sistema de resolução de correferências

Na etapa de construção do sistema, o meio como algumas informações são obtidas é diferente da etapa de construção do classificador: nessa etapa, a extração de sintagmas é realizada por meio da ferramenta CoGrOO [SIL13], assim como as informações de gênero e número, necessárias para as respectivas *features*. Por utilizar recursos próprios da implementação, as *features 'Parcial_String_Match'* e *'Alias'* não sofreram alterações, sendo exatamente iguais às descritas na subseção 5.2.1.

Para as *features* 'Categ_semântica_Igual' e 'Categ_semântica_Diferente' optou-se por não utilizar a ferramenta NERP-CRF. O motivo por trás da decisão está na inviabilidade de embarcar a ferramenta junto ao sistema, de forma que sua execução seja linear, sem intervenção por parte do usuário. Desta forma, utilizou-se apenas as informações existentes no repositório Repentino e nas três listas auxiliares. Abrindo mão da alta precisão, nesta etapa, caso a entidade seja encontrada em um dos recursos a categoria é aferida imediatamente. Para as *features* Distância >5, 10 e 15 o meio de extração foi o mesmo mencionado na subseção 5.2.1.

5.3. Construção e Seleção do Modelo de Aprendizado

Vários modelos de classificação foram gerados antes de se chegar ao escolhido. O primeiro deles, antes da implementação da *feature* 'Distância', acabou por não ter um retorno satisfatório na classificação dos pares [FON13a]. Isto é, mesmo com seus 70.7% de acerto, o modelo classificava muitos pares negativos como positivos. O problema de falsos positivos foi o responsável pela exclusão da maioria dos classificadores gerados em experimentos anteriores. Depois das primeiras tentativas, optou-se pelo conjunto de *features* descritas na Tabela 2. Na Tabela 3, podemos visualizar os resultados de experimentos realizados com alguns algoritmos de classificação, objetivando medir o desempenho de cada um deles. Obviamente, todos os

experimentos existentes na Tabela 3 foram realizados com o mesmo conjunto de dados.

Para o treinamento desses modelos foram utilizados 5431 pares, sendo 1576 positivos e 3855 negativos. Como podemos notar, não foi utilizada uma quantidade de pares exatamente balanceada. Dadas as restrições de obtenção de pares, os 50 textos do corpus Summ-it retornaram 3855 pares negativos e aproximadamente 400 pares positivos. Para não cortar a quantidade de pares negativos a fim de igualá-la à quantidade de positivos e, consequentemente, correr o risco de perder características importantes que possam existir em alguns pares, optou-se por replicar os pares positivos até que chegassem a uma quantidade aceitável para gerar o modelo, preservando, assim, todos pares negativos.

-	Precision A	Recall A	F-Measure A	Precision B	Recall B	F-Measure B	Correctly Classified Instances
SimpleCart	76.8%	88.9%	82.4%	95.1%	89.0%	92.0%	88.97%
BFTree	76.4%	88.9%	82.2%	95.1%	88.8%	91.9%	88.82%
REPTree	75.4%	88.9%	81.6%	95.1%	88.1%	91.5%	88.36%
J48	76.3%	88.9%	82.1%	95.1%	88.7%	91.8%	88.78%
LBR	79.5%	87.3%	83.2%	94.6%	90.8%	92.7%	89.79%
NaiveBayes	76.8%	84.1%	80.3%	93.3%	89.6%	91.4%	88.01%
Random Forest	76.7%	89.8%	82.8%	95.5%	88.9%	92.1%	89.15%
Multilayer Perceptron	79.1%	89.8%	84.1%	95.6%	90.3%	92.9%	90.16%

Tabela 3: Precisão dos classificadores [FON13b].

Como podemos visualizar na Tabela 3, o algoritmo com melhores resultados foi o Multilayer Perceptron, seguido pelo LBR, Random Forest e SimpleCart. Nas Tabelas 4, 5, 6 e 7 podemos visualizar as respectivas matrizes de confusão para os quatro algoritmos mencionados, em que ‘A’ representa a classe de pares Positivos (correferêntes), e ‘B’, a classe de pares negativos (não correferentes).

-	A	B
A	1416	160
B	374	3481

Tabela 4: Multilayer Perceptron.

-	A	B
A	1376	200
B	354	3501

Tabela 5: LBR.

-	A	B
A	1416	160
B	429	3426

Tabela 6: Random Forest.

-	A	B
A	1401	175
B	424	3431

Tabela 7: SimpleCart

Mapeando as linhas e colunas das matrizes de confusão acima (Tabelas 4, 5, 6 e 7), temos:

- AA: Verdadeiros Positivos
- AB: Falsos Negativos
- BA: Falsos Positivos
- BB: Verdadeiros Negativos

Para a escolha do algoritmo de classificação a ser utilizado no sistema, levou-se em consideração dois fatores: sua taxa de acerto, isto é, quão preciso o algoritmo é na classificação de ambas as classes (A e B), bem como a sua portabilidade. Como o sistema de resolução de correferências é dependente de muitos recursos, adotou-se uma política de usabilidade: a ideia é optar sempre por recursos que permitam ser embarcados ao sistema. Dessa forma a execução da ferramenta torna-se algo simples e linear, sem muitas dependências. Para a utilização dos algoritmos Multilayer Perceptron, LBR ou Random Forest, seria necessário salvar o modelo de classificação gerado e, por meio da API do Weka [BOU13], utilizar o classificador selecionado. Pensando nessas dependências, optou-se pelo algoritmo SimpleCart. Como o algoritmo trabalha com árvore de decisão, pode ser facilmente implementado diretamente no código fonte do sistema, descartando a chamada de outros recursos. Na Figura 11 podemos visualizar a árvore de decisão gerada pelo SimpleCart, utilizada no sistema.

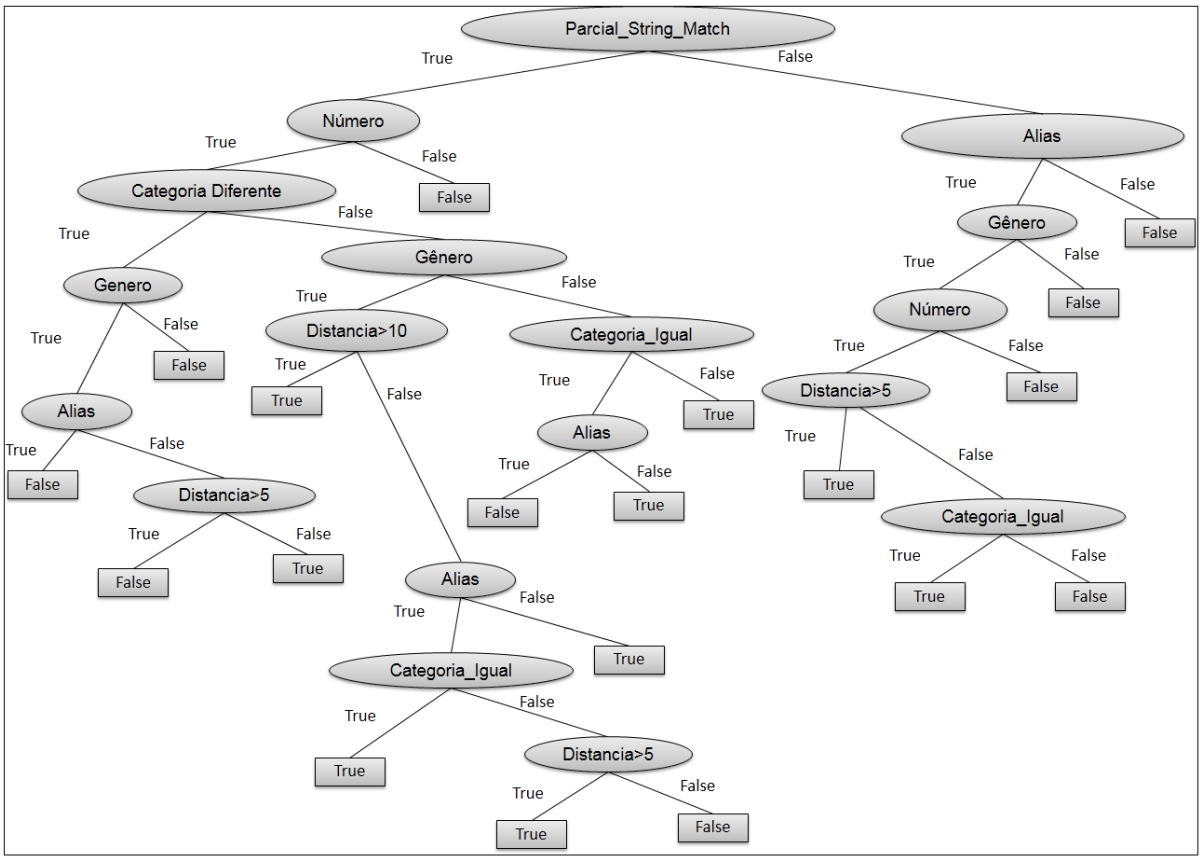

Figura 11: Árvore de decisão gerada pelo algoritmo SimpleCart

6. CORP

O CORP[1] (Coreference Resolution for Portuguese) é um sistema de resolução de correferências para a língua portuguesa, construído totalmente com recursos *open source*. O objetivo dessa ferramenta é auxiliar nas mais diversas tarefas de PLN. Conforme [GAB11] aponta, a resolução de correferências pode prover ganhos significativos para tarefas da área de Processamento da Linguagem Natural. Um bom exemplo disso é a extração de relações entre entidades nomeadas [ABR13]. Identificando as várias formas de referenciarmos a mesma entidade em um determinado texto, é possível tornar mais eficiente o processo de extração de relação entre entidades. Por exemplo, considere a seguinte sentença: "José da Silva reside próximo à Cidade Baixa, em Porto Alegre. Silva está no primeiro ano de seu mestrado na PUC-RS.". Identificando e criando uma relação de correferência entre as entidades 'José da Silva' e 'Silva', é possível inferir uma relação direta entre as entidades 'Silva' e 'Cidade Baixa' (Silva reside na Cidade Baixa em Porto Alegre), assim como é possível dizer que 'José da Silva' tem relação com PUC-RS (José da Silva é aluno da PUC-RS).

6.1. Arquitetura do Sistema

Nesta Seção, é apresentada a arquitetura da ferramenta CORP, em que relataremos como ocorre o processo de resolução de correferência passo a passo.

[1] Recurso disponível em http://www.inf.pucrs.br/linatural/corp.html

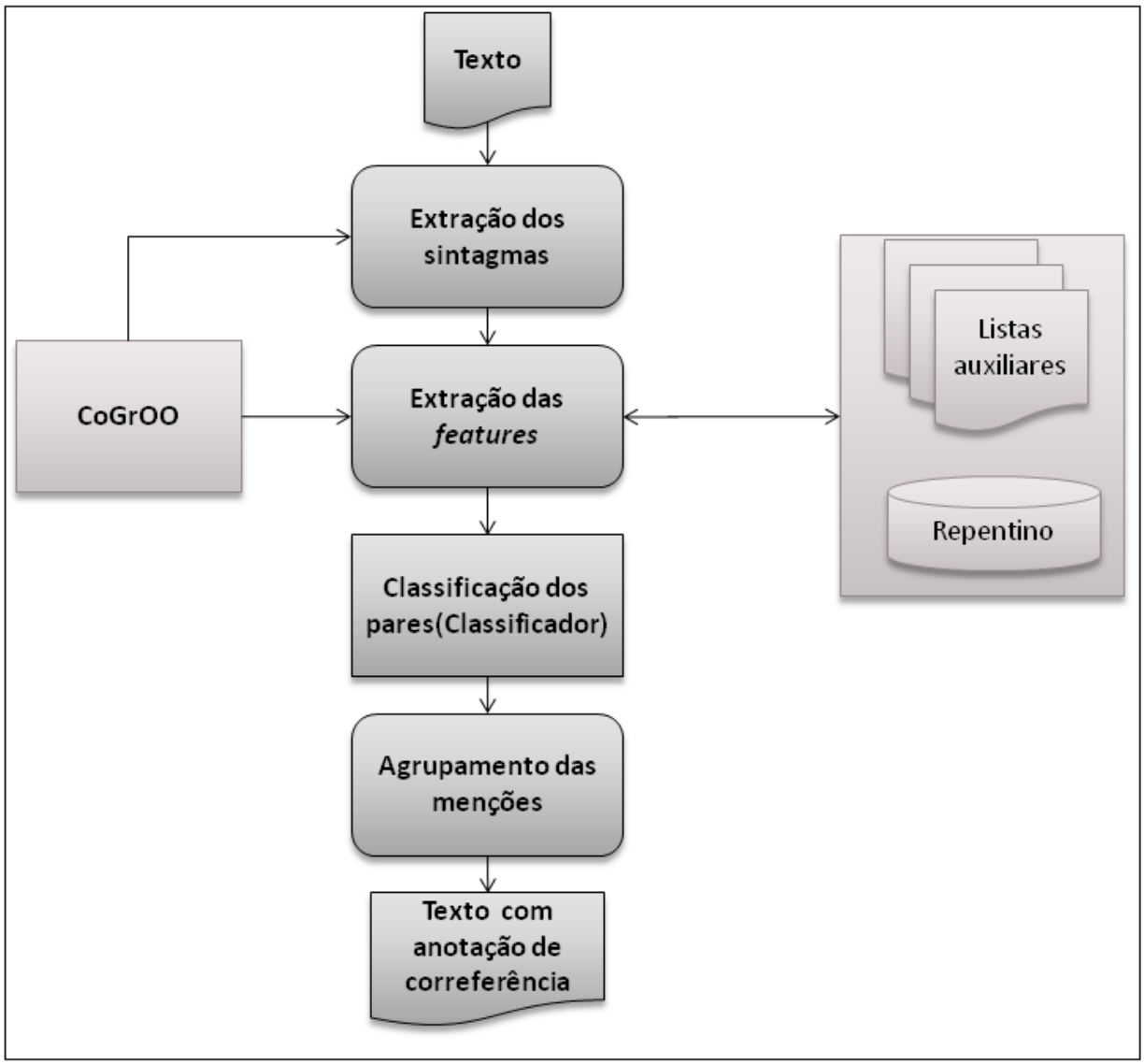

Figura 12: Arquitetura CORP

Dado um texto puro, o sistema efetua a chamada da API CoGrOO [SIL13], extraindo todos os sintagmas nominais do texto. Em seguida, são gerados os pares de sintagmas. A geração dos pares é feita da seguinte forma: Cada sintagma faz par com o próximo, por exemplo: dados os elementos SN1, SN2, SN3 e SN4, os pares formados serão: {SN1-SN2, SN1-SN3, SN1-SN4, SN2-SN3, SN2-SN4, SN3-SN4} (Cada sintagma faz par com o próximo da lista, mas nunca com um sintagma anterior), semelhante ao método utilizado por [COR10] e [STR07]. Extraídos os sintagmas e gerados os pares candidatos à correferência, o sistema faz a extração dos atributos/*features* desses pares, retornando "true" ou "false" para cada atributo (Tabela 2). A descrição de cada *feature* e do recurso utilizado nesta etapa encontra-se no capítulo 5, subseção 5.5.2. Após a extração dos atributos o sistema armazena os dados como vetores de características (Tabela 8).

SN1	SN2	P_StrMatch	Alias	Gênero	Número	C.Sem. Igual	C.Sem. Dif	Dist>5	Dist>10	Dist>15	Corref
Carlos Nobre	Nobre	True	False	True	True	True	False	False	False	False	?

Tabela 8: Vetor de características CORP

Ao fim do processo, uma lista de vetores é submetida à árvore de decisão (classificador), que, por sua vez dá um valor ("true" ou "false") ao campo "coref", de acordo com os valores dos outros atributos. Tendo os pares já classificados, o sistema gera as cadeias de correferência. A técnica consiste em agrupar pares de sintagmas que possuem pelo menos um dos sintagmas em comum. Após o processo, é gerado um arquivo .XML, contendo as anotações de correferência (Figura 13).

```
- <Documento_hub-822.txt>
  - <Texto>
      Euro segue em alta face ao dólar O euro segue a reforçar os ganhos em relação à divisa dos Estados
      Unidos esta segunda-feira, cotando sobre 1,474 dólares, seis cêntimos acima dos máximos da sessão
      anterior. A tendência reflecte as perspectivas do mercado de uma nova descida nas taxas de juro
      daReserva Federal dos EUA com o objectivo de inverter o cenário de abrandamento da economia.
      Por outro lado, números do Fundo Monetário Internacional divulgados na passada sexta-feira
      indicam que a divisa europeia já pesa 26,4% do total das reservas externas em todo mundo no
      terceiro trimestre, mais do que no trimestre anterior e a comparar com 24,4% um ano antes. Na
      passada sexta-feira, o Banco Central Europeu (BCE) fixou a taxa de câmbio oficial a 1,4692 dólares,
      um valor que corresponde a mais 17% face ao câmbio média calculado para 2006, quando a moeda
      única europeia se fixou em 1,2556 dólares, segundo dados do BCE para a média anual.
    </Texto>
  - <Cadeia_ID0>
      <SN ID="13" Cat="EN_LOC" Sentença="1" Sintagma="Estados Unidos"/>
      <SN ID="40" Cat="EN_LOC " Sentença="2" Sintagma="EUA"/>
    </Cadeia_ID0>
  - <Cadeia_ID1>
      <SN ID="84" Cat="EN_ORG" Sentença="4" Sintagma="BCE"/>
      <SN ID="111" Cat="EN_ORG" Sentença="0" Sintagma="BCE"/>
    </Cadeia_ID1>
</Documento_hub-822.txt>
```

Figura 13: Saída CORP

Na Figura 13, podemos ver a saída de um arquivo texto (livre de anotações) proveniente do corpus HAREM [FRE10], submetido como entrada à ferramenta. O CORP estrutura sua saída hierarquicamente da seguinte forma: Primeiramente, é disposta a marcação contendo o nome do documento **<Documento_X></Documento_X>**, onde "X" é o nome do documento em questão. Dentro dessa marcação, temos **<Texto></Texto>**, que contém o texto original. Na anotação **<Cadeia_IDX></Cadeia_IDX>** (onde "X" representa o "ID" correspondente à cadeia), encontramos

as cadeias de correferência geradas pela ferramenta. E, dentro de cada cadeia, para cada sintagma existente nela, temos "**SN ID=**", "**Cat=**", "**Sentença=**" e "**Sintagma=**" – respectivamente, o "ID" do sintagma, sua categoria semântica, sentença do texto em que o sintagma encontra-se e o próprio sintagma.

7. AVALIAÇÃO DO CORP

A resolução de correferências é vista como uma tarefa intermediária, que pode ser utilizada como parte de outros sistemas. Dessa forma, é possível avaliar a tarefa tanto intrínseca como extrinsecamente. A avaliação intrínseca é feita por meio da comparação das cadeias obtidas por um sistema automático com as cadeias anotadas manualmente. Já a avaliação extrínseca é realizada por meio da utilização de outros sistemas, verificando-se qual é a variação no desempenho de um sistema final com a adição do processo de resolução de correferências.

Neste capítulo é apresentado o método de avaliação desenvolvido para o sistema proposto nesta dissertação. O CORP foi avaliado intrinsecamente, utilizando como referência o corpus do Harem [FRE10].

7.1. Corpus de Avaliação

O método utilizado para a avaliação do CORP é baseado na comparação das cadeias de correferência obtidas pelo sistema com as cadeias anotadas manualmente, existentes no corpus do Harem [FRE10]. O corpus do Harem (Capítulo 4, Seção 7) é composto por 129 textos, anotados com informação de correferência, contendo também a categoria de cada entidade nomeada. Na Figura 14 podemos ver como estão dispostas as anotações de correferência.

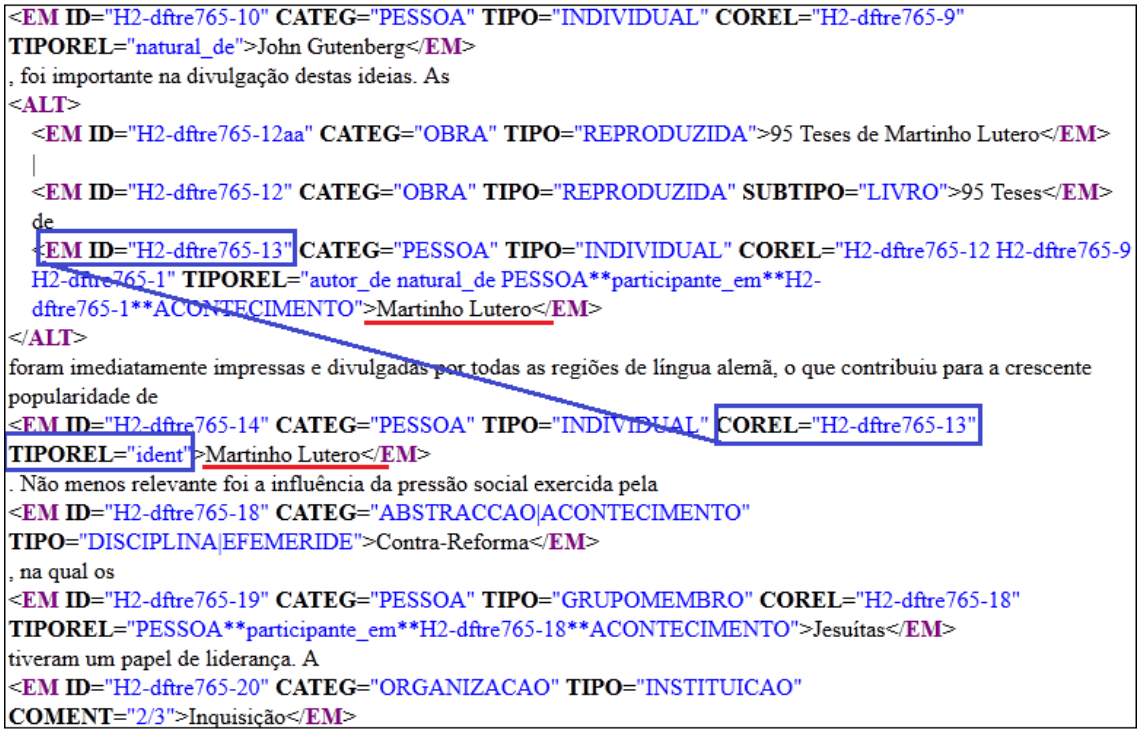

Figura 14: Anotação de correferência no corpus do HAREM [HAR08]

Para facilitar a comparação dos resultados, foi construído um programa utilizando a linguagem Java. O programa basicamente discretiza as informações do corpus anotado e do CORP, de forma que possam ser comparadas. A estratégia de captura deu-se da seguinte forma: cada entidade nomeada está disposta entre a marcação "<EM </EM>", dessa forma foi bastante simples efetuar a captura por meio de expressões regulares. Cada entidade possui um "ID" único. Para formar as cadeias de correferência, utilizou-se a anotação "COREL" (Figura 14). Após a captura e a estruturação das cadeias de correferência, foram aplicados filtros, excluindo das cadeias as informações de correferência cujo tipo de relação fosse diferente de identidade ("TIPOREL=ident") e sua categoria semântica não fosse igual a: Pessoa, Local ou Organização. A avaliação do CORP foi realizada utilizando as métricas descritas na seção 7.2.

7.2. Métrica Utilizada

Os trabalhos que abordam a resolução de correferência concentram seus esforços na avaliação intrínseca, pois é mais fácil de ser realizada e de se obter resultados que sejam comparáveis e reprodutíveis [SIL11]. Essa comparação é realizada por meio de medidas de avaliação que visam quantificar o desempenho de um sistema de resolução de correferências.

Entre as métricas existentes, a mais utilizada na avaliação desses sistemas é a medida proposta por [VIL95]. Essa métrica de desempenho é a mesma utilizada nos trabalhos de [SIL11], [LEE11] e [FER12].

A métrica proposta por [VIL95], a MUC, busca aferir um valor de precisão e cobertura para dada cadeia de correferência, documento ou corpus. Como podemos ver nos exemplos a seguir.

(7.2.1) Marcos passou no vestibular. Ele está muito feliz com a notícia... O estudante mostrou seu conhecimento ficando em primeiro colocado na lista dos classificados.

Para calcular as medidas de precisão propostas por [VIL95], considera-se que um sistema automático obteve como resultado a cadeia "Marcos" e "Ele". Para avaliar o desempenho desse sistema, seu resultado será comparado com a cadeia anotada manualmente ("Marcos", "Ele" e "O estudante"). Considera-se que cada cadeia de correferência é o conjunto das ligações entre as menções e seu antecedente. Para a cadeia automática tem-se {"Marcos" - "Ele"} e para a cadeia de referência temos: {"Marcos" - "Ele", "Ele " - "O estudante"}. O cálculo de precisão e cobertura em nível de cadeia de correferência é dado pelas seguintes fórmulas:

$$Precisão = \frac{N° \ de \ ligações \ corretas}{N° \ de \ ligações \ da \ cadeia \ automática}$$

$$Cobertura = \frac{N° \ de \ ligações \ corretas}{N° \ de \ ligações \ da \ cadeia \ de \ referência}$$

Para o exemplo (7.2.1), temos uma precisão de 100% e cobertura de 50%:

$$Precisão = \frac{1}{1} = 1 \ ou \ 100\%$$

$$Cobertura = \frac{1}{2} = 0.5 \ ou \ 50\%$$

(7.2.2) Para simplificar, vamos pensar nas cadeias de correferência como conjuntos mais

simples, onde cada letra maiúscula é um link. Temos então": Cadeia de referência= {A, B, C, D, E, F} e Cadeia automática= {A, B, C, H}. A precisão e a cobertura são dadas por:

$$Precisão = \frac{3}{4} = 0.75\ ou\ 75\%$$

$$Cobertura = \frac{3}{6} = 0.50\ ou\ 50\%$$

(7.2.3) Cadeia de referência= {A, B, C} e Cadeia automática= {A, B, C, D, E, F}. Temos então:

$$Precisão = \frac{3}{6} = 0.50\ ou\ 50\%$$

$$Cobertura = \frac{3}{3} = 1\ ou\ 100\%$$

(7.2.4) Cadeia de referência= {B, C, D} e Cadeia automática= {A, B, C, D, G, H}:

$$Precisão = \frac{3}{6} = 0.50\ ou\ 50\%$$

$$Cobertura = \frac{3}{3} = 1\ ou\ 100\%$$

(7.2.5) Cadeia de referência= {A} e Cadeia automática= {A, C}:

$$Precisão = \frac{1}{2} = 0.50\ ou\ 50\%$$

$$Cobertura = \frac{1}{1} = 1\ ou\ 100\%$$

(7.2.6) Cadeia de referência= {A, C} e Cadeia automática= {A}:

$$Precisão = \frac{1}{1} = 1\ ou\ 100\%$$

$$Cobertura = \frac{1}{2} = 0.50\ ou\ 50\%$$

Até agora, vimos como calcular a precisão e a cobertura de cada cadeia; porém, geralmente um documento é composto por mais de uma cadeia de correferência. Para calcular a precisão e cobertura das cadeias de correferência em nível de documento temos a seguinte fórmula:

$$Precisão_{Documento} = \frac{\sum_{i=1}^{N°\ Cadeias} (Precisão)}{N°\ Cadeias}$$

$$Cobertura_{Documento} = \frac{\sum_{i=1}^{N°\ Cadeias} (Cobertura)}{N°\ Cadeias}$$

Assumindo que os exemplos (7.2.1), (7.2.2), (7.2.3), (7.2.4), (7.2.5) e (7.2.6) pertençam ao mesmo documento, os cálculos de precisão e cobertura por nível de documento são dados por:

$$Precisão_{Documento} = \frac{1 + 0.75 + 0.5 + 0.5 + 0.5 + 1}{6} = 0.70\ ou\ 70\%$$

$$Cobertura_{Documento} = \frac{0.5 + 0.5 + 1 + 1 + 1 + .05}{6} = 0.75\ ou\ 75\%$$

Podemos também calcular a precisão e cobertura para uma coleção de documentos (corpus). Para tal cálculo, a fórmula segue o mesmo padrão hierárquico:

$$Precisão_{Corpus} = \frac{\sum_{i=1}^{N° \ Documentos} (Precisão_{Documento})}{N° \ Documentos}$$

$$Cobertura_{Corpus} = \frac{\sum_{i=1}^{N° \ Documentos} (Cobertura_{Documento})}{N° \ Documentos}$$

7.3. Validação e Testes

Tendo em "mãos" a métrica e as informações para avaliar o CORP, foram realizados quatro experimentos. Cada experimento foi realizado com um filtro diferente. Como a arquitetura do CORP lhe permite não informar uma categoria a uma entidade caso ele não identifique, os experimentos foram realizados visando explorar o impacto que a categoria semântica pode ter nos resultados. A comparação das cadeias deu-se em nível de *matching* exato. Isto é, dentro de cada cadeia de correferência, foram considerados corretos apenas os sintagmas nominais exatamente iguais aos do corpus de referência.

Experimento 1: O primeiro experimento foi realizado levando em consideração todos os pares de saída do sistema. Isto é, as cadeias foram geradas levando em consideração todos os pares classificados como correferentes. (E1-todos-pares)

Experimento 2: No segundo experimento, para gerar as cadeias de correferência, foram considerados pares em que pelo menos uma categoria semântica tivesse sido aferida. Na Tabela 9, podemos visualizar alguns exemplos de pares considerados e não considerados, em que "EN_?" significa que a categoria semântica não foi aferida. (E2-min-1-PLO)

Categoria semântica Sintagma 1	**Categoria semântica Sintagma 2**	**Par válido?**
EN_ORG	EN_ORG	Sim
EN_ORG	EN_?	Sim
EN_SER	EN_LOC	Sim
EN_?	EN_?	Não
EN_SER	EN_ORG	Sim
EN_?	EN_SER	Sim

Tabela 9: Tipos de pares considerados no experimento 2.

Experimento 3: No terceiro experimento foram considerados apenas pares em que ambas as categorias foram aferidas. (Tabela 10).(E3-2-ent-PLO)

Categoria semântica Sintagma 1	Categoria semântica Sintagma 2	Par válido?
EN_ORG	EN_ORG	Sim
EN_ORG	EN_?	Não
EN_SER	EN_LOC	Sim
EN_?	EN_?	Não
EN_SER	EN_ORG	Sim
EN_?	EN_SER	Não
EN_SER	EN_?	Não

Tabela 10: Tipos de pares considerados no experimento 3.

Experimento 4: Para o quarto experimento utilizou-se os mesmos filtros utilizados pelo terceiro experimento, com a diferença de considerar como par correferente apenas pares em que os sintagmas nominais possuem a mesma classe semântica. Na Tabela 11, podemos visualizar alguns exemplos de pares considerados e não considerados. (E4-PLO-ident)

Categoria semântica Sintagma 1	Categoria semântica Sintagma 2	Par válido?
EN_ORG	EN_ORG	Sim
EN_ORG	EN_?	Não
EN_SER	EN_LOC	Não
EN_?	EN_?	Não
EN_SER	EN_SER	Sim
EN_?	EN_SER	Não
EN_LOC	EN_LOC	Sim
EN_LOC	EN_SER	Não

Tabela 11: Tipos de pares considerados no experimento 4.

Os filtros utilizados no experimento quatro foram os que surtiram melhores efeitos. Na Tabela 12, podemos visualizar os resultados obtidos em cada teste realizado. Os resultados obtidos são provenientes da comparação direta com o corpus de referência [HAR08] e [FRE10]. Como podemos

ver, os resultados são satisfatórios à afirmação de [COR10], para o qual trabalhos que utilizam categoria semântica para a resolução de correferência podem obter melhores resultados. No CORP, embora a diferença não seja tão grande entre um experimento e outro, essa tese foi comprovada.

Experimento	Precisão	Cobertura	Medida-F	Número de cadeias
1	60.32%	67.35%	63.64%	768
2	71.52%	58.34%	64.26%	477
3	74.88%	56.83%	64.62%	415
4	77.97%	59.38%	67.42%	408

Tabela 12: Avaliação CORP

Apesar de não ser possível uma comparação direta com os outros sistemas, devido às restrições do CORP, como trabalhar com apenas nomes próprios e com as categorias do tipo Pessoa, Local e Organização para a língua portuguesa, o CORP obteve bons resultados, visto os principais trabalhos, como o de [SIL11]: para a resolução de correferência em mono documento, utilizando a mesma métrica, o trabalho de [SIL11] obteve uma cobertura de 45.9% e uma precisão de 49.12%. Na Tabela 13, podemos visualizar os resultados dos principais trabalhos relacionados utilizando a métrica MUC.

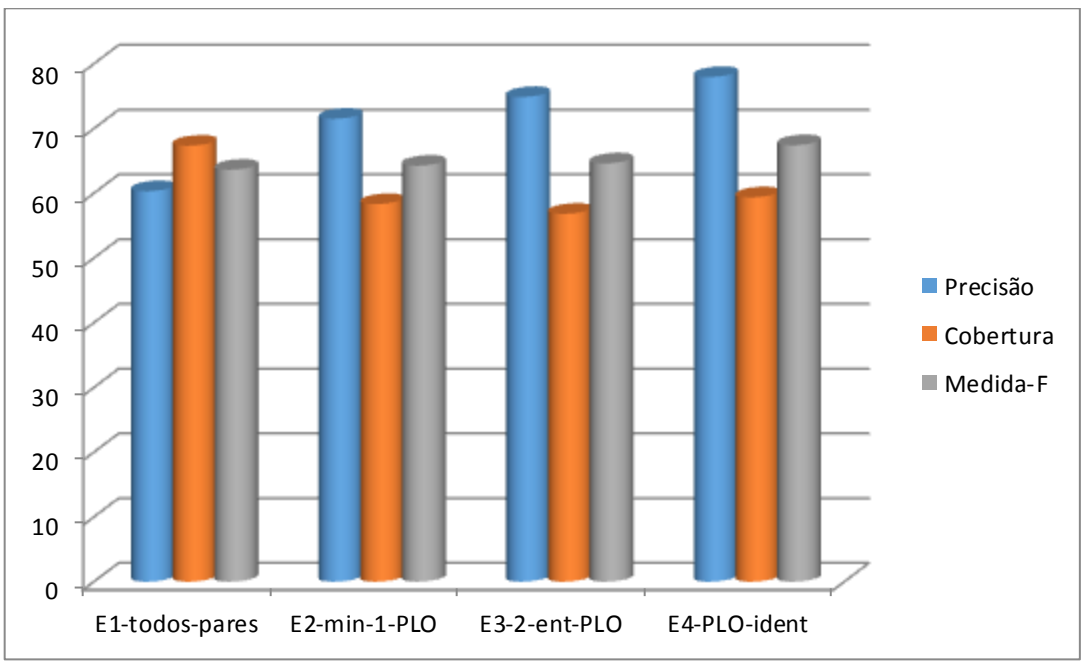

Figura 15: Resultados dos quatro experimentos realizados no CORP.

Sistema	Idioma	Precisão	Cobertura	Medida-F
Silva[SIL11]	Português	49.12%	45.9%	47.45%
CORP E4	Português	77.97.%	59.38%	67.42%
Fernandes[FER12]	Inglês	75.91%	65.83%	70.51%
Fernandes[FER12]	Chinês	70.58%	52.69%	60.34%
Fernandes[FER12]	Árabe	49.69%	43.63%	46.46%
Lee[LEE11]	Inglês	59.3%	62.8%	61.00%

Tabela 13: Resultados não comparativos, baseados na métrica MUC[VIL95].

7.4. Análise de Erros

Esta seção tem por objetivo analisar os erros que acontecem durante o processo de resolução de correferência. Pequenos erros que podem se acumular e degradar a saída do sistema. Por meio dessa análise é possível entendermos melhor os resultados obtidos pelo CORP nas avaliações realizadas e até mesmo projetar melhorias futuras para a ferramenta.

Figura 16: Análise de erros exemplo 1.

Na Figura 16, podemos visualizar um exemplo de cadeia de correferência incorreta. Como a *Parcial_String_Match* é uma *feature* de classificação bastante relevante, às vezes o classificador pode retornar como par correferente os casos em que os sintagmas concordam parcialmente, como podemos notar na cadeia "ID=0", em que: "**Jorge** Wanderley" e "**Jorge** Luiz Borges" possuem em comum a *string* "Jorge". Para esse caso, o sistema assumiu que "Jorge Luiz Borges" deveria formar cadeia com "Jorge Wanderley", e não com a cadeia de "ID=6", "Borges". Ainda na Figura 16, podemos notar outro caso de falso positivo: para "**Carlo** Emilio Gadda" e "Gian **Carlo** Roscioni", a cadeia foi gerada pelo mesmo processo descrito anteriormente, tendo "Carlo" como *string* em comum. Obviamente, não basta a *feature Parcial_String_Match* retornar "true" para tornar um par positivo. Em ambos os casos, as *features* **M_Gênero**, **M_Número** e **Categ_semântica_Igual** retornam o valor "true". Na Figura 17, podemos visualizar os possíveis caminhos utilizados pelo classificador.

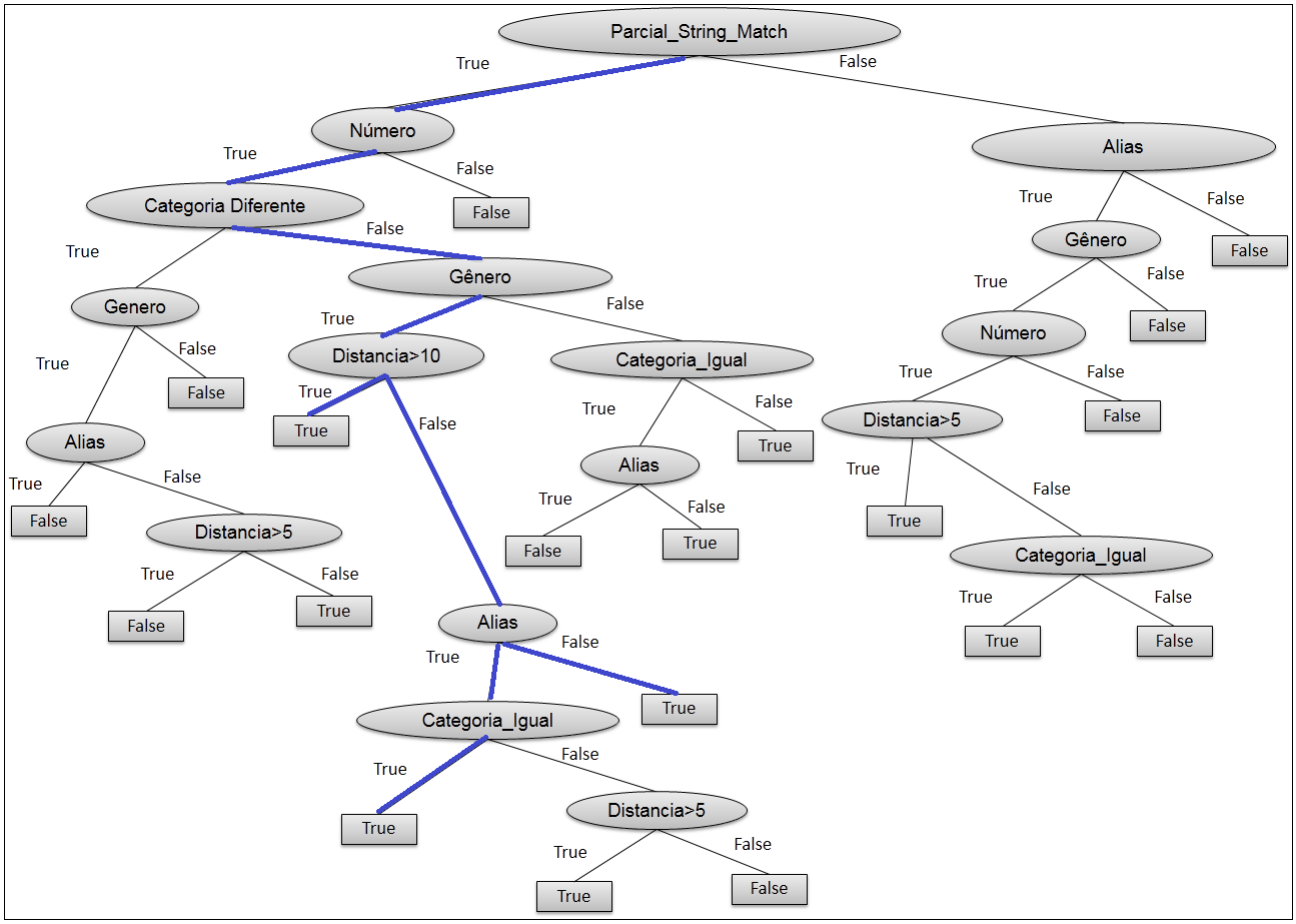

Figura 17: Possíveis caminhos da árvore para o exemplo existente na Figura 16.

Outro erro que ocorre em alguns casos é a classificação incorreta das entidades nomeadas, como podemos visualizar na Figura 18.

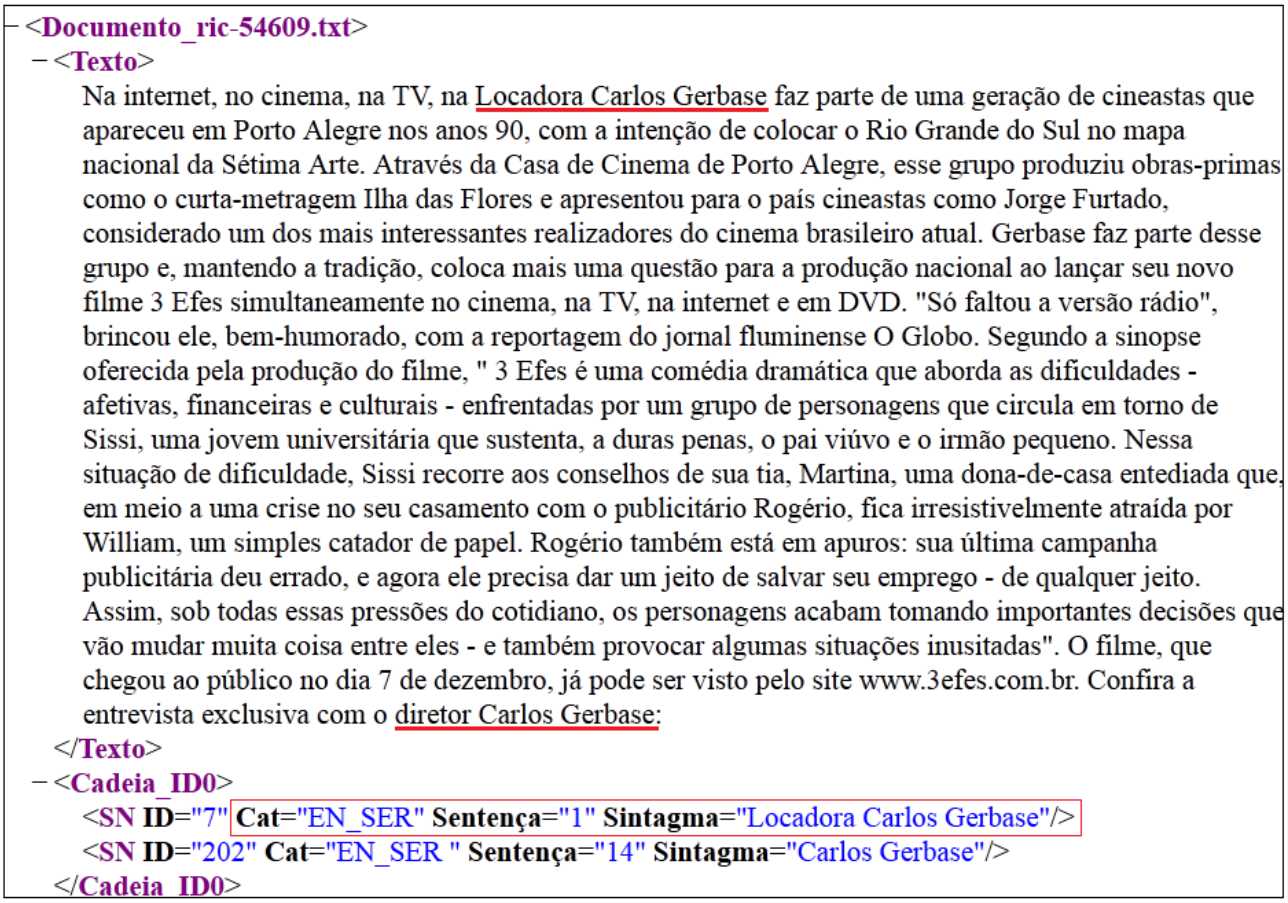

Figura 18: Análise de erros exemplo 2.

O erro por ambiguidade durante o processo de classificação das entidades nomeadas também é um desafio. Como podemos ver na Figura 18, o sistema aferiu a categoria Pessoa (EN_SER) ao sintagma “Locadora Carlos Gerbase”. O problema ocorrido neste caso foi uma análise sintática feita sob um trecho de texto escrito incorretamente. Como podemos notar, a palavra “Locadora” está escrita em letra maiúscula no meio de uma frase, o que nos sugere um nome próprio. No caso de “Locadora” ser uma entidade, seria do tipo Organização, enquanto o sistema classificou-a incorretamente como Pessoa. Nesse caso, temos dois erros: o de análise sintática, causado pela escrita incorreta e o da incorreta classificação da entidade. O erro proveniente dessa classificação pode ser notado, se analisarmos a Figura 19.

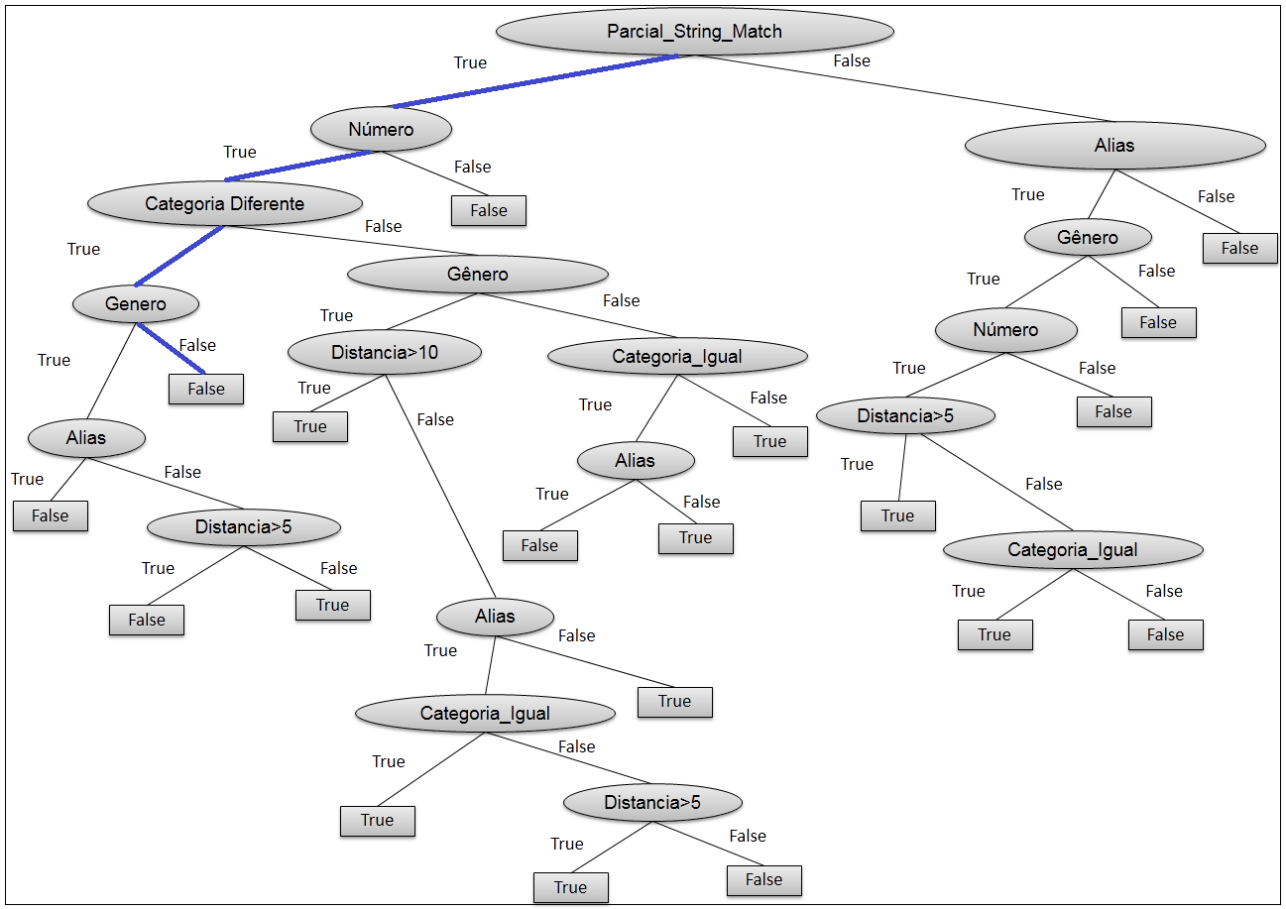

Figura 19: Caminho correto para a classificação do par de sintagma da Figura 17.

Como podemos visualizar na Figura 19, caso a categoria da entidade nomeada “Locadora Carlos Gerbase” tivesse sido aferida corretamente como organização, esse par de sintagmas teria sido classificado como par não correferente (resposta correta), dado que o sintagma “Locadora Carlos Gerbase” possui gênero feminino e o sintagma “Carlos Gerbase” possui gênero masculino. Esse exemplo mostra a importância do conhecimento semântico para a tarefa de resolução de correferências.

Outro erro constatado no processo de resolução de correferências do CORP encontra-se na geração das cadeias de correferência, como podemos ver na Figura 20.

```
−<Cadeia_ID43>
    <SN ID="96" Cat="EN_ORG" Sentença="5" Sintagma="Delta do Mississipi"/>
    <SN ID="113" Cat="EN_ORG" Sentença="5" Sintagma="Delta"/>
  </Cadeia_ID43>
−<Cadeia_ID130>
    <SN ID="13" Cat="EN_SER" Sentença="1" Sintagma="Jeff Mills"/>
    <SN ID="342" Cat="EN_SER" Sentença="18" Sintagma="Jeff Mills"/>
    <SN ID="347" Cat="EN_SER" Sentença="18" Sintagma="DJ"/>
    <SN ID="383" Cat="EN_SER" Sentença="1" Sintagma="Mills"/>
    <SN ID="444" Cat="EN_SER" Sentença="18" Sintagma="Mills"/>
    <SN ID="446" Cat="EN_SER" Sentença="21" Sintagma="DJ"/>
    <SN ID="480" Cat="EN_SER" Sentença="19" Sintagma="Mills"/>
    <SN ID="492" Cat="EN_SER" Sentença="21" Sintagma="Mills"/>
    <SN ID="554" Cat="EN_SER" Sentença="22" Sintagma="Mills"/>
    <SN ID="609" Cat="EN_SER" Sentença="28" Sintagma="Mills"/>
    <SN ID="620" Cat="EN_SER" Sentença="32" Sintagma="DJ"/>
    <SN ID="653" Cat="EN_SER" Sentença="31" Sintagma="Mills"/>
    <SN ID="758" Cat="EN_SER" Sentença="34" Sintagma="Mills"/>
    <SN ID="18" Cat="EN_LOC" Sentença="1" Sintagma="Detroit"/>
    <SN ID="32" Cat="EN_LOC" Sentença="2" Sintagma="Detroit"/>
    <SN ID="56" Cat="EN_LOC" Sentença="3" Sintagma="Uma Cidade Dificil Detroit"/>
    <SN ID="144" Cat="EN_LOC" Sentença="3" Sintagma="Detroit"/>
    <SN ID="199" Cat="EN_LOC" Sentença="8" Sintagma="Detroit"/>
    <SN ID="240" Cat="EN_LOC" Sentença="10" Sintagma="Detroit"/>
    <SN ID="300" Cat="EN_LOC" Sentença="12" Sintagma="Detroit"/>
    <SN ID="660" Cat="EN_LOC" Sentença="16" Sintagma="Detroit"/>
    <SN ID="707" Cat="EN_LOC" Sentença="34" Sintagma="Detroit"/>
    <SN ID="777" Cat="EN_LOC" Sentença="37" Sintagma="Detroit"/>
    <SN ID="137" Cat="EN_LOC" Sentença="7" Sintagma="Cidade Motor"/>
    <SN ID="169" Cat="EN_ORG" Sentença="8" Sintagma="General Motors"/>
    <SN ID="350" Cat="EN_LOC " Sentença="18" Sintagma="Durban"/>
  </Cadeia_ID130>
```

Figura 20: Cadeia gerada incorretamente (arquivo 2ght33.txt [HAR08])

Na Figura 20, podemos notar que a cadeia “ID=130” possui claramente duas cadeias de correferência distintas agrupadas incorretamente, sendo elas: “Detroit” e “Jeff Mills”. Analisando a saída do sistema, constatou-se que o CORP estava agrupando ambas as cadeias em uma só, por meio do sintagma “DJ”. Isto é, as cadeias foram agrupadas pelo fato de o sistema ter classificado como correferente o par cujos sintagmas são “Detroit” e “Dj”. Como o sintagma “Dj” é uma correferência de “Jeff Mills”, o sistema juntou ambas as cadeias. Por meio do quarto experimento, foi possível eliminar esse tipo de problema, tendo em vista que, para determinado par ser considerado correferente, ambos os sintagmas devem possuir a mesma categoria semântica. Nesse caso, “Detroit” é um local e “Dj” é uma profissão, isto é, como são menções a classes distintas de entidades, respectivamente, Local e Pessoa, o par pode ser excluído sem problemas,

pois com certeza é um falso positivo. Os resultados do quarto experimento provam essa teoria, pois além de melhorar a precisão do sistema, os filtros aplicados melhoraram também sua cobertura.

Como a validação do sistema ocorre em nível de cadeias, o fato de duas cadeias serem consideradas apenas uma degrada consideravelmente a precisão e a cobertura do sistema, dado o fato que uma cadeia terá uma precisão bastante baixa pelo fato de existirem muitos elementos incorretos, ao mesmo tempo em que a outra cadeia não existirá, diminuindo a cobertura da ferramenta.

Na Figura 21, temos a saída do mesmo documento, que utiliza as métricas propostas pelo quarto experimento (Tabela 12).

Figura 21: Cadeia gerada utilizando os filtros do quarto experimento.

8. CONSIDERAÇÕES FINAIS

Para conduzir esta pesquisa, foi realizado um estudo da base teórica a respeito de resolução de correferências e das iniciativas propondo métodos e protótipos. Em seguida, foram identificadas técnicas a serem utilizadas para concepção do modelo. A partir desses estudos, foi proposto um método automático de resolução de correferências a partir de documentos de texto puro (arquivos livres de anotação), avaliado por meios bem consolidados na área de Processamento da Linguagem Natural.

Este trabalho tratou apenas nomes próprios, do tipo pessoa, local e organização. Apesar desse escopo limitado, ele pode servir de base para um modelo mais abrangente, que trate correferências para todas as categorias de entidades nomeadas, incluindo outros sintagmas nominais e pronomes.

8.1. Publicações

Os artigos relacionados a seguir foram produzidos e publicados durante o período do mestrado e seu conteúdo se refere ao trabalho conduzido:

- Geração de features para resolução de correferência: Pessoa, Local e Organização [FON13a]. Publicado no 9th Brazilian Symposium in Information and Human Language Technology, 2013 (STIL2013). Este artigo apresenta resultados parciais do trabalho realizado no decorrer do mestrado, relatando um estudo sobre a importância da correta seleção e implementação de *features*, utilizadas por um sistema de resolução de correferências baseado em aprendizado de máquina.

- Resolução de Correferência em Língua Portuguesa: Pessoa, Local e Organização [FON13b]. Publicado no X National Meeting on Artificial and Computational Intelligence, 2013 (ENIAC2013). Este artigo relata o andamento do trabalho desta dissertação até o final do terceiro semestre, apresentando experimentos realizados que culminaram na escolha do classificador utilizado pelo CORP.

8.2. Contribuições deste Estudo

Além das contribuições científicas anteriores a esta dissertação [FON13a] e [FON13b], a principal contribuição deste trabalho, além do modelo proposto, está na apresentação de um sistema de resolução de correferências para a língua portuguesa livre e funcional. Por meio de revisão bibliográfica, constatou-se que até o momento não existia um sistema de resolução de correferências com essas características. Quando falamos de um sistema funcional, nos referimos a um sistema que, dado um arquivo texto puro, livre de anotações, resolva correferências desse texto, tendo como saída as anotações de correferência dessa entrada. Além disso, para a concepção do CORP, optou-se pela utilização de apenas recursos livres. Dessa forma foi possível obter uma ferramenta *open source* que pode ser distribuída e modificada por todo meio acadêmico, sem qualquer restrição, comumente existente quando utilizamos *softwares* proprietários [BIC00].

8.3. Desafios

Um dos principais desafios deste trabalho ocorre na geração dos pares candidatos à correferência. Como a ideia é gerar um modelo de resolução de correferências baseado em aprendizado de máquina, a primeira etapa deste trabalho é treinar um modelo de classificação, o qual recebe pares de sintagmas, e, por meio de suas características, consiga prever se esses pares são correferentes ou não. A primeira tarefa é algo relativamente simples, pois as informações de correferência estão contidas no corpus de treinamento, bastando organizá-las e submetê-las ao aprendizado.

Gerado o modelo, o sistema deve ser capaz de, por meio de inúmeros pares de sintagmas positivos e negativos, prever quais pares são correferentes e quais não são. O interesse nesse ponto é saber quais pares são, de fato, correferentes. Como a quantidade de pares negativos é muito maior, é natural (e, também, um grande problema) haver pares negativos classificados como positivos e positivos classificados como negativos. Isso ocorre pelo fato de não haver (com exceção da *Parcial_String_Match*) *features* que sejam tão determinantes a ponto de classificar determinado par de SN com uma alta taxa precisão. Um meio de contornar esse problema de classificação incorreta foi limitar a abrangência do sistema, levando em consideração apenas pares que possuam nomes próprios em ambos os sintagmas. Outro desafio dessa produção está na limitação de recursos. Além do fato de a língua portuguesa possuir poucos recursos, o presente trabalho optou pela utilização de recursos *open source* apenas.

8.4. Limitações

Assim como qualquer recurso computacional, o CORP possui suas limitações. Uma delas ocorre no tamanho do corpus utilizado para o treinamento do classificador. Conforme vimos anteriormente, o corpus Summ-it é formado por 50 (cinquenta) textos jornalísticos. Além disso, para o treinamento do modelo foi necessário balancear a quantidade de pares, por meio de replicação dos pares positivos. Esse tipo de técnica é bastante utilizada, porém pode tornar o modelo tendencioso na classificação.

Outra limitação se dá na geração dos pares candidatos à correferência. Como visto na seção 8.3, a quantidade de pares negativos por documento sempre será maior que a quantidade de pares positivos. Esse tipo de problema limita muito a obtenção de bons resultados. Baseando-se nessa limitação, foi tomada a decisão de trabalhar apenas com nomes próprios em ambos os sintagmas. Isso reduziu a abrangência do sistema, mas consequentemente aumentou a qualidade dos resultados obtidos, tornando-os no mínimo satisfatórios.

8.5. Trabalhos Futuros

Como trabalhos futuros, espera-se aumentar a abrangência do CORP, resolvendo anáforas e correferência pronominal. Para isso, é preciso encontrar meios mais eficazes para a criação de pares candidatos a correferência, bem como encontrar uma ferramenta de reconhecimento e classificação de entidades nomeadas que permita ser embarcada ao sistema, de forma que sua execução seja imperceptível aos olhos do usuário.

REFERÊNCIAS

[ABR05] Abreu S., **Análise de Expressões Referenciais em Corpus Anotado da Língua Portuguesa**, dissertação de mestrado, Universidade do Vale do Rio dos Sinos, 2005.

[ABR13] Abreu S. C., Bonamigo T. L. and Vieira R., **A review on Relation Extraction with an eye on Portuguese**, Pages 1-19 In: Journal of the Brazilian Computer Society, 2013.

[AMA13] Amaral D., **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.

[BIC00] Bick E., **The Parsing System "Palavras" - automatic grammatical analysis of portuguese in a constraint grammar framework**, Tese de Doutorado, Department of Linguistics, University of Århus, DK, 2000.

[BOU13] Boukckaert R., Frank E., Hall M., Kirkby K., Reutemann P., Seewald A. and Scuse D., **Weka Manual for version 3.6.9**, The University of Waikato, 2013.

[BRU08] Bruckschen M., Muniz F., Souza J., Fuchs J., Infante K., Muniz M., Gonçalves. P., Vieira R., Aluísio S., **Anotação Linguística em XML do Corpus PLN-BR**, USP, 2008.

[CAR08] Cardoso N., **REMBRANDT – Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto**, chapter 11, pages 195-211. Linguateca, 1 edition. ISBN 9789892016566, 2008.

[COL07] Collovini S., Carbonel T., Fuchs J., Coelho J., Rino L., Vieira R., **Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática.** In: V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXVII Congresso da SBC, Rio de Janeiro, 2007.

[CON11] CoNLL2011, **Conference on computational natural language learning**, Disponível em: http://conll.cemantix.org/2011/. Acesso em: 05/08/2012.

[COR10] Coreixas T., **Resolução de Correferência e Categorias de Entidades Nomeadas**, Dissertação de Mestrado, Pontifícia Universidade Católica Do Rio Grande Do Sul, 2010.

[FER12] Fernandes E., Santos C., Milidiú R., **Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution**, Conference on computational natural language learning, 2012.

[FRE10] Freitas C., Mota C., Santos D., Oliveira H., Carvalho P., **Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese**, Linguateca, FCCN, 2010.

[FRE13] Freeling, **An Open Source Suite of Language Analyzers**, Disponível em: http://nlp.lsi.upc.edu/freeling/, Acesso em: 07/10/2013

[FON13a] Fonseca E. B., Vieira R., Vanin A., Geração **de Features para Resolução de Correferência: Pessoa, Local e Organização**, The 9th Brazilian Symposium in Information and Human Language Technology, 2013.

[FON13b] Fonseca E. B., Vieira R., Vanin A., **Resolução de Correferência em Língua Portuguesa: Pessoa, Local e Organização**, X National Meeting on Artificial and Computational Intelligence, 2013.

[LEE11] 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, Conference on computational natural language learning, 2011.

[GAB11] Gabbard R., Freedman M., Weischedel R. M., **Coreference for Learning to Extract Relations: Yes, Virginia, Coreference Matters**. In: Proceedings 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pages 288–293, Portland, Oregon, 2011.

[HAR08] HAREM ,**Reconhecimento de entidades mencionadas em português**. Disponível em: http://www.linguateca.pt/HAREM/, 10/10 2013.

[LTA13] Language Tasks, **Reconhecedor de Entidades Nomeadas**, Disponível em: http://ltasks.com/documentacao/ , Acesso em: 07/09/2013

[LXT12] LX-Tagger, **Language Resources and Technology for Portuguese University of Lisbon**, NLX-Natural Language and Speech Group Disponível em: http://lxcenter.di.fc.ul.pt/tools/en/LXTaggerEN.html, Acesso em: 05/12/2012.

[MAI08] Maia L. C. G., **Uso de Sintagams Nominais Na Classificação Automática De Documentos Eletrônicos**, Tese de Doutorado, Universidade Federal de Minas Gerais. Escola de Ciência da Informação. Departamento Organização e Uso da Informação, 2008.

[MAN99] Manning C. D., Schütze H. , **Foundations of Statistical Natural Language Processing**, The MIT Press Cambridge, Massachussetts London, England, computational linguistics-Statistical methods. ISBN 0-262-13360-l, Page 299, 1999.

[MAZ08] Maziero E. G., Pardo T. A. S., Felipo A. D. e Silva B. C., **A Base de Dados Lexical e a Interface Web do TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil**. In Workshop em Tecnologia da Informação e da Linguagem Humana – TIL, Vilha Velha – Es, 2008.

[MIL95] Miller G., WordNet: A Lexical Database for English. In: Communications of the ACM Vol. 38, No. 11: 39-41, 1995.

[PRA11] Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R., Xue N., **Modeling Unrestricted Coreference in OntoNotes**, CoNLL Shared Task, 2011.

[REP05] Repentino, **REPositório para reconhecimento de ENtidades Nomeadas**, Disponivel em: http://labclup.letras.up.pt/repentino/faq.html, 2005, Acesso em: 10/03/2013.

[SAN11] Santos C., Carvalho D., **Rule and Tree Ensembles for Unrestricted Coreference Resolution**, 15th Conference on Computational Natural Language Learning, 2011.

[SIL11] Silva J., **Resolução de Correferência em Múltiplos Documentos Utilizando Aprendizado Não Supervisionado**, Dissertação de Mestrado, USP, 2011.

[SIL13] Silva W. D. C. M., **Aprimorando o Corretor Gramatical CoGrOO**, Dissertação de Mestrado, Instituto de Matemática e Estatística da Universidade de São Paulo, 2013.

[SOO01] Soon W., NG H. and Lim D. ,**A Machine Learning Approach to Coreference Resolution of Noun Phrases**. Computational Linguistics, 2001.

[STR07] Strube M., Ponzetto S. P., **Knowledge derived from Wikipedia for computing semantic relatedness**, In: Journal of Articial Intelligence Research, v. 30, number 1, pp. 181-212, 2007.

[VIE01] Vieira R., Strube V., **Lingüística computacional: princípios e aplicações.** Pontifícia Universidade Católica do Rio Grande do Sul e Centro de Ciências da Comunicação, Centro de Ciências Exatas e Tecnológicas, Universidade do Vale do Rio dos Sinos, 2001.

[VIE08] Vieira R., Gonçalves P. N. , Souza J. G. C., **Processamento computacional de anáfora e Correferência**, Revista de Estudos da Linguagem, Belo Horizonte, v. 16, n. 1, p. 263-284, jan./jun 2008.

[VIL95] 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 – MUC6 ’95, page45, Morristown, NJ USA, Association for Computational Linguistics. ISBN 1558604022. doi: 10.3115/1072399.1072405, 1995.

[WIK13] WIKPEDIA, **A Enciclopédia Livre,** Disponivel em: http://pt.wikipedia.org/wiki/, Acesso em: 15/03/2013.