PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

FACULDADE DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RESOLUÇÃO DE CORREFERÊNCIA E CATEGORIAS DE ENTIDADES NOMEADAS

TATIANE COREIXAS DE MORAES

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

Orientador: Profa. Dra. Renata Vieira

Porto Alegre

2010

TERMO DE APRESENTAÇÃO

FICHA CATALOGRÁFICA

M827r Moraes, Tatiane Coreixas
Resolução de correferência e categorias de entidades nomeadas / Tatiane Coreixas de Moraes. – Porto Alegre, 2010.
75 f.

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

DEDICATÓRIA

Dedico este trabalho:
Ao meu esposo, Edimar, amor da minha vida.
Aos meus pais (*in memorian*), Ceni e Franklin, por me ensinarem a lutar pelos meus sonhos.

"É melhor tentar e falhar,
que preocupar-se e ver a vida passar;
é melhor tentar, ainda que em vão,
que sentar-se fazendo nada até o final.
Eu prefiro na chuva caminhar,
que em dias tristes em casa me esconder.
Prefiro ser feliz, embora louco,
que em conformidade viver ..."

(Martin Luther King)

AGRADECIMENTOS

Agradeço a Nossa Senhora e ao meu anjinho da guarda, por iluminarem meu caminho.

Ao meu esposo, Edimar, pela dedicação, compreensão, apoio e, principalmente, pelo amor e carinho durante todo o tempo. Se não fosse por seu companheirismo, não estaria onde estou. Você é e sempre será alguém muito especial para mim.

Aos meus pais, por estarem ao meu lado de forma muito especial e por me mostrarem que sempre devemos batalhar por aquilo que acreditamos e que nunca é tarde para lutar por nossos sonhos.

À minha orientadora e amiga, Profa. Renata Vieira, pelos conselhos, incentivos e apoio para a realização deste trabalho.

Aos meus familiares, pela torcida que tudo desse certo.

Aos meus amigos de hoje e de sempre, por entenderem a minha ausência nos encontros de grupo para ficar estudando.

Aos amigos conquistados no grupo de PLN, por tudo. Para Anderson, Larissa, Clarissa, Mírian, Douglas e Patrícia, obrigada pelo ombro amigo e os ouvidos nos momentos de desabafo, como também pelos momentos de gargalhadas.

À minha colega e amiga de faculdade e pós, Ana, pelo apoio antes e depois de entrar no mestrado. Seus conselhos foram essenciais e, por isso, estou aqui hoje.

Aos colegas da UFSCar e USP pelas dicas e pelo convívio durante o período do sanduíche. Muito obrigado, com carinho, a minha amiga e companheira no sanduíche, Larissa: a estadia lá não seria a mesma coisa sem a sua companhia.

Aos professores do pós, por compartilharem seus conhecimentos e experiências.

A CAPES pelo apoio durante o período de realização deste trabalho.

RESOLUÇÃO DE CORREFERÊNCIA E CATEGORIAS DE ENTIDADES NOMEADAS

RESUMO

Define-se correferência como a relação entre diversos componentes linguísticos com uma mesma entidade de mundo. A resolução automática de correferência textual está inserida num contexto muito importante na área de Processamento da Linguagem Natural, pois vários sistemas necessitam dessa tarefa. O nível de processamento linguístico depende do conhecimento de mundo, e isso ainda é um desafio para a área. Esse desafio estimulou e tornou-se o objeto de estudo desta dissertação. Nesse sentido, analisamos o papel das categorias de entidades nomeadas e, através de aprendizado de máquina, verificamos as condições de resolução em diferentes categorias. Os resultados dos experimentos demonstraram que o conhecimento de mundo, representado nas categorias de entidades nomeadas, auxilia nessa tarefa, pois o percentual de retorno do sistema com base nas categorias teve uma melhora de 17% em comparação com a versão sem as categorias.

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

COREFERENCE RESOLUTION AND CATEGORIES OF NAMED ENTITIES

ABSTRACT

Coreference is defined as the relationship of linguistic expressions with one same entity of the world. Automatic coreference resolution is inserted in a very important context in the area of Natural Language Processing, because many systems require this task. This level of language processing depends on world knowledge, and this is still a challenge for the area. This challenge has stimulated and became the subject of this dissertation. Accordingly, we analyzed the role of categories of named entities and, through machine learning, we checked the conditions for resolution of different categories.

The results of the experiments showed that world knowledge, represented by categories of named entities, helps in this task, since the percentage of return of the system based on the categories improved in about 17% when compared to the version without the categories.

Keywords: Coreference resolution, natural language processing, named entities, machine learning.

LISTA DE FIGURAS

Figura 1 – Fragmento de texto de exemplo do *corpus*..... 17
Figura 2 – Exemplo de texto explicativo sobre anáfora. 18
Figura 3 – Árvore sintática gerada para a frase: A opinião é do agrônomo Miguel Guerra, da UFSC. 21
Figura 4 – Trechos de marcação de correferência do HAREM. 25
Figura 5 – Árvore de decisão induzida pelo algoritmo J48. 32
Figura 6 – Exemplo de cadeia do tipo Outro. 41
Figura 7 – Exemplo de cadeia do tipo Pessoa. 41
Figura 8 – Exemplo de marcação incorreta de nome próprio. 42
Figura 9 – Exemplo de cadeia com etiquetas semânticas de outras categorias. 42
Figura 10 – Trecho de um arquivo ARFF. 44
Figura 11 – Visão geral do sistema. 46
Figura 12 – Trecho do arquivo de *markables.* 48
Figura 13 – Exemplo de cadeia de correferência. 48
Figura 14 – Exemplo de sintagmas para a geração dos pares negativos. 49
Figura 15 – Trecho do arquivo *phrases*..... 50
Figura 16 – Trecho do arquivo de *pos*..... 51
Figura 17 – Trecho do arquivo de *tokens.* 52
Figura 18 – Trecho do arquivo de entrada para a API do Weka. 53
Figura 19 – Visão descritiva do protótipo. 54
Figura 20 – Árvore de decisão gerada pelo algoritmo J48 para os testes. 56
Figura 21 – Exemplo de cadeia de Acontecimento com etiquetas semânticas de outras categorias. 59
Figura 22 – Exemplo de cadeia com marcação de várias categorias. 62
Figura 23 – Saída do Weka para o experimento de teste do *baseline*. 63
Figura 24 – Saída do Weka para o experimento utilizando a categoria Pessoa. 64
Figura 25 – Saída do Weka para o experimento utilizando a categoria Acontecimento. 65
Figura 26 – Saída do Weka para o experimento utilizando a categoria Local. 66
Figura 27 – Saída do Weka para o experimento utilizando a categoria Organização. 67
Figura 28 – Saída do Weka para o experimento utilizando o conjunto de categorias escolhidas. 68

LISTA DE TABELAS

Tabela 1 – Marcação de correferência do ACE. 22

Tabela 2 – Tabela dos resultados do experimento de Strube e Ponzetto [STR07]. 30

Tabela 3 – Conjunto de *keywords* por categoria do ACE. 34

Tabela 4 – Etiquetas semânticas da representação de seres humanos. 37

Tabela 5 – Grupo de etiquetas semânticas para cada categoria do HAREM. 38

Tabela 6 – Total de cadeias por categoria do conjunto de textos do *corpus.* 40

Tabela 7 – Resultados dos dados de teste do *baseline*. 58

Tabela 8 – Resultados dos dados de teste da categoria Pessoa. 58

Tabela 9 – Resultados dos dados de teste da categoria Local. 58

Tabela 10 – Resultados dos dados de teste da categoria Acontecimento. 59

Tabela 11 – Resultados dos dados de teste da categoria Organização. 59

Tabela 12 – Resultados do conjunto de teste de Pessoa, Local, Acontecimento e Organização. 60

LISTA DE SIGLAS

ACE – *Automatic Content Extraction*

ACROPOS - *Automatic Coreference ResOlution system for POrtugueSe*

API – *Application Programming Interface*

ARFF – *Attribute-Relation File Format*

EN – Entidade Nomeada

EM – Entidade Mencionada

ER – Expressão Referencial

HAREM – Avaliação de Reconhecedores de Entidades Mencionadas

PLN – Processamento da Linguagem Natural

MMAX – *Multi-Modal Annotation in XML*

MUC – *Message Understanding Conference*

Recorcaten – REsolução de COrreferência por CATegorias de Entidades Nomeadas

SN – Sintagma Nominal

XCES – *Corpus Encoding Standard for XML*

XML – *eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO 14

1.1. Motivação 14

1.2. Objetivo do trabalho 15

1.2.1. Objetivo geral 15

1.2.2. Objetivos específicos 15

1.3. Hipótese 16

1.4. Organização da dissertação 16

2. FUNDAMENTAÇÃO TEÓRICA 17

2.1. Resolução de correferência 17

2.2. Categorias de entidades nomeadas 19

2.2.1. Avaliações conjuntas 19

2.2.1.1. Avaliação ACE 19

2.2.1.2. Avaliação HAREM 23

2.3. Considerações sobre este capítulo 26

3. TRABALHOS RELACIONADOS 27

3.1. O trabalho de Soon et al. 27

3.2. O trabalho de Strube e Ponzetto 28

3.3. O trabalho de Souza 31

3.4. Os trabalhos de Ng 32

3.5. Considerações sobre este capítulo 35

4. RECURSOS UTILIZADOS 36

4.1. Analisador sintático PALAVRAS 36

4.2. Summ-it 39

4.2.1. Estudo do *corpus* de acordo com as categorias do HAREM 40

4.3. Weka 43

4.4. Considerações sobre este capítulo 45

5. O PROTÓTIPO DESENVOLVIDO 46

***5.1.* Características do *Baseline* 46**

5.2. Características do sistema de correferência Recorcaten 54

5.3. Considerações sobre este capítulo 55

6. ANÁLISE DOS RESULTADOS 56

6.1. Considerações sobre este capítulo 61

7. CONSIDERAÇÕES FINAIS 69

7.1. Contribuições 70

7.2. Limitações 70

7.3. Trabalhos futuros 70

REFERÊNCIAS BIBLIOGRÁFICAS 72

ANEXO A - TABELA DE *FEATURES* DE NG E CARDIE[NG02] 74

APÊNDICE A - TABELA DAS CATEGORIAS POR CADEIA POR TEXTO 75

1. INTRODUÇÃO

1.1. Motivação

A área de Processamento da Linguagem Natural (PLN) está se destacando nos últimos anos devido à necessidade de otimizar a grande quantidade de informação disponível na internet. Com isso, várias aplicações estão ganhando destaque, como a sumarização automática e a extração de informação.

Uma tarefa envolvida nessas aplicações é a resolução de correferência. Esta tarefa diz respeito ao processo de identificar automaticamente e corretamente as expressões correferentes, que são expressões que se referem a uma mesma entidade de mundo [SOO01]. Por necessitar do conhecimento de mundo, essa tarefa ainda apresenta grandes desafios para a área. Em se tratando da língua portuguesa, a disponibilidade de recursos é ainda muito reduzida.

Em trabalhos que abordam a língua inglesa, nota-se o uso de um conjunto de categorias de entidades, em especial, nas avaliações conjuntas da área. Uma definição para conjunto de categorias é que especificam grupos que apresentam as mesmas características. Acredita-se que para o português também se conseguirá melhores resultados para a resolução de correferência daqueles existentes hoje se for considerada a observação de tipos de categorias.

Uma avaliação conjunta é o ACE [ACE08a], a qual define que suas tarefas abranjam expressões linguísticas formadas por nomes próprios, bem como sua retomada por nomes comuns ou grupos nominais (por exemplo, "o candidato democrata"), além de expressões temporais (horas e datas) e expressões numéricas (percentuais e monetárias). No ACE essas expressões são denominadas "*Named Entities*" [ACE08a]. Esta avaliação trata de textos da língua inglesa.

Já a avaliação conjunta para o português, o HAREM [HAR08], foca em expressões linguísticas referenciadas num texto apenas por um nome próprio (como nomes de Pessoas, Organizações, Lugares), e também expressões temporais e numéricas.

No HAREM essas expressões são denominadas “Entidades Mencionadas” [SAN07]. Neste trabalho, será adotada a proposta do ACE, que inclui na análise os

substantivos comuns e sintagmas nominais relacionados aos nomes próprios identificados e, por isso, se fará uso da nomenclatura Entidades Nomeadas - ENs (*Named Entities*).

No que se refere às definições das categorias de ENs, este trabalho seguirá as definições propostas pela avaliação HAREM.

Com base nesse cenário, parte-se da hipótese 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.

A informação semântica é dada no presente trabalho através da análise automática feita pelo analisador sintático PALAVRAS [BIC00].

1.2. Objetivo do trabalho

1.2.1. Objetivo geral

O objetivo deste trabalho é propor e avaliar métodos para a resolução de correferência para a língua portuguesa, com foco nas categorias de ENs e utilizando informações semânticas fornecidas pelo analisador sintático PALAVRAS relativas às categorias de ENs.

1.2.2. Objetivos específicos

No que se refere aos objetivos específicos, pode-se listar:

- Levantamento das *features* propostas por outros trabalhos, tanto para a língua inglesa como para a portuguesa;
- Seleção de um conjunto de *features* de acordo com o estudo realizado;
- Modelagem e desenvolvimento de um protótipo;
- Realização de experimentos;
- Comparação da resolução considerando o conjunto completo de expressões com o conjunto limitado por categorias;
- Avaliação dos resultados.

1.3. Hipótese

Este trabalho considera como hipótese que o uso de categorias de ENs pode auxiliar na tarefa de resolução de correferência, uma vez que categorias de ENs são representadas por um conjunto mais coeso de tipos semânticos.

1.4. Organização da dissertação

A presente dissertação está estruturada da seguinte forma: o capítulo 2 apresenta a fundamentação teórica do trabalho, como os conceitos de correferência e a descrição das conferências ACE e HAREM. Já o capítulo 3 traz a descrição dos trabalhos relacionados, realizando um levantamento das *features* apresentadas em cada um deles. O capítulo 4 lista os recursos utilizados para a realização dos experimentos. O capítulo 5 apresenta as características técnicas do protótipo desenvolvido, enquanto que o capítulo 6 descreve os experimentos realizados, como também uma avaliação dos seus resultados. O capítulo 7 traz as considerações finais, com as contribuições, limitações e trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo irá abordar os conceitos base para a elaboração desta dissertação, que são a resolução de correferência e as categorias de ENs, de acordo com duas avaliações conjuntas: o ACE e o HAREM.

2.1. Resolução de correferência

Resolução de correferência é uma tarefa importante em diversas aplicações de PLN, pois dela depende uma compreensão adequada dos textos. O principal desafio da tarefa está na identificação automática e correta de expressões correferentes, que são as expressões que se referem a uma mesma entidade.

> 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...

Figura 1 – Fragmento de texto de exemplo do *corpus*.

No fragmento de texto da Figura 1, retirado do texto CIENCIA_2000_6389.txt do *corpus* Summ-it (descrito no Capítulo 4), 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. Este 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 habilidades cognitivas

humanas complexas, de difícil reprodução por sistemas computacionais. Por exemplo, 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?

O conjunto dessas expressões referenciais relativas a uma mesma entidade de mundo denomina-se cadeia de correferência. Este conjunto é responsável pela construção coesa de um texto e por isso sua importância, já que a coesão é responsável pela compreensão textual.

Para uma melhor compreensão de correferência, é preciso definir anáfora, já que seus conceitos estão relacionados.

Anáfora se define como toda a retomada de uma ideia já introduzida por uma entidade mencionada anteriormente. Pode-se dizer que quando uma entidade é citada pela primeira vez em um texto, se faz o processo de evocação da entidade. A expressão que faz a retomada da entidade é dita anafórica, e a expressão a quem ela se refere é chamada de antecedente. A relação entre essas expressões denomina-se relação de correferência. No fragmento do exemplo anteriormente citado, **Miguel Guerra** é o antecedente e **Guerra** é a anáfora.

Assim, 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.

A Figura 2 (retirada de [ABR05]) ilustra um exemplo de anáfora sem a relação de correferência.

> O Eurocenter oferece **cursos de Japonês** na bela cidade de Kanazawa. **Os cursos** têm quatro semanas de duração. **As aulas do nível avançado** incluem refeições típicas e passeios a pontos turísticos.

Figura 2 – Exemplo de texto explicativo sobre anáfora.

De acordo com o exemplo, a expressão anafórica **Os cursos** retoma uma expressão já citada no discurso, **cursos de Japonês** (antecedente), sendo que essas

duas expressões fazem menção à mesma entidade, são expressões correferenciais e anafóricas. Já a expressão **As aulas do nível avançado** não é correferente a nenhum termo, mas apresenta significado na expressão **cursos de Japonês**, sendo uma expressão anafórica, mas não correferente.

Geralmente, para o desenvolvimento do processo de resolução, é preciso realizar, além da análise sintática das frases, também a análise semântica e de contexto, o que envolve conhecimento de mundo. Uma combinação dessas informações se faz necessária para realizar a identificação das expressões correferentes. Porém, capturar um modelo do contexto é ainda um grande desafio para a área. Já a identificação de tipos semânticos é uma tarefa bem desenvolvida que pode ser adotada como parte da resolução de correferência.

A seção a seguir irá apresentar a descrição mais detalhada das categorias de ENs de acordo com duas avaliações conjuntas na área.

2.2. Categorias de entidades nomeadas

Nesta seção são apresentadas categorias de ENs, tal como abordadas nas avaliações conjuntas ACE [ACE08a] e HAREM [HAR08].

2.2.1. Avaliações conjuntas

Avaliações conjuntas são eventos com o intuito de incentivar e avaliar os sistemas desenvolvidos para resolver uma ou mais tarefas específicas, de acordo com os objetivos da avaliação.

Para este trabalho foram estudadas duas avaliações, o ACE e o HAREM, com o foco em tarefas relacionadas ao tema desta dissertação.

2.2.1.1. Avaliação ACE

O ACE é uma avaliação conjunta internacional com o intuito de abordar a compreensão da linguagem humana por meio da tecnologia, provendo o reconhecimento da linguagem textual.

Uma das tarefas definidas por essa avaliação diz respeito à identificação e relacionamento de entidades nomeadas de acordo com categorias pré-estabelecidas. Entidades não pertencentes às categorias pré-definidas não devem ser consideradas.

A tarefa de reconhecimento dos relacionamentos entre as entidades corresponde à resolução de correferência. No que se refere à forma de tratamento dada ao problema pelos sistemas, não há restrições em usar e buscar conhecimento extratextual, isto é, pode-se utilizar de dados externos aos textos como meio de apoio para o reconhecimento, como realizar consultas à internet e bases de dados, como a WordNet[1].

O ACE trata também dos sintagmas nominais, além dos nomes próprios, na tarefa de relacionamento entre as entidades.

O sintagma consiste em um conjunto de elementos que constituem uma unidade significativa dentro da oração, mantendo relações de dependência e ordem. Sua organização é referente a um elemento central, denominado de núcleo. De acordo com a classificação do núcleo é determinado o nome para o sintagma, por exemplo, no sintagma verbal, o verbo é o seu núcleo; já no sintagma adjetival, o núcleo é um adjetivo. O sintagma nominal (SN) pode ter como núcleo um nome ou um pronome substantivo [KOC04].

A estrutura básica de uma sentença é formada por um sintagma nominal e um sintagma verbal. Além desses elementos, uma oração pode ter outros sintagmas complementares, como, por exemplo, o sintagma preposicional e o sintagma adjetival. O sintagma preposicional é constituído de uma preposição seguido de um SN. Um sintagma preposicional ocorre dentro de um SN, sintagma verbal ou sintagma adjetival.

Retomando o exemplo da página 17 (Figura 1), pode-se definir a seguinte estrutura sintática para a frase: A opinião é do agrônomo Miguel Guerra, da UFSC.

[1] http://wordnetweb.princeton.edu/perl/webwn

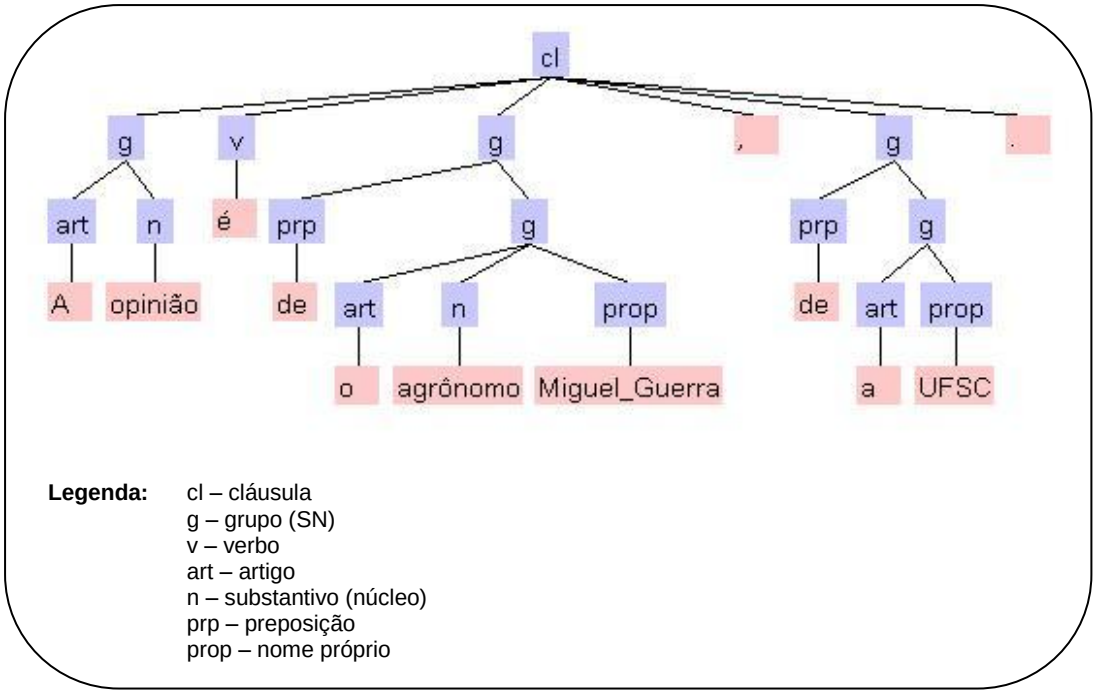

Figura 3 – Árvore sintática gerada para a frase: A opinião é do agrônomo Miguel Guerra, da UFSC.

Por meio desta ilustração, é possível identificar a composição de um SN (no exemplo marcado pela etiqueta “g”). Essa estrutura foi criada através da ferramenta de geração de estrutura sintática desenvolvida por Bick [BIC09], que se encontra disponível para a língua portuguesa na *Web*[2].

Esta dissertação tratará os sintagmas nominais para a tarefa de resolução de correferência.

No que se refere às categorias de ENs, o ACE as define da seguinte forma (os exemplos são retirados do *corpus* Summ-it, descrito no Capítulo 4):

1. FAC (instalação), que contempla aeroportos e outras construções
    a. Exemplo: “A estação...”;
2. GPE (entidade geopolítica) abrange continentes, países ou estados
    a. Exemplo: “O Brasil”;
3. LOC (localização), que se refere a endereços, fronteiras e regiões gerais
    a. Exemplo: “...o único país que se recusa a aceitar a determinação européia”;

[2] http://visl.sdu.dk/visl/pt/parsing/automatic/trees.php

4. ORG (organização), diz respeito a empresas, órgãos governamentais, dentre outros
    a. Exemplo: “...o CCNE (Comitê_Consultivo_Nacional_de_Ética) , órgão que orienta o governo francês sobre aspectos éticos de a biotecnologia”;
5. PER (pessoa), que se refere tanto a grupos como a indivíduos
    a. Exemplo: “...Adalberto_Veríssimo , de a ONG_Imazon”.

Observando a descrição de cada categoria, percebe-se que uma categoria possui um conjunto de características, as quais expressam uma delimitação das possibilidades de identificação para uma EN. Esse conjunto pode permitir a definição de um conhecimento de mundo em relação à tarefa de correferência, já que uma categoria representa um grupo de definições específicas.

Para realizar a marcação de correferência, a avaliação definiu que cada entidade é identificada por uma etiqueta. Após, outra etiqueta irá identificar se é um referente (REF) ou um atributo de referência (ATR). ATR é toda a entidade que faz ligação com o referente e, por isso, o mesmo deverá acompanhar a etiqueta do referente, gerando, assim, o elo de correferência. O subtipo também é indicado, por exemplo, individual (ind), grupo (group) [ACE08b]. A Tabela 1 mostra um exemplo dessa marcação.

Tabela 1 – Marcação de correferência do ACE.

Texto	**Tipo**	**Ref-Atr**
Fernando Henrique e Rigoto Ex-presidente e ex-governador	PER.GROUP	E1-REF
Fernando Henrique	PER.IND	E2-REF
Ex-presidente	PER.IND	E2-ATR
Rigoto	PER.IND	E3-REF
Ex-governador	PER.IND	E3-ATR

2.2.1.2. Avaliação HAREM

O HAREM é uma avaliação conjunta que tem o objetivo de avaliar sistemas reconhecedores de entidades nomeadas ou entidades mencionadas (EMs), de acordo com a sua terminologia própria. EMs são definidas como nomes próprios considerando-se o contexto no qual o nome está inserido. O HAREM, em sua segunda edição (2008), propôs uma nova trilha na avaliação referente ao reconhecimento de relações entre EMs, a ReRelEM (Reconhecimento de Relações entre Entidades Mencionadas), uma tarefa próxima à resolução de correferência que, no entanto, não considera os sintagmas nominais com o núcleo substantivo comum.

A seguir, serão descritas as categorias definidas no HAREM [SAN07] (os exemplos são retirados do *corpus* Summ-it, descrito no Capítulo 4):

1. Pessoa: refere-se a uma pessoa ou grupo de pessoas
    a. Exemplo: “Os pesquisadores...”.
2. Organização: refere-se a entidades que apresentam características definidas por uma estrutura organizacional
    a. Exemplo: “...o LIP (Laboratório_de_Instrumentação_e_Partículas) , de o Instituto_de_Física...”.
3. Acontecimento: abrange eventos que descrevam um conjunto de atividades ou ações
    a. Exemplo: “...a Declaração_de_Helsinque...”.
4. Lugar: entidades que referenciam um local específico
    a. Exemplo: “Portugal...”.
5. Abstração: entidades que exprimam idéias
    a. Sem exemplos no *corpus*.
6. Obra: entidades que foram construídas pelo homem e que tenham um nome próprio
    a. Exemplo: “...o veículo...”.

7. Valor: refere-se a entidades absolutas ou relativas. Itens numéricos utilizados para marcar ordem no texto não devem ser considerados
    a. Sem exemplos no *corpus*.
8. Coisa: refere-se a objetos (com ou sem forma determinada), ou a um conjunto de objetos
    a. Exemplo: “...madeira brasileira...”.
9. Tempo: entidades que se referem explicitamente à data ou hora. Nomes de meses com a primeira letra em maiúscula são considerados membros dessa categoria. Não devem ser considerados valores que identifiquem idade
    a. Sem exemplos no *corpus*.
10. Outro: entidades que não atenderam a nenhuma das categorias acima listadas
    a. Exemplo: “...água_potável...**”**.

A tarefa de classificação do HAREM se diferencia do ACE por identificar todos os nomes próprios do texto (classificando-os como Outro, se não for de nenhuma das categorias definidas anteriormente).

Outra diferença é que o HAREM trata apenas nomes próprios, enquanto o ACE considera todas as referências ocorridas no texto em relação a uma categoria, e irá relacionar a expressão "o pesquisador" com um nome próprio previamente introduzido no texto, "José Steiner", por exemplo.

Para a marcação de correferência, o HAREM determina que cada entidade receba um número como identificação (ID). Quando ocorre a relação de correferência, esse ID deverá ser informado na marcação *corel*, sendo esta a responsável em criar as relações entre as entidades. Caso ocorra que a referência seja entre mais de uma entidade, deve-se colocar o ID de todas as entidades, separado por espaço, conforme exemplo da Figura 4 (as marcações <EM> e </EM> delimitam a EN).

Em 9 de Setembro de 1895, foi organizado em <EM ID=”15”> New York </EM> o <EM ID=”16” COREL=”15” TIPOREL=”ocorre_em”> Congresso Americano de Bowling </EM> (”<EM ID=”17” COREL=”16 15” TIPOREL=”ident ocorre_em”>ABC</EM> - <EM ID=”18” COREL=”16 15” TIPOREL=”ident ocorre_em”> American Bowling Congress </EM>”), sediado em <EM ID=”19” COREL=”15 16 17 18” TIPOREL=”incluido sede_de”> Milwaukee </EM>, com o objectivo de aplicar medidas correctivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

<EM ID="FG51”> João Steiner </EM>, astrofísico da USP, durante a (...),

explicou <EM ID="FG560"COREL="FG51”> Steiner </EM>.

Figura 4 – Trechos de marcação de correferência do HAREM.

Para realizar a marcação de correferência, também será necessário realizar a identificação do tipo de relação entre as entidades correlacionadas.

A avaliação definiu as seguintes identificações e como devem ser aplicadas:

- Identidade – todas as relações, exigindo igualdade de categoria, tipo e subtipo.
- Inclusão – todas as relações, exceto valor, exigindo igualdade de categoria.
- Ocorre_em – relação entre acontecimento e local, e relação entre organização e local.
- Outra – várias relações que não se encaixam nas definições anteriores. Um exemplo é dado pelas relações familiares entre entidades do tipo Pessoa.

A relação de identidade ocorre entre entidades mencionadas da mesma categoria. A relação de inclusão é estabelecida quando uma entidade faz parte de outra entidade. A relação *ocorre_em* infere relação de localização, por isso, deve ser interpretada como *localizada_em* quando estiver tratando a relação entre categorias do tipo Organização e Local [SAN07].

Abaixo, a descrição de como a marcação deverá ocorrer para identificar as relações entre entidades.

- identidade (sem TIPOREL ou TIPOREL=”ident”)
- inclusão (TIPOREL=”inclui” ou TIPOREL=”incluido”)
- ocorre_em (TIPOREL=”ocorre_em” ou TIPOREL=”sede_de”)

Os exemplos da Figura 4 mostram o modelo de retorno dos sistemas, fazendo uso da marcação de relação e de correferência. É importante considerar essas definições, pois os sistemas são avaliados com base nessas informações e formato.

2.3. Considerações sobre este capítulo

A partir da compreensão do que é correferência, percebe-se o quanto é importante sua resolução para diversos sistemas computacionais, como a sumarização automática, já que sua resolução está diretamente relacionada com a clareza e a coerência textual.

Observando, também, as definições das características para determinar uma categoria, pode-se inferir que elas podem ser utilizadas como forma de conhecimento de mundo. Neste trabalho, conhecimento de mundo é o contexto do assunto em que a EN está inserida. Assim, cada categoria é vista a partir de suas peculiaridades semânticas, o que poderá possibilitar melhores resultados na tarefa de resolução de correferência.

Como essas avaliações conjuntas atuam focando o uso das categorias de ENs, torna-se interessante desenvolver sistemas que atendam aos objetivos dessas avaliações e, com isso, tentar atingir melhores resultados em relação à língua portuguesa.

3. TRABALHOS RELACIONADOS

Este capítulo irá descrever um conjunto de trabalhos relacionados compreendendo um estudo das *features* para a resolução de correferência. Um conjunto de *features* que seja significativo para a tarefa de resolução, com base nos resultados dos trabalhos estudados, é identificado.

3.1. O trabalho de Soon et al.

O trabalho de Soon et al. [SOO01] trata da tarefa de resolução de correferência abordando os vários tipos de sintagmas nominais. Cabe destacar que esse trabalho trata da tarefa em textos de qualquer domínio em língua inglesa.

Essa foi a primeira proposta que uniu os conceitos de aprendizado de máquina e processamento de *corpus*. Essa abordagem necessita de um *corpus* anotado com informações das cadeias de correferência dos sintagmas nominais. Essas informações são usadas no processo de aprendizado para a tarefa de resolução de correferência, definida pelo MUC-6[3] e MUC-7. Uma diferença da proposta dos autores está em identificar todas as entidades e não apenas as que são estabelecidas pelo MUC-6 e MUC-7.

O processo de resolução ocorre através do aprendizado com base em *features* para classificar um par de termos como anafórico ou não. Os autores usaram 12 *features*, que são:

1. Str_Match - ambos os termos possuem a mesma grafia;
2. Alias - um termo é sigla do outro;
3. I_Pronoun - se o antecedente é pronome;
4. J_Pronoun - se a anáfora é pronome;
5. Def_NP - se a anáfora começa pelo artigo *the*;
6. Dem_NP - se a anáfora começa por *this*, *that*, *these* ou *those*;
7. Number - ambos os termos são numerais;

[3] Conferência substituída pelo ACE.

8. Gender - ambos os termos possuem o mesmo gênero;
9. Proper_Name - se os termos são nomes próprios;
10. Appositive - se a anáfora é aposto do antecedente;
11. Dist - número de frases que separam os termos;
12. SemClass - se os termos possuem a mesma categoria semântica. Essa informação é extraída da WordNet.

Para quase todas as *features* são considerados os valores "verdadeiro" ou "falso", sendo que a *feature* 12 apresenta, além desses valores, o valor "desconhecido", caso não identifique uma categoria semântica dentre as definidas. Outra *feature* que traz um valor diferente é a 11, sendo atribuído um valor numérico, já que se trata da distância entre a anáfora e o antecedente.

Um dos experimentos realizados foi de 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 valor de retorno diferente de zero. De acordo com os autores [SOO01], esse resultado demonstra que as *features* são importantes para a tarefa, pois isoladamente apresentam retorno na classificação dos pares.

Por esse ser o primeiro trabalho com *features*, ele é comumente considerado o *baseline* na área de PLN para a resolução de correferência.

3.2. O trabalho de Strube e Ponzetto

Outro trabalho que descreve uma solução para a resolução de correferência em língua inglesa baseada em aprendizado de máquina é o de Strube e Ponzetto [STR07]. Essa proposta baseia-se em Soon et al., contudo, faz o uso mais elaborado de bases de conhecimento externas, sendo estas a Wikipédia[4] e a WordNet.

A utilização de bases de conhecimento na resolução de correferência permite a busca de informação semântica, a qual auxilia no processo de desambiguação de entidades nomeadas. Por exemplo, um texto apresenta a entidade Bento Gonçalves. Como identificar se a entidade está se referindo à cidade Bento Gonçalves do Rio Grande

[4] http://www.wikipedia.org/

do Sul, à Avenida de Porto Alegre ou ao personagem que fez parte da Revolução Farroupilha? Poder-se-ia resolver a questão com a ajuda da identificação de palavras próximas à entidade em conjunto com consultas a bases de conhecimento.

Para essa tarefa, são extraídos pares de termos (representados por ER_i e ER_j), os quais passarão por um processo de análise para determinar se são relacionados ou não. Esse processo de análise contempla as seguintes *features* [STR07], além das 12 *features* apresentadas em Soon et al. [SOO01]:

1. I_Semrole - o papel semântico (informação fornecida pelo *parser* Assert) do antecedente (ERi);
2. J_Semrole - o papel semântico (informação fornecida pelo *parser* Assert) da anáfora (ERj);
3. WN_Similarity_Best - a maior pontuação de similaridade entre todos os *synsets* dos pares (baseada na taxonomia da WordNet – tamanho do caminho percorrido na taxonomia);
4. WN_Similarity_Avg - a média da pontuação de similaridade;
5. I/J_Gloss_Contains - atribui-se o valor indefinido se não há páginas da Wikipédia intituladas com $ER_{i/j}$, verdadeiro se ER_i estiver presente no título e ER_j no primeiro parágrafo do texto (e vice-versa), caso contrário, falso;
6. I/J_Related_Contains - atribui-se o valor indefinido se não há páginas da Wikipédia intituladas com $ER_{i/j}$, verdadeiro se ER_i estiver presente no título e ER_j presente em pelo menos um *hyperlink* (e vice-versa), senão, falso;
7. I/J_Categories_Contains - atribui-se o valor indefinido se não há páginas da Wikipédia intituladas com $ER_{i/j}$, verdadeiro se ER_i estiver presente no título e ER_j na lista de categorias da página de ER_j (e vice-versa), senão, falso;
8. Gloss_Overlap - a média (medida de similaridade) entre o primeiro parágrafo do texto da página de $ER_{i/j}$, calculada pela equação:

   a. $$\text{lesk\_wikipedia}(t_1 t_2) = \tanh \frac{(\text{overlap}(t_1, t_2))}{(\text{length}(t_1) + \text{length}(t_2))}$$

   onde “t_1” representa o primeiro texto e “t_2” o segundo. “*length*” significa o tamanho de “t_1” e “t_2”, respectivamente. Essa medida é para normalizar e fornecer um escore de relacionamento entre os textos. E

"tanh" representa a função tangente hiperbólica, utilizada para minimizar os *outliers* (valores discrepantes).

9. Wiki_Relatedness_Best - a pontuação mais alta do relacionamento de C_{ERi} e C_{ERj} (onde C é o conjunto de categorias extraídas a partir das páginas de $ER_{i/j}$). Essa pontuação é calculada a partir do caminho percorrido no gráfico de categorias da Wikipédia;
10. Wiki_Relatedness_Avg - a média dessa pontuação.

As *features* 3 e 4 são referentes à WordNet, enquanto que as *features* 5, 6, 7, 8, 9 e 10 à Wikipédia.

Para determinar a página da Wikipédia que será usada para a definição das *features*, o autor fez uso de um algoritmo [STR07] que recupera páginas da Wikipédia e as desambigua semanticamente. Para isso, é realizada uma consulta através dos termos (ER_1 e ER_2), retornando todos os redirecionamentos (por exemplo, a palavra **carro** redireciona para **automóvel**). Através do mesmo algoritmo, resolve-se a ambiguidade das páginas, retornando a página que representará cada termo.

Com o objetivo de comparar as *features* do trabalho de Soon et al. e as que se referem ao uso de base de conhecimento, os autores realizaram um experimento processando o *corpus* disponibilizado pelo ACE do ano de 2003 utilizando apenas as *features* de Soon et al. e, depois, com o conjunto todo de *features*, isto é, incluindo as *features* relacionadas com a Wikipédia e a WordNet. Os resultados obtidos demonstraram que o uso de base de conhecimento (informação semântica) proporciona melhores resultados à tarefa de resolução de correferência. Os valores são detalhados na Tabela 2.

Tabela 2 – Tabela dos resultados do experimento de Strube e Ponzetto [STR07].

	Cobertura	Precisão	Medida-F
Baseline	50,5%	82,0%	62,5%
WordNet	59,1%	82,4%	68,8%
Wikipédia	58,3%	81, 9%	68,1%

3.3. O trabalho de Souza

Em Souza [SOU08], o sistema ACROPOS é apresentado com objetivo de automatizar a resolução de correferência para a língua portuguesa. O sistema seleciona as cadeias de correferência de um texto através do aprendizado dos pares de expressões anafóricas.

Esse é o único trabalho que trata da resolução de sintagmas nominais para a língua portuguesa. Há outros trabalhos que tratam apenas da resolução de pronomes, tais como [CHA07], [CHA08] e [COE05]. Esse sistema não observa as diferentes categorias referenciais, observadas no reconhecimento de entidades nomeadas.

Os pares de sintagmas são gerados de acordo com a metodologia apresentada em [SOO01], a qual faz a associação entre os sintagmas da cadeia da seguinte forma: SN1-SN2, SN2-SN3, SN3-SN4 para os pares positivos. Isto é, o primeiro sintagma da cadeia forma par com o segundo, o segundo com o terceiro e assim sucessivamente. E para os pares negativos, o segundo sintagma da cadeia (SN2) faz par com todos os sintagmas existentes entre ele e o primeiro sintagma (SN1).

Também de acordo com a proposta de Soon et al. [SOO01], o autor fez uso de *features*, sendo que algumas foram adaptadas para atender às características do *corpus* disponível. Abaixo estão listadas as *features* definidas para a tarefa:

1. Cores-Match - comparação dos núcleos do par de sintagmas;
2. Distance - distância em número de frases entre os dois sintagmas;
3. Antecedent-is-pronoun - verificação se o antecedente é pronome;
4. Anaphora-is-pronoun - se a anáfora for um pronome receberá o valor verdadeiro;
5. Both-proper-names - se os sintagmas são nomes próprios;
6. Gender-agreement - concordância de gênero;
7. Number-agreement - concordância de número;
8. Both-subject - se os sintagmas são sujeitos;
9. Semantic-agreement - concordância semântica (se possuem etiquetas semânticas idênticas);

10. Same-semantic-group - mesmo grupo semântico (caso possuam etiquetas semânticas que pertençam ao mesmo grupo).

Na *feature* 10, o autor selecionou as etiquetas de mesmo grupo de acordo com a definição de [BIC00], criador do analisador sintático, o qual será detalhado no Capítulo 4. O analisador é responsável pelo fornecimento das etiquetas semânticas.

Também de acordo com Soon et al. [SOO01], Souza trabalhou com aprendizado de máquina. A árvore de decisão gerada pelo algoritmo de aprendizado J48 é ilustrada na Figura 5:

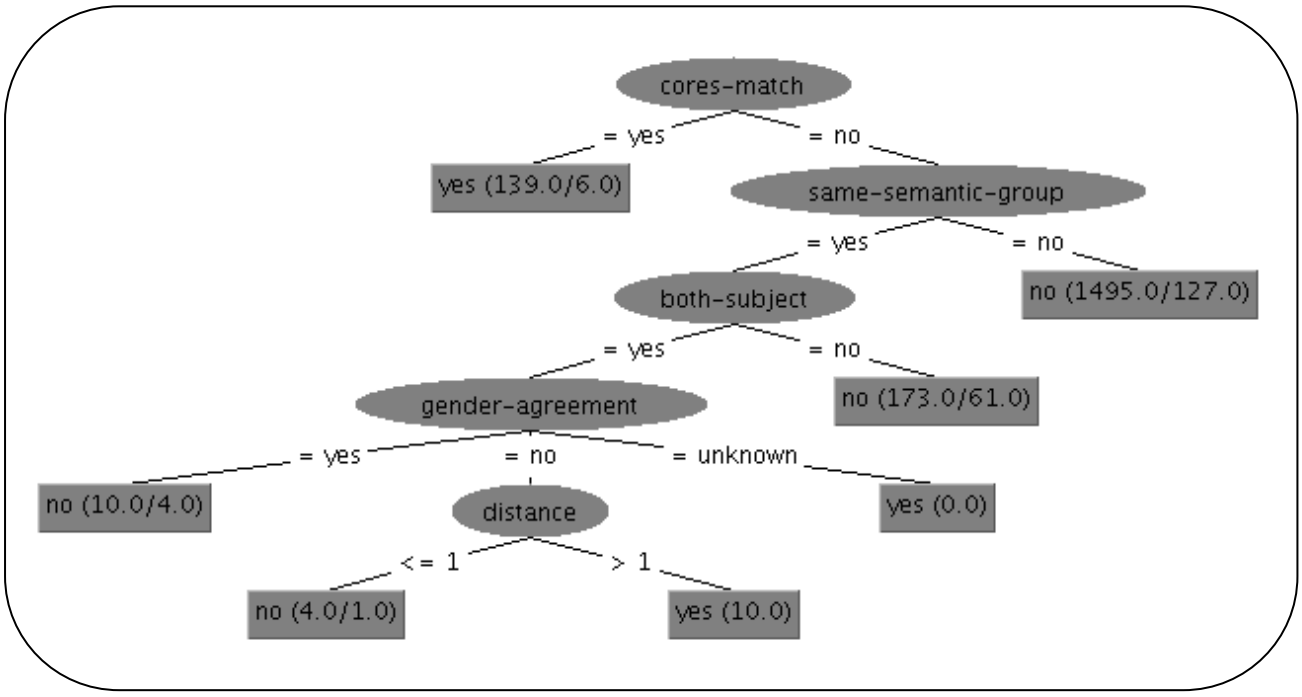

Figura 5 – Árvore de decisão induzida pelo algoritmo J48.

A partir desse retorno, nota-se que as *features* “cores-match”, “same-semantic-group”, “both-subject”, “gender-agreement” e “distance” foram as que apresentaram maior relevância, já que o algoritmo de aprendizado as utilizou para gerar a árvore de decisão.

3.4. Os trabalhos de Ng

Ng e Cardie [NG02] propuseram um conjunto de 37 *features*, classificadas de acordo com 4 grupos: lexical, gramatical, semântico e posicional. Por exemplo, a *feature*

“strings iguais” é classificada como lexical, enquanto a *feature* “aposto” como gramatical. Essas novas *features* estão listadas no Anexo A (serão listadas apenas as que diferem do *baseline*).

De acordo com os resultados, as *features* que apresentaram maior importância para identificar se um par é anafórico ou não foram as *features* “*alias*” e a “mesmo núcleo”. Essas mesmas *features* também foram destacadas no trabalho de Soon et al. [SOO01].

Outro trabalho de Ng [NG05] apresenta uma lista menor de *features* (24) do que aquela proposta pelo autor anteriormente, trazendo uma classificação também diferente. Neste trabalho a classificação está dividida da seguinte forma: *features* que descrevem o candidato a antecedente, *features* que descrevem o pronome e *features* que descrevem o relacionamento entre eles. Por intermédio desse conjunto de *features*, o autor quer realizar a tarefa apenas para a resolução de pronomes.

Em outra proposta que se refere à resolução de correferência, Ng [NG07] faz uso de sete tipos de *features* para determinar a classe semântica. Para realizar a identificação da classe semântica, busca a informação na WordNet, relacionando os nomes das classes definidas pelo ACE com as existentes na WordNet. Para a extração de informação gramatical fez utilização do *parser* Minipar. As *features* são as seguintes:

1. Word - para cada palavra de um SN é criado a *feature* Word, menos para as *stopwords*;
2. Subj_verb - se um determinado SN está envolvido com a relação de verbo subordinado, é criada então a *feature*;
3. Verb_obj - se um SN possui a relação de objeto com um verbo;
4. EN - faz uso do BBN's IdentiFinder (sistema que determina o tipo de um EN de acordo com as regras do MUC) para o reconhecimento de EN. Associar o tipo de EN segue as regras do ACE;
5. Wn_Class - para cada *keyword* (listadas na Tabela 3) da *feature* Word é determinado se um núcleo de um SN é hipônimo da *feature* Word na WordNet.
6. Induced_class - Com o uso do IdentiFinder, coloca-se como rótulo o tipo de EN em cada EN e com o Minipar (que é o analisador sintático) extrai-se as informações de aposto. O autor faz o cálculo da probabilidade de

substantivos comuns que co-ocorrerem com cada tipo de EN baseado no aposto, e se essa probabilidade for maior que 0,7, então, atribui-se a essa *feature* o valor do tipo de EN.

7. Neighbor - com base na pesquisa na área de semântica lexical, que diz que há uma similaridade na distribuição dos SNs, criou-se essa *feature*, que para cada SN são escolhidos dez SNs vizinhos e similares semanticamente. Essa similaridade semântica é provida de um tesauro desenvolvido por Lin's [NG07].

A Tabela 3 apresenta a relação entre as categorias do ACE com as *keywords* da WordNet.

Tabela 3 – Conjunto de *keywords* por categoria do ACE.

Categorias ACE	***Keyword***
Pessoa	Pessoa
Organização	Grupo social
Facility	Estabelecimento, construção, lugar de trabalho
GPE	País, governo, cidade, sociedade, comunidade, centro, ilha, administração
Localização	Área de terra, região, área com água, área geográfica, formação geológica

Essas *keywords* foram escolhidas através dos experimentos realizados pelo autor na base de treino levando em consideração a WordNet e as classes semânticas do ACE. O trabalho de Ng é a primeira proposta a considerar categorias de ENs de acordo com os conceitos estabelecidos por uma avaliação conjunta.

Em outro trabalho [NG09], o autor desenvolveu uma medida para determinar se um termo é anafórico, ou seja, ajudar a definir se um par é anafórico ou não. O autor fez uso das 37 *features* do seu trabalho [NG02] para o processo de classificação do par. O método de geração dos pares segue a proposta de Soon et al., tanto para os positivos quanto para os negativos. A proposta é gerar pares de exemplos certos e errados e, depois do processo de aprendizado, classificar se um par é correferente ou não.

Nesse trabalho, o autor utilizou-se do *corpus* do ACE como base de dados. Para método de comparação, o autor fez uso de dois trabalhos seus anteriores, [NG02] [NG04],

como também de outros autores que atuaram na mesma área com relação a sua proposta de uso da medida de probabilidade da anáfora [NG09].

De acordo com o autor, a medida-F de seu sistema proposto foi maior do que todos os outros sistemas utilizados para comparação (59,4%).

3.5. Considerações sobre este capítulo

Nota-se que os trabalhos para o inglês são mais elaborados em comparação aos trabalhos que tratam do português. Um dos motivos é o fato de atualmente existir mais recursos disponíveis para o inglês do que para o português. Um exemplo disso seria a WordNet. Os resultados dos experimentos de Strube e Ponzetto [STR07] mostram que o percentual de acerto usando as *features* para a WordNet é superior em relação ao experimento sem as mesmas.

Em relação à evolução das *features* percebe-se que as primeiras focavam-se no processamento morfológico. À medida que a informação semântica através do acesso a bases de conhecimento passou a ser utilizada, novas *features* foram criadas com o uso desse conhecimento. Considerando os resultados alcançados nos trabalhos, é possível constatar que aqueles que englobaram as *features* propostas com o uso de base de conhecimento conseguiram obter resultados melhores.

No que diz respeito à informação semântica, é preciso analisar quais *features* são possíveis de implementar de acordo com os recursos disponíveis para a língua portuguesa, pois, por exemplo, não há uma base de informação como a WordNet que esteja bem definida para o português.

Um recurso semântico prático que poderia ser utilizado são as informações semânticas fornecidas pela análise do analisador sintático PALAVRAS, descrito em detalhes no Capítulo 4.

Com base nos estudos anteriores e nas possibilidades práticas de implementação, identificou-se as seguintes *features* a serem implementadas, detalhadas no Capítulo 5: mesmo gênero, mesmo número, mesmo núcleo, mesmo núcleo semântico e mesmo conjunto semântico. A escolha por essas features deve-se ao fato das mesmas apresentarem relevância a tarefa de resolução de acordo com os trabalhos relacionados, como também, incluem o conhecimento de mundo à tarefa.

4. RECURSOS UTILIZADOS

Este capítulo irá apresentar os recursos que foram escolhidos para a realização deste trabalho. A escolha ocorreu considerando a disponibilidade dos mesmos para a língua portuguesa.

4.1. Analisador sintático PALAVRAS

O analisador sintático PALAVRAS possui o objetivo de processar textos da língua portuguesa, realizando as análises morfossintáticas e semânticas. Foi desenvolvido por Bick, como tese de seu doutorado [BIC00]. Sua saída é um único arquivo com todas as informações do processamento.

Essa ferramenta participou da avaliação conjunta HAREM, obtendo índices bons em sua avaliação. Para a tarefa de identificação, o analisador alcançou os melhores resultados entre os participantes (80%), enquanto para a tarefa de marcação semântica, conseguiu os melhores resultados em 2 dos 4 cenários (cenários podem ser descritos como conjuntos de parâmetros pré-definidos para a realização dos testes) disponibilizados pela avaliação (63% para os dois cenários). Para ambas as tarefas foram considerados os resultados da medida-F [HAR08].

Uma das etapas do analisador é a identificação das ENs e sua classificação de acordo com premissas definidas pelo autor. Para isso, ocorre a associação de cada entidade identificada com etiquetas semânticas.

Será apresentado a seguir como o autor realizou a marcação semântica no que diz respeito à desambiguação, já que essa informação é relevante para a tarefa de resolução de correferência.

O autor definiu uma lista com várias etiquetas para realizar as marcações, tanto para informações morfossintáticas como semânticas. A lista com o conjunto completo de etiquetas utilizadas é encontrada em [BIC09].

No total 157 etiquetas semânticas foram definidas, classificadas de acordo com um conjunto de características definidas pelo autor. A seguir, na Tabela 4, um exemplo de conjunto de etiquetas semânticas e as características atribuídas para a representação de seres humanos.

Tabela 4 – Etiquetas semânticas da representação de seres humanos.

Etiqueta semântica	**Características**
H	Humana
HH	Grupo de humanos
Hattr	Substantivos com atribuição humana com final –ista, -ante
Hbio	Relacionadas à biologia humana (como raça, idade)
Hfam	Termo relacionado com família (como pai, noivo)
Hideo	Ideologia (por exemplo, religião)
Hmyth	Místico, mitológico (como duendes)
Hnat	Nacionalidade (por exemplo, brasileira)
Hprof	Profissão (como pesquisador), e também hobbies e esportes (como alpinista)
Hsick	Doenças humanas (como asmático, diabético)
Htit	Termos relacionados com títulos (por exemplo, senhora, rei)

O analisador realiza o processo de desambiguação criando uma árvore de derivação. Cada final de caminho da árvore equivale a uma classe, sendo denominada pelo autor como protótipo. No exemplo da Tabela 4, o protótipo definido é o de humanos. Um protótipo é um conjunto de etiquetas semânticas que determinam uma classe. Por exemplo, para definir se uma classe pertence ao protótipo animal precisa apresentar as seguintes características: é concreto, animado, não-humano e se move.

Por intermédio dos protótipos, o autor determina a possível classe de uma palavra. Mesmo não fazendo uso de bases de conhecimento, como dicionários, o autor conseguiu

estabelecer um conjunto de características de forma a criar, indiretamente, o conhecimento de mundo, auxiliando-o no processo de desambiguação [BIC00].

Depois de determinado o protótipo, é necessário realizar a desambiguação entre as possíveis etiquetas do conjunto. Para isso, o autor faz uso de uma gramática de restrições criada por ele, de forma a atender as peculiaridades da língua portuguesa. Para realizar essa tarefa de desambiguação, faz-se necessário verificar o retorno da análise da palavra vizinha. Por exemplo, o verbo *correr* com a característica de movimento necessita de um complemento que também apresente a mesma característica, ou seja, de movimento [BIC00].

Tendo em vista que a ferramenta participou do HAREM, o autor estabeleceu uma associação entre algumas das etiquetas retornadas pelo analisador com as categorias definidas pela avaliação conjunta.

A Tabela 5 mostra a relação das etiquetas semânticas processadas pelo analisador PALAVRAS com as categorias do HAREM. Essas marcações foram definidas pelo autor [BIC09].

Tabela 5 – Grupo de etiquetas semânticas para cada categoria do HAREM.

Categoria	**Etiqueta Semântica**
Pessoa	Hprof, groupind, groupofficial, hum, official, H, Htitle, member
Abstracao	brand, genre, school, idea, plan, author, absname, disease
Organizacao	admin, org, inst, media, party, suborg, Linst
Lugar	top, civ, address, site, virtual, road, Ltop, Lciv, Lh
Obra	tit, pub, product, V, artwork, Vair, Vwater
Coisa	object, common, mat, class, plant, currency
Acontecimento	occ, event, history
Valor	quantity, prednum, currency
Tempo	date, hour, period, cyclic

Nota-se que para realizar a associação, um conjunto de características, que expressa uma relação entre as marcações, é considerado.

Esse processo de associação ocorre para todas as categorias estabelecidas na avaliação conjunta. Assim, pode-se inferir que as categorias de ENs representam um conhecimento de mundo.

4.2. Summ-it

O Summ-it é um *corpus* de textos em língua portuguesa com anotações linguísticas [COL07]. Esse recurso é muito importante para essa área de estudos, por proporcionar uma base de conhecimento rica em informações linguísticas relacionadas ao discurso.

O *corpus* é composto por cinquenta textos jornalísticos do caderno de Ciências da Folha de São Paulo retirados do *corpus* PLN-BR[5]. 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.

O analisador sintático utilizado para processar o *corpus* foi o PALAVRAS. 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 *token*s, composto pelo *token* e seu respectivo ID; outro com as informações dos sintagmas, isto é, qual o ID do *token* inicial e final do sintagma e outro com as informações sintáticas-semânticas associadas ao ID do *token*. Os arquivos estão em formato XML. A escolha pela extensão XML deu-se devido a este ser um formato padrão.

A nomenclatura adotada para identificar os arquivos foi: *título_do_texto-tipo_arquivo.xml*, aonde *tipo_arquivo* é definido da seguinte forma: *token*s para *token*s, *phrases* para estruturas sintagmáticas e *pos* para sintático-semântico. Sendo assim, para cada texto são encontrados três arquivos oriundos do analisador e mais um arquivo, o de *markables*, com as informações dos sintagmas nominais do texto. O arquivo de *markables* foi criado manualmente, anotado através da ferramenta MMAX.

Dos 5047 *markables* (sintagmas nominais anotados), a maior parte corresponde a sintagmas nominais com nome núcleo (95,15%). Já os pronomes representam apenas 4,82% [COL07].

---

[5] Projeto destinado ao desenvolvimento de recursos e ferramentas para a recuperação de informação em bases textuais em português do Brasil. Disponível em http://www.inf.pucrs.br/~linatural/pln-br.html

Cabe destacar que a anotação do *corpus* não levou em consideração a identificação das categorias de ENs. Como os sistemas de resolução de correferência são avaliados nos contextos de conferências, como ACE e HAREM, observa-se a importância de uma análise de *corpus* em relação a essas categorias. Além disso, as categorias podem servir como uma estrutura inicial para definição de características semânticas que podem auxiliar o processo de resolução automática de correferência.

4.2.1. Estudo do *corpus* de acordo com as categorias do HAREM

Com base no retorno do analisador PALAVRAS e da relação das etiquetas semânticas do mesmo com as categorias da avaliação conjunta, realizou-se uma análise das cadeias do *corpus*, em que se constatou que do total das 589 cadeias, há 84 cadeias da categoria Pessoa, por exemplo. O Apêndice A apresenta a quantidade de cadeias por texto do *corpus*. Na Tabela 6 pode-se verificar o total de cadeias por categoria do *corpus*.

Tabela 6 – Total de cadeias por categoria do conjunto de textos do *corpus.*

Categorias						
PESSOA	ORGANIZACAO	ACONTECIMENTO	LOCAL	OBRA	COISA	OUTRO
84 (14,26%)	31 (5,26%)	29 (4,92%)	50 (8,49%)	13 (2,21%)	1 (0,17%)	381 (64,68%)

É compreensível que a categoria Outro apresente um valor maior que das demais categorias, pois o *corpus* é composto pelo caderno de ciências, abordando assuntos relacionados a animais. Acrescenta-se, também, que o conjunto de etiquetas que Bick [BIC09] associou para cada categoria do HAREM é menor que o total retornado no processamento do analisador. Por isso, foi atribuído para a categoria Outro um grupo muito heterogêneo de etiquetas semânticas.

Abaixo (Figura 6 e 7) são apresentados exemplos de cadeias do *corpus* Summ-it.

uma pele biônica

pele -> an

a pele-robô futura

pele-robô -> ac anbo

a tecnologia

tecnologia -> domain

Legenda:
an – substantivos anatômicos
ac anbo – abstração de anatomia vegetal
domain – domínio em geral

Figura 6 – Exemplo de cadeia do tipo Outro.

ele e seus colegas

colegas -> Hfam

Mattson e colegas

colegas -> Hfam

a equipe

equipe -> HH

os cientistas

cientistas -> Hprof

Legenda:
Hfam – refere-se à família de pessoas
HH – grupo de pessoas
Hprof – profissão relacionada a pessoas

Figura 7 – Exemplo de cadeia do tipo Pessoa.

Através desses exemplos é possível constatar que a marcação semântica da categoria Pessoa apresenta uma semelhança (começam pela letra H), o que automaticamente possibilita sua comparação. Já a categoria Outro não apresenta essa semelhança, o que a torna mais complexa de se resolver de forma automática. Cabe salientar, também, que as categorias apresentam um conjunto de características mais definido em relação à categoria Outro.

Apesar de o analisador fazer uso de informação semântica para desambiguação, nota-se que o processo realizado possui algumas falhas. Observando o *corpus*, foi possível detectar que a desambiguação de nomes próprios nem sempre é correta. A Figura 8 mostra um exemplo disso, no qual o analisador marcou um evento que deveria ser classificado na categoria Acontecimento como categoria Organização (inst).

o Big_Bang , explosão que teria dado origem a o Universo

Big_Bang -> inst

o Big_Bang

Big_Bang -> inst

Legenda:
inst – instituição

Figura 8 – Exemplo de marcação incorreta de nome próprio.

a China

China -> org

o governo chinês

governo -> HH

a China

China -> civ

Legenda:
org – instituição, empresa
HH – grupo de pessoas
civ – cidade, estado

Figura 9 – Exemplo de cadeia com etiquetas semânticas de outras categorias.

Já a Figura 9 apresenta um exemplo no qual a entidade **a China** teve sua marcação semântica atribuída à categoria Organização (org) e, depois, para a anáfora com mesmo núcleo, foi atribuída uma etiqueta de outra categoria (categoria Local - civ). Uma possível explicação pode estar relacionada à vacuidade semântica para a entidade China.

4.3. Weka

O Weka[6] 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 é 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 como modularidade, polimorfismo, encapsulamento, reutilização de código, dentre outros. Além disso, é um software de domínio público.

O método de aprendizado de máquina consiste em aprender através de exemplos, dos quais características diferenciam exemplos de determinadas classes. Para isso, como dados de entrada, é necessário fornecer características que representam os exemplos. Essas características são conhecidas na área como *features*.

O Weka possui um formato próprio de arquivo, o ARFF. Antes de aplicar os dados a qualquer algoritmo do pacote Weka, estes devem ser convertidos para o formato, que é composto por basicamente duas partes. A primeira contém uma lista de todos os atributos, onde se deve definir o tipo do atributo ou os valores que ele pode representar, sendo que se ocorrer a utilização de valores, estes devem estar entre “{ }”, separados por vírgulas. A segunda parte consiste das instâncias, ou seja, os registros a serem minerados com o valor dos atributos para cada instância, separados por vírgula. A ausência de um item em um registro deve ser atribuída pelo símbolo “?”.

Na Figura 10, um exemplo de arquivo ARFF do Weka.

[6] http://www.cs.waikato.ac.nz/ml/weka/

```
% ARFF file for the coreference data with some numeric features
%
@relation coreference
@attribute dist numeric
@attribute alias {true, false}
@attribute gender { true, false }
@attribute coref? { yes, no }
@data
%%
2 instances
%
2, true, false, no
1, true, true, yes
```

Figura 10 – Trecho de um arquivo ARFF.

São vários os métodos implementados no Weka. Para este trabalho, será utilizado o algoritmo J48, o qual retorna uma árvore de decisão com as *features* relevantes que foram identificadas por meio da análise dos exemplos.

A árvore de decisão é muito utilizada para a inferência indutiva [MIT97]. É um método de aproximação de função de valor distinto, sendo suas funções de aprendizado representadas por uma árvore, a árvore de decisão. A construção da árvore é *top-down*, ou seja, do nodo raiz para os ramos. Primeiro, verifica-se qual dentre os atributos será o nodo raiz. A escolha é de acordo com o valor de entropia. Entropia é a medida da impureza da coleção dos exemplos de treino [MIT97]. Através dessa medida é possível medir a eficácia dos atributos na classificação da base de treino.

Assim, quanto menor a entropia mais significativo é o atributo, então, mais ao topo da árvore ele será colocado. Os outros nodos da árvore também são escolhidos de acordo com seu valor de entropia, sendo primeiro montado o ramo da árvore do lado esquerdo e depois do lado direito. Esses valores são calculados de acordo com os testes dos atributos na base de treino. Os atributos mais utilizados nos testes são escolhidos para compor a árvore.

Pode ocorrer que alguns atributos fiquem de fora da árvore pelo fato de não apresentarem uma significância nos testes, isto é, não foram utilizados de forma a ajudar na resolução do problema.

Já o método escolhido para avaliar a árvore de decisão foi a validação cruzada. Essa abordagem faz uso de pedaços do pacote de treino para treinar e de um outro para testar (conhecidos como *k-folds*) [MIT97]. Por exemplo, se a validação cruzada for

configurada para dez, significa que as instâncias de treino serão quebradas em dez partes, agrupando-se em pacotes. Em um pacote de dez instâncias, nove são usadas para treino e uma para teste, dentro do processo de treino. Do total de pacotes obtidos pela divisão, 1 é selecionado para teste e o restante para treino. Por exemplo, num arquivo de 1000 instâncias, têm-se 100 pacotes com dez instâncias cada, sendo que 99 pacotes serão utilizados para treino e 1 pacote para teste. Essas informações de testes são usadas para medir o erro da validação [MIT97].

4.4. Considerações sobre este capítulo

Os recursos utilizados neste trabalho foram escolhidos de acordo com sua disponibilidade para o português. Por isso, apesar de compreender que o *corpus* é de tamanho pequeno, é o que há de disponível com as informações necessárias, principalmente a marcação manual das cadeias de correferência para a composição das características dos exemplos e, também, para permitir a avaliação posterior do sistema.

O uso de aprendizado de máquina foi escolhido por ser o que é utilizado em outros trabalhos da área, como Soon et al. [SOO01], Strube e Ponzetto [STR07] e Ng e Cardie [NG02].

Cabe destacar que este trabalho não fez nenhum tipo de correção da marcação do analisador. Essa postura foi adotada pois a intenção é trabalhar com os dados extraídos de forma automática através dos recursos que se possui para a tarefa. Assim como a classificação, a identificação das entidades nomeadas realizada pelo analisador também não é corrigida ou revisada, manual ou automaticamente.

5. O PROTÓTIPO DESENVOLVIDO

Para realizar a tarefa de resolução de correferência, um protótipo foi desenvolvido em duas versões.

Uma, que será o *baseline* do trabalho, e outra, que fará uso das categorias de ENs. Para a segunda versão será dado o nome de Recorcaten, que significa “REsolução de CORrefência por CATegorias de ENs”.

A primeira etapa do trabalho foi levantar os requisitos do protótipo e modelá-lo de forma a possibilitar sua reutilização.

Uma visão geral do sistema é apresentada na Figura 11:

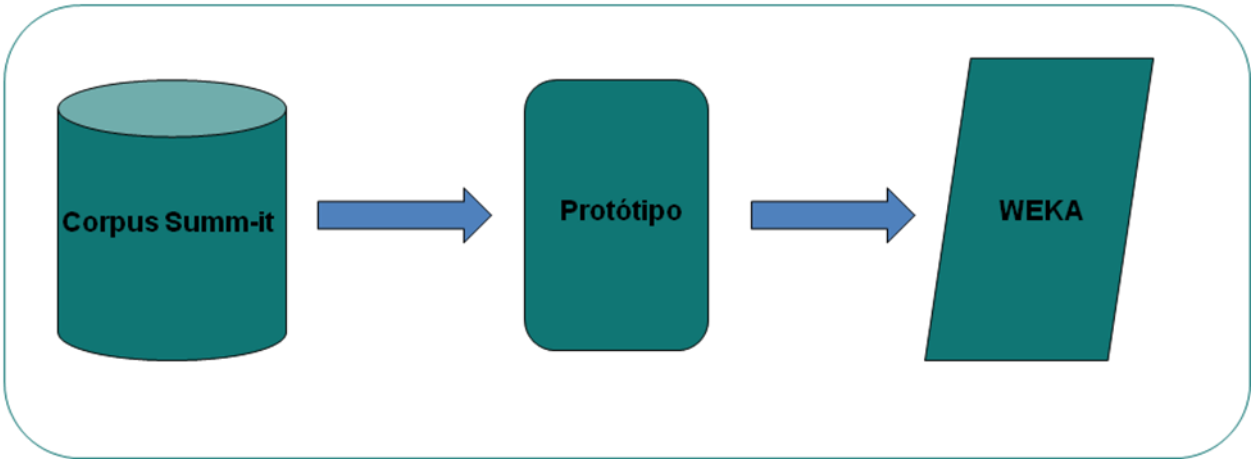

Figura 11 – Visão geral do sistema.

A entrada do protótipo é o *corpus* Summ-it processado pelo analisador PALAVRAS. Após sua execução, é gerado o arquivo ARFF, que será a entrada para o Weka, finalizando o ciclo de execuções.

A seguir serão descritas as características específicas tanto do *baseline* como do Recorcaten.

5.1. Características do *Baseline*

A primeira fase de implementação tem a função de gerar os pares de sintagmas sem considerar as categorias de ENs.

A linguagem escolhida para a implementação foi a Java, já que é a linguagem utilizada pelo grupo de pesquisa para o desenvolvimento de outras aplicações, bem como por ser uma linguagem disponível para uso e desenvolvimento de forma gratuita.

O protótipo foi modelado de modo a possibilitar a reutilização de parte de suas funções, por isso, decidiu-se segmentá-lo em um conjunto de classes. Assim, foi criada uma classe para a leitura dos arquivos do *corpus*, empregando a API JDOM[7] para a leitura da extensão XML, uma para a geração dos pares, uma classe específica para cada *feature*, e outra para gerar o arquivo TXT. Foi utilizada a API Weka[8] para gerar o arquivo ARFF. Essa divisão possibilita a inclusão de uma nova *feature* ou uma exclusão a qualquer momento, já que o processo de invocação das classes ocorre no programa principal.

Como apresentado no Capítulo 4, o *corpus* é composto por um conjunto de arquivos. O processamento do sistema necessita buscar dados em todos os arquivos, pois cada um possui uma informação específica necessária para obter as informações comparativas das *features*.

A seguir será detalhada a função de cada arquivo do *corpus* para o sistema.

O arquivo de *markables* é composto por informações de todos os sintagmas nominais do texto. Uma informação importante é o conjunto de sintagmas que fazem parte das cadeias de correferência, pois o processamento do sistema é dependente dela. O elemento *member* do arquivo diz a qual cadeia o sintagma pertence. Com essa informação é possível determinar os sintagmas que farão parte dos pares positivos e negativos posteriormente, pois se agrupam os sintagmas que pertencem ao mesmo valor de *member*. Esse é o primeiro arquivo processado pelo sistema.

A Figura 12 mostra um exemplo do arquivo *markables*.

[7] http://www.jdom.org/

[8] Esta está inserida juntamente com o pacote de instalação do Weka.

Texto: CIENCIA_2000_6389_*markables*.xces.xml

```
<cesAna>
<struct from="t1" to="t6" type="markable">
        <feat name="id" value="markable_1" />
        <feat name="np_n" value="yes" />
        <feat name="np_form" value="def-np" />
        <feat name="status" value="new" />
</struct>
<struct from="t4" to="t6" type="markable">
        <feat name="id" value="markable_2" />
        <feat name="np_n" value="yes" />
        <feat name="np_form" value="def-np" />
        <feat name="status" value="new" />
        <feat name="member" value="set_14" />
</struct>
<struct from="t15" to="t17" type="markable">
        <feat name="id" value="markable_3" />
        <feat name="np_n" value="yes" />
        <feat name="np_form" value="bare-np" />
</struct>
```

Figura 12 – Trecho do arquivo de *markables.*

A geração dos pares dar-se-á da seguinte forma: os pares positivos serão gerados a partir dos sintagmas pertencentes à mesma cadeia, tendo como base a proposta de Strube e Ponzetto [STR07], o qual coloca que todos os elementos de uma cadeia farão par com todos. A cadeia formada por SN1, SN2, SN3 e SN4, por exemplo, formará os seguintes pares positivos: SN1-SN2, SN1-SN3, SN1-SN4, SN2-SN3, SN2-SN4, SN3-SN4 (cada SN representa um sintagma da cadeia).

SN1 - ele e seus colegas
SN2 - Mattson e colegas
SN3 - a equipe
SN4 - os cientistas

Figura 13 – Exemplo de cadeia de correferência.

No exemplo da Figura 13, o sintagma **ele e seus colegas** fará par com **Mattson e colegas**, **a equipe** e **os cientistas**.

Decidiu-se utilizar a proposta de Strube e Ponzetto [STR07] nos pares positivos, pois esta proporciona um conjunto maior de exemplos positivos se comparada com a de Soon et al. [SOO01], a qual é implementada no trabalho de Souza [SOU07].

Já os pares negativos serão agrupados de acordo com a proposta de Soon et al. [SOO01], onde o segundo membro faz par com os sintagmas existentes entre ele e o primeiro membro da cadeia, sendo representado por SN1 e SN2, respectivamente. Por exemplo, cada sintagma entre SN1 e SN2 é representado por uma letra do alfabeto (a, b, c, d...), então, SN2 fará os seguintes pares: SN2-a, SN2-b, SN2-c, SN2-d. Um exemplo pode ser observado na Figura 14, que ilustra um recorte do conjunto de sintagmas entre os dois elementos da cadeia.

SN1 - ele e seus colegas
a - os doentes
b - a morte
c - Mattson
SN2 - Mattson e colegas

Figura 14 – Exemplo de sintagmas para a geração dos pares negativos.

Assim, o sintagma **Mattson e colegas** formará par negativo com **Mattson**, **a morte** e **os doentes**.

O segundo arquivo a ser processado é o arquivo de *phrases*, composto pelas informações dos *token*s que compõem as estruturas sintagmáticas e o seu núcleo. Nesse arquivo há uma marcação – *markable_ref* – que associa a estrutura ao sintagma dos *markables*. Essa marcação foi realizada apenas para os sintagmas pertencentes às cadeias de correferência. Assim, os demais sintagmas não foram associados. Através dessa marcação consegue-se obter o núcleo do sintagma. É através do núcleo que as comparações dos pares serão realizadas. Cabe destacar que essa marcação será utilizada para a geração dos pares positivos.

A Figura 15 mostra um trecho do arquivo de *phrases*.

```
Texto: CIENCIA_2000_6389-phrase.xml

<cesAna>

<struct from="t1" to="t17" type="phrase">
        <feat name="id" value="phr0" />
        <feat name="cat" value="s" />
</struct>
<struct from="t1" to="t17" type="phrase">
        <feat name="id" value="phr1" />
        <feat name="cat" value="fcl" />
        <feat name="function" value="STA" />
</struct>
<struct from="t4" to="t6" type="phrase">
        <feat name="id" value="phr2" />
        <feat name="cat" value="np" />
        <feat name="markable_ref" value="markable_2" />
        <feat name="head" value="t5" />
        <feat name="function" value="DP" />
</struct>
```

Figura 15 – Trecho do arquivo *phrases*.

Para os pares negativos será utilizada a informação da marcação *from*, obtida no arquivo de *markables*, como forma de obter a informação do núcleo. A escolha dessa associação se fez necessária pois nem todos os sintagmas pertencentes aos pares negativos possuem a marcação *markable_ref*. Optou-se em utilizar apenas a informação do *from*, pois nem sempre os valores do *from-to* do arquivo de *markables* correspondem aos do arquivo de *phrases*. Isso ocorre porque o arquivo de *markables* foi marcado manualmente enquanto a marcação do *phrases* foi feita automaticamente e sem revisão.

O número de pares negativos acaba sendo maior que o de pares positivos. Como forma de equalizar os exemplos, os pares positivos serão repetidos até alcançarem o valor aproximado do total de exemplos negativos. Esse procedimento é necessário para não ocasionar um aprendizado tendencioso pelo algoritmo de aprendizado na classificação das *features* para a criação da árvore de decisão, já que o algoritmo gera a árvore de acordo com a significância de cada *feature* [WIT05].

As informações sintáticas e semânticas são encontradas no arquivo de *pos*. Ou seja, nesse arquivo busca-se as informações referentes ao gênero, número e marcação semântica retornadas pelo analisador sintático.

De acordo com a Figura 16, o valor de gênero é obtido na etiqueta *gender*, o de número na etiqueta *number* e a informação semântica nas etiquetas *semantic* ou *complement*. Por intermédio da marcação *tokenref* é que será identificado o núcleo do sintagma, obtido no arquivo de *phrases* anteriormente explicado.

Salienta-se que essas informações serão a base de composição das *features*.

```
Texto: CIENCIA_2000_6389-pos.xml
<cesAna>
<struct type="pos">
        <feat name="id" value="pos1" />
        <feat name="class" value="art" />
        <feat name="tokenref" value="t1" />
        <feat name="canon" value="o" />
        <feat name="gender" value="F" />
        <feat name="number" value="S" />
</struct>
<struct type="pos">
        <feat name="id" value="pos2" />
        <feat name="class" value="n" />
        <feat name="tokenref" value="t2" />
        <feat name="canon" value="discussão" />
        <feat name="gender" value="F" />
        <feat name="number" value="S" />
        <feat name="semantic" value="talk" />
</struct>
<struct type="pos">
        <feat name="id" value="pos3" />
        <feat name="class" value="prp" />
        <feat name="tokenref" value="t3" />
        <feat name="canon" value="sobre" />
        <feat name="complement" value="np-close" />
</struct>
```

Figura 16 – Trecho do arquivo de *pos*.

A seguir são listadas as *features* que foram implementadas no *baseline*:

1. MNucleo_Semantico - se os núcleos do par são da mesma marcação semântica;
2. MGenero - se os núcleos do par apresentam o mesmo gênero (masculino/feminino);
3. MNumero – se os núcleos do par apresentam o mesmo número (singular/plural);
4. MNucleo - se os núcleos são idênticos;
5. MConj_Semantico - se os núcleos do par pertencem ao mesmo conjunto de marcação semântica.

Os valores atribuídos a todas as *features* foram definidos em *true* quando verdadeiro e *false* quando falso.

As *features* “MNucleo”, “MGenero” e “MNumero” são classificadas em nível gramatical e apresentam um papel importante na tarefa, pois estão diretamente associadas com a relação da anáfora com o antecedente.

Já as *features* “MNucleo_Semantico” e “MConj_Semantico” introduzem o conhecimento semântico na tarefa de resolução de correferência.

Para finalizar, o arquivo de *token*s é utilizado para comparação do núcleo do sintagma (Figura 17), localizado através da marcação ID, a qual possui armazenado o valor de *token*, que é a informação associada nos outros arquivos.

Texto: CIENCIA_2000_6389-*token*.xml

```
<cesAna>
<struct to="1" type="token" from="0">
        <feat name="id" value="t1" />
        <feat name="base" value="A" />
</struct>
<struct to="11" type="token" from="2">
        <feat name="id" value="t2" />
        <feat name="base" value="discussão" />
</struct>
<struct to="17" type="token" from="12">
        <feat name="id" value="t3" />
        <feat name="base" value="sobre" />
</struct>
<struct to="19" type="token" from="18">
        <feat name="id" value="t4" />
        <feat name="base" value="a" />
</struct>
<struct to="33" type="token" from="20">
        <feat name="id" value="t5" />
        <feat name="base" value="biotecnologia" />
</struct>
```

Figura 17 – Trecho do arquivo de *tokens.*

O sistema realiza a leitura dos arquivos do *corpus* acima listados. Primeiramente, faz-se a classificação dos pares positivos e negativos, armazenando como referência a identificação do *markable* e do valor da marcação do *from*. Em sequência, faz-se a busca no arquivo de *phrases* para descobrir o núcleo do sintagma. Com essa informação, busca-se o valor do *token* no arquivo de *tokens* e no arquivo de *phrases* os valores de

gênero, número e etiqueta semântica. Após, é realizado o processo de teste das *features*, armazenando o *markable* de referência e o valor de retorno das *features*.

Com as informações armazenadas, é gerado um arquivo no formato TXT para ser a base de entrada para a classe que gera o arquivo ARFF, sendo este o arquivo de entrada do Weka [WIT05], como já mencionado no Capítulo 4.

A Figura 18 ilustra um exemplo do arquivo TXT criado para os pares positivos.

Texto: CIENCIA_2000_6389.xml

markable_4 transgenia	, markable_64 engenharia_genética	, false , true , true , false , false
markable_95 agrônomo	, markable_11 Guerra	, false ,false , true , false , false
markable_95 agrônomo	, markable_22 Guerra	, false, false , true , false , false
markable_95 agrônomo	, markable_42 agrônomo	, true , true , true , true , true
markable_11 Guerra	, markable_22 Guerra	, true , true , true , true , true
markable_11 Guerra	, markable_42 agrônomo	, false, false , true , false , false
markable_22 Guerra	, markable_42 agrônomo	, false, false , true , false , false
markable_21 país	, markable_38 país	, true , true , true , true , true
markable_83 micropropagação	, markable_33 ela	, false , true , false ,false ,false
markable_89 presidente	, markable_61 Ele	, true , true , false , false , false
markable_99 Embrapa	, markable_57 empresa	, false , true , true , false , false

Legenda:

1ª coluna: referência do primeiro sintagma
2 ª coluna: núcleo do primeiro sintagma
3 ª coluna: referência do segundo sintagma
4 ª coluna: núcleo do segundo sintagma
5 ª coluna: *feature* MNucleo_Semantico
6 ª coluna: *feature* MGenero
7 ª coluna: *feature* MNumero
8 ª coluna: *feature* MNucleo
9 ª coluna: *feature* MConj_Semantico

Figura 18 – Trecho do arquivo de entrada para a API do Weka.

É indispensável definir corretamente as informações das colunas, pois a API para gerar o arquivo ARFF irá realizar a conversão do arquivo de acordo com as informações repassadas, sendo definidas pela identificação das colunas.

A Figura 19 apresenta uma visão detalhada do processo do protótipo, ilustrando o que foi descrito neste capítulo.

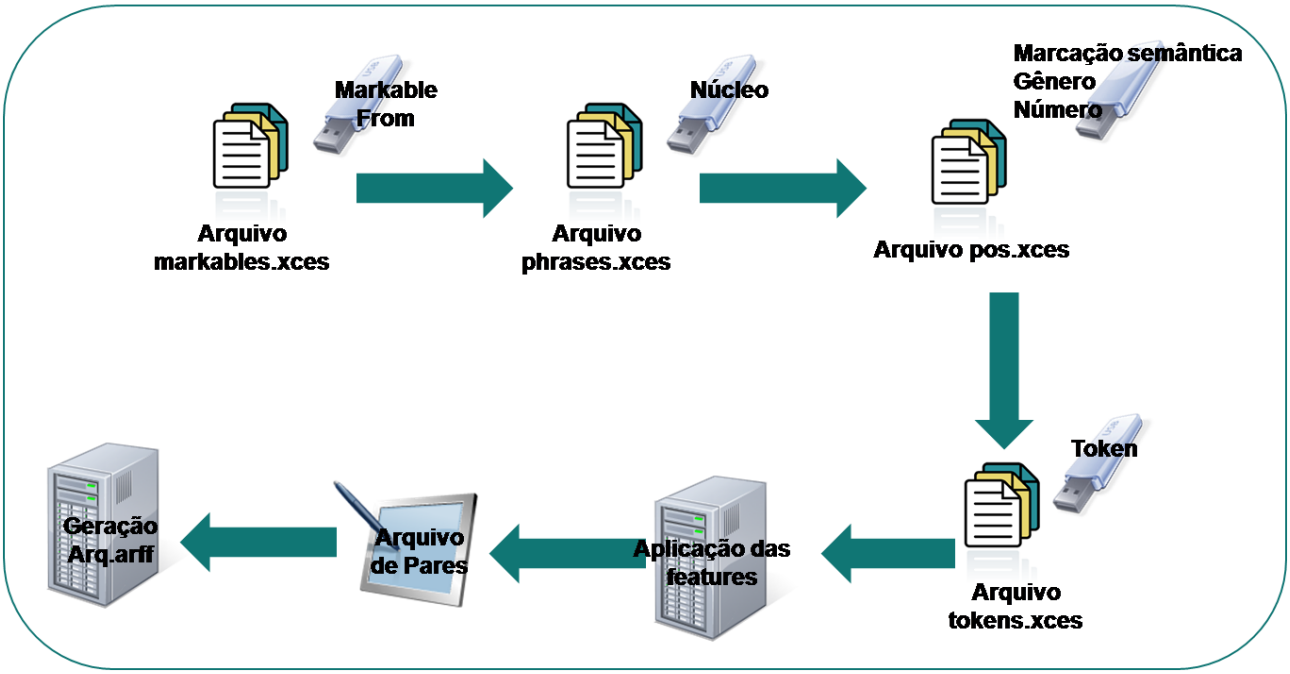

Figura 19 – Visão descritiva do protótipo.

5.2. Características do sistema de correferência Recorcaten

Essa versão do protótipo considera as categorias de ENs na escolha das cadeias. Assim, os pares gerados, tanto positivos quanto negativos, serão apenas daquelas cadeias pré-definidas através da seleção das categorias.

Para realizar a seleção das cadeias de acordo com as categorias, fez-se necessário alterar a classe que gera os pares para que considerasse a informação das categorias. Primeiro, identifica-se qual é a categoria da cadeia. Caso a mesma pertença à categoria definida (o Capítulo 6 irá apresentar detalhadamente as definições das categorias para o conjunto de experimentos), os pares de sintagmas da cadeia serão agrupados.

O método utilizado para agrupar os pares é o mesmo do *baseline*, ou seja, os pares positivos serão gerados entre todos os membros de uma cadeia (SN1-SN2, SN1-

SN3, SN2-SN3), enquanto que para os pares negativos o segundo sintagma da cadeia formará par com todos os sintagmas existentes entre ele e o primeiro sintagma da cadeia.

O arquivo TXT de saída do sistema continua no mesmo formato do *baseline*, com as mesmas *features* já citadas. Este arquivo será a entrada para a Java API Weka criar o arquivo ARFF.

5.3. Considerações sobre este capítulo

Apesar de demorada, a tarefa de levantamento de requisitos e modelagem do protótipo foi fundamental para que possibilitasse grande parte de reutilização das classes entre a primeira versão e a segunda.

A diferença entre o *baseline* e o Recorcaten refere-se ao processo de escolha das cadeias que serão utilizadas para gerar os pares, tanto os positivos como negativos. Essa modificação é responsável pela inclusão do conhecimento de mundo na tarefa de resolução de correferência.

Assim, o *baseline* compreende todas as cadeias disponíveis no *corpus*, enquanto que o Recorcaten faz uso apenas das cadeias associadas às categorias específicas.

Para escolha do conjunto de *features* considerou-se o retorno do algoritmo J48 do trabalho de Souza [SOU08] e as colocações dos trabalhos de Soon et al. [SOO01], Strube e Ponzetto [STR07] e Ng e Cardie [NG02]. Optou-se por um conjunto pequeno de *features*, mas que todas fossem importantes de forma a constarem na árvore de decisão, pois o intuito é realizar experimentos com determinadas categorias considerando sempre o mesmo conjunto de *features*.

A descrição desses experimentos será apresentada no capítulo a seguir.

6. ANÁLISE DOS RESULTADOS

Para realizar a avaliação dos resultados serão utilizadas as medidas de cobertura, precisão e medida-F. A medida-F é a média entre a cobertura e a precisão. A escolha dessas medidas foi por serem as medidas utilizadas nos trabalhos da área.

A partir do conjunto de 50 textos do *corpus*, foi utilizado um subconjunto composto por 31 textos dos 50, para a geração do arquivo de treino e o restante dos textos, 19, criou-se outro subconjunto para a geração do arquivo de teste. Essas definições de conjuntos serão utilizadas tanto para o *baseline* como para o Recorcaten.

O primeiro experimento foi gerar, tanto o arquivo de treino como o de teste, a partir do *baseline*.

O arquivo de treino foi composto por 6849 exemplos positivos (pares de sintagmas correferentes) e 7383 exemplos negativos (pares de sintagmas não correferentes). Esse treino serviu para induzir a árvore de decisão, a qual será utilizada para os testes futuros.

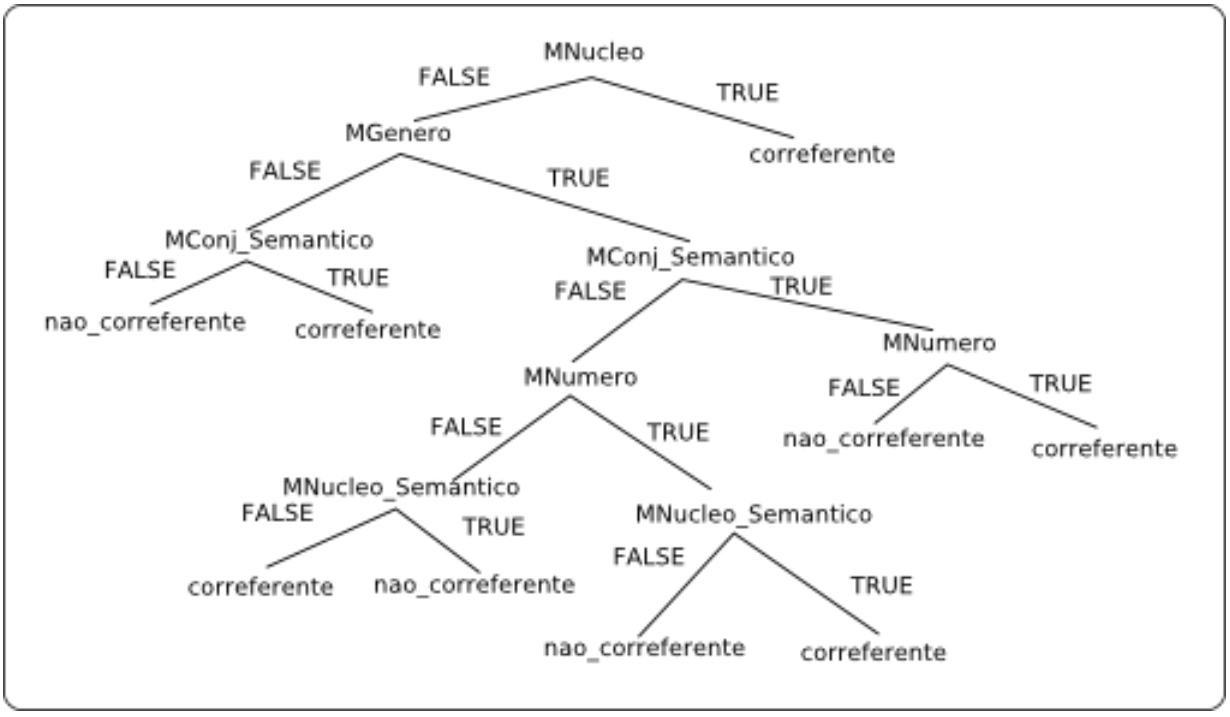

Figura 20 – Árvore de decisão gerada pelo algoritmo J48 para os testes.

Percebe-se pela árvore gerada na Figura 20 que todas as *features* escolhidas foram utilizadas pelo algoritmo de aprendizado de máquina, sendo que a *feature*

“MNucleo” é a raiz da árvore, o que significa ser a *feature* mais significativa, ou seja, a mais importante no processo de definir se um par é correferente ou não.

O arquivo de teste do *baseline* foi composto por um total de 1311 pares positivos e 1358 pares negativos, sendo que do total foram selecionados 540 corretos na classificação dos pares positivos e 771 dos pares negativos. Esse arquivo foi processado com base na árvore de decisão ilustrada na Figura 20. Seu retorno de acerto total foi de 53%. A Tabela 7 mostra as informações dos percentuais das medidas de cobertura, precisão e medida-F, obtidos na classificação dos pares positivos e negativos.

O segundo experimento foi gerar um arquivo de teste com base na segunda versão do sistema, a Recorcaten. Esta versão faz um filtro por categoria na geração dos pares, a qual foi aplicada ao subconjunto de teste. A categoria escolhida nesse experimento foi a categoria Pessoa. O arquivo teve um total de 196 pares positivos e 201 pares negativos. O percentual de retorno foi de 69%. Retomando, a base de treino é a mesma utilizada para o *baseline*, isto é, fez-se uso da árvore de decisão definida pelo arquivo de treino do *baseline*. A Tabela 8 apresenta os resultados.

Os próximos experimentos foram realizar os testes com as categorias Acontecimento, Local e Organização separadamente. A escolha por essas categorias deu-se em consideração a quantidade de cadeias do *corpus*. O Apêndice A lista a quantidade de cadeias por texto do *corpus*. O critério de seleção das categorias foi considerar a quantidade total de cadeias por categoria do *corpus.* Aquelas que apresentaram um valor maior que 20 cadeias foram selecionadas (não considerando a categoria Outro).

Notou-se que a categoria Local (Tabela 9) atingiu um percentual de 98% na classificação em geral, sendo que classificou todos os pares positivos corretamente. Já as categorias Acontecimento (Tabela 10) e Organização (Tabela 11) alcançaram um percentual de 44%, não conseguindo classificar nenhum par correferente corretamente.

Comparando apenas o percentual da medida-F deste resultado com o obtido com o arquivo de teste do *baseline*, percebe-se que ocorreu um aumento de acerto. Ao analisar mais detalhadamente o retorno do Weka, percebe-se que a classificação correta dos pares correferentes foi maior na versão do sistema Recorcaten do que do *baseline*. O sistema Recorcaten acertou 119 do total, ou seja, o percentual de 65%, enquanto que o *baseline* classificou 540, isto é, 46%.

A partir desse resultado, infere-se que o uso das categorias ajudou na tarefa, já que contribuiu para aumentar o percentual de classificação correta dos pares correferentes. Não obstante, também contribuiu para melhorar o acerto dos pares não correferentes, que passou de 48% no *baseline* para 71% no Recorcaten.

Ao observar as Tabelas, nota-se que tanto a medida de cobertura quanto a de precisão aumentaram, o que significa que ocorreu um aumento tanto na cobertura do *corpus* como na precisão de sua classificação e não apenas o aumento de uma única medida.

A Figura 23 mostra a saída de retorno do teste do *baseline* completa do Weka. Essa figura apresenta a árvore de decisão, a matriz de confusão e todas as informações que a ferramenta retorna para o processo realizado de classificação. Já a Figura 24 apresenta os resultados obtidos com a categoria Pessoa. Ambas as Figuras encontram-se no final deste capítulo.

Tabela 7 – Resultados dos dados de teste do *baseline*.

	Pares classificados corretos	**Cobertura**	**Precisão**	**Medida-F**
Pares Positivos	540 (total 1311)	41%	52%	46%
Pares Negativos	771 (total 1358)	64%	53%	58%

Tabela 8 – Resultados dos dados de teste da categoria Pessoa.

	Pares classificados corretos	**Cobertura**	**Precisão**	**Medida-F**
Pares Positivos	119 (total 196)	60%	72%	65%
Pares Negativos	155 (total 201)	77%	66%	71%

Tabela 9 – Resultados dos dados de teste da categoria Local.

	Pares classificados corretos	**Cobertura**	**Precisão**	**Medida-F**
Pares Positivos	88 (total 88)	100%	97%	98%
Pares Negativos	87 (total 90)	97%	100%	98%

Tabela 10 – Resultados dos dados de teste da categoria Acontecimento.

	Pares classificados corretos	**Cobertura**	**Precisão**	**Medida-F**
Pares Positivos	0 (total 40)	0%	0%	0%
Pares Negativos	36 (total 41)	88%	47%	61%

Tabela 11 – Resultados dos dados de teste da categoria Organização.

	Pares classificados corretos	**Cobertura**	**Precisão**	**Medida-F**
Pares Positivos	0 (total 54)	0%	0%	0%
Pares Negativos	56 (total 60)	93%	51%	66%

As Figuras 25, 26 e 27 (apresentadas no final deste capítulo) mostram o resultado completo do retorno do Weka para as categorias Acontecimento, Local e Organização, respectivamente.

O fato de não haver ocorrido a classificação correta de nenhum par correferente está relacionado com a marcação do conjunto de etiquetas semânticas da cadeia. Isto é, uma cadeia não apresenta o mesmo conjunto de etiquetas de uma categoria específica.

Ao analisar a árvore (Figura 20), percebe-se que a etiqueta semântica é determinante para definir se um par é correferente ou não. Observe o exemplo de uma cadeia na Figura 21.

o nascimento de a bezerra Vitoriosa

nascimento -> event

o resultado de um experimento realizado por a Embrapa (Empresa_Brasileira_de_Pesquisa_Agropecuária)

resultado -> ac

Legenda:
event – evento
ac – abstração contável (lazer, alternativa)

Figura 21 – Exemplo de cadeia de Acontecimento com etiquetas semânticas de outras categorias.

A marcação para a cadeia da Figura 21 é da categoria Acontecimento, contudo, o outro sintagma possui marcação semântica de outra categoria (ac). O par gerado nessa cadeia é correferente, no entanto, ao usar a árvore de decisão (Figura 20) para classificá-lo, o mesmo seria considerado não correferente (não pertence ao mesmo conjunto semântico da categoria Acontecimento, o que induz a classificação de não correferente na árvore).

Em outro experimento, agrupou-se todas as categorias selecionadas para o trabalho (Pessoa, Acontecimento, Local e Organização) para gerar um único arquivo de teste. Os resultados desse experimento foram de acerto na classificação correta dos pares de 70,25%, sendo que 64% foi para os pares correferentes. A Tabela 12 mostra os resultados das medidas de cobertura, precisão e medida-F desse experimento.

Tabela 12 – Resultados do conjunto de teste de Pessoa, Local, Acontecimento e Organização.

	Pares classificados corretos	**Cobertura**	**Precisão**	**Medida-F**
Pares Positivos	207 (total 378)	55%	78%	64%
Pares Negativos	334 (total 392)	85%	66%	75%

Observando o percentual de retorno dos testes de todas as categorias com a categoria Pessoa, a diferença é pequena. Nota-se que os pares positivos não obtiveram um ganho significativo de acerto, contudo, os pares negativos tiveram um percentual de 4% de aumento, o que acabou elevando a medida-F final do processo de classificação.

Ao realizar uma revisão manual do retorno do protótipo de ambas as versões foi possível constatar alguns problemas, principalmente no que se refere ao processo de geração dos pares.

Um deles é a constatação de que alguns sintagmas do *corpus* não possuem a marcação do núcleo e, por consequência, acaba-se descartando a geração de um par, assim como de uma cadeia, principalmente na versão Recorcaten, já que a seleção da cadeia ocorre através da categoria da mesma.

Esse problema ocasionou um impacto na quantidade dos pares considerando as categorias, tendo em vista que a quantidade de cadeias por categoria não possui um valor

elevado, o que acabou gerando um conjunto pequeno de exemplos tanto para os pares positivos como para os negativos.

Como já mencionado anteriormente, a questão de uma cadeia possuir marcações de etiquetas de categorias diferentes proporcionou alterações no processo de classificação dos pares positivos. Uma das alterações constatadas foi ao analisar a árvore de decisão (Figura 20), já que uma de suas ramificações, a *feature* “MNucleo_Semantico”, se estiver com o valor *true*, irá informar que o par não é correferente.

6.1. Considerações sobre este capítulo

De acordo com os resultados, observa-se que o uso de categorias proporcionou uma melhora no percentual de acerto ao definir se um par é anafórico ou não.

Com isso, pode-se inferir que o uso de conhecimento semântico na tarefa de resolução de correferência é relevante para o processo de resolução, retornando um maior número de acerto na classificação correta dos pares.

Através dos experimentos também foi possível constatar a importância do conhecimento de mundo para a tarefa. Como visto, algumas categorias (Acontecimento e Organização) não apresentaram um retorno satisfatório na classificação dos pares correferentes porque o processo de desambiguação não foi realizado da forma correta. Algumas marcações de etiquetas dessas categorias foram utilizadas para identificar outras categorias, como a de Pessoa.

A Figura 22 ilustra essa colocação. Mesmo apresentando sintagmas com núcleos iguais (Nasa), o analisador sintático informou etiquetas semânticas de categorias diferentes (neste caso, da categoria Organização (org) e Pessoa (hum), respectivamente).

a Nasa

Nasa -> org

a agência espacial Americana

agência -> HH

a Nasa

Nasa -> org

a Nasa

Nasa -> hum

a Nasa

Nasa -> org

Legenda:
org – organização
HH – grupo de pessoas
hum – pessoa

Figura 22 – Exemplo de cadeia com marcação de várias categorias.

Assim, ressalta-se a necessidade de bases de conhecimento mais estruturadas e com sinônimos, como a WordNet, para complementar e apoiar a tarefa de resolução de correferência.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Correferencia
Instances:    14232
Attributes:   6
              MNucleo_Semantico
              MGenero
              MNumero
              MNucleo
              MConj_Semantico
              Classe
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------
MNucleo = false
|   MGenero = false
|   |   MConj_Semantico = false: nao_correferente (6377.0/1716.0)
|   |   MConj_Semantico = true: correferente (200.0/50.0)
|   MGenero = true
|   |   MConj_Semantico = false
|   |   |   MNumero = false
|   |   |   |   MNucleo_Semantico = false: correferente (2226.0/966.0)
|   |   |   |   MNucleo_Semantico = true: nao_correferente (132.0/48.0)
|   |   |   MNumero = true
|   |   |   |   MNucleo_Semantico = false: nao_correferente (2619.0/1218.0)
|   |   |   |   MNucleo_Semantico = true: correferente (351.0/144.0)
|   |   MConj_Semantico = true
|   |   |   MNumero = false: nao_correferente (17.0/3.0)
|   |   |   MNumero = true: correferente (186.0/33.0)
MNucleo = true: correferente (2124.0/30.0)

Number of Leaves  :       9
Size of the tree :        17

Time taken to build model: 1.73 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        1417               53.091  %
Incorrectly Classified Instances      1252               46.909  %
Kappa statistic                          0.0579
Mean absolute error                      0.4539
Root mean squared error                  0.5292
Relative absolute error                 90.8449 %
Root relative squared error            105.8417 %
Total Number of Instances             2669

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
 0.412    0.354      0.529      0.412    0.463        0.6     correferente
 0.646    0.588      0.532      0.646    0.583        0.6     nao_correferente

=== Confusion Matrix ===

   a   b   <-- classified as
 540 771 |   a = correferente
 481 877 |   b = nao_correferente
```

Figura 23 – Saída do Weka para o experimento de teste do *baseline*.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Correferencia
Instances:    14232
Attributes:   6
              MNucleo_Semantico
              MGenero
              MNumero
              MNucleo
              MConj_Semantico
              Classe
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------
MNucleo = false
|   MGenero = false
|   |   MConj_Semantico = false: nao_correferente (6377.0/1716.0)
|   |   MConj_Semantico = true: correferente (200.0/50.0)
|   MGenero = true
|   |   MConj_Semantico = false
|   |   |   MNumero = false
|   |   |   |   MNucleo_Semantico = false: correferente (2226.0/966.0)
|   |   |   |   MNucleo_Semantico = true: nao_correferente (132.0/48.0)
|   |   |   MNumero = true
|   |   |   |   MNucleo_Semantico = false: nao_correferente (2619.0/1218.0)
|   |   |   |   MNucleo_Semantico = true: correferente (351.0/144.0)
|   |   MConj_Semantico = true
|   |   |   MNumero = false: nao_correferente (17.0/3.0)
|   |   |   MNumero = true: correferente (186.0/33.0)
MNucleo = true: correferente (2124.0/30.0)

Number of Leaves  :       9
Size of the tree :        17

Time taken to build model: 1.69 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         274               69.0176 %
Incorrectly Classified Instances       123               30.9824 %
Kappa statistic                          0.379
Mean absolute error                      0.4064
Root mean squared error                  0.4584
Relative absolute error                 81.3124 %
Root relative squared error             91.6677 %
Total Number of Instances              397

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.607    0.229      0.721      0.607      0.659      0.713    correferente
                 0.771    0.393      0.668      0.771      0.716      0.713    nao_correferente

=== Confusion Matrix ===

   a   b   <-- classified as
 119  77 |   a = correferente
  46 155 |   b = nao_correferente
```

Figura 24 – Saída do Weka para o experimento utilizando a categoria Pessoa.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Correferencia
Instances:    14232
Attributes:   6
              MNucleo_Semantico
              MGenero
              MNumero
              MNucleo
              MConj_Semantico
              Classe
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------
MNucleo = false
|   MGenero = false
|   |   MConj_Semantico = false: nao_correferente (6377.0/1716.0)
|   |   MConj_Semantico = true: correferente (200.0/50.0)
|   MGenero = true
|   |   MConj_Semantico = false
|   |   |   MNumero = false
|   |   |   |   MNucleo_Semantico = false: correferente (2226.0/966.0)
|   |   |   |   MNucleo_Semantico = true: nao_correferente (132.0/48.0)
|   |   |   MNumero = true
|   |   |   |   MNucleo_Semantico = false: nao_correferente (2619.0/1218.0)
|   |   |   |   MNucleo_Semantico = true: correferente (351.0/144.0)
|   |   MConj_Semantico = true
|   |   |   MNumero = false: nao_correferente (17.0/3.0)
|   |   |   MNumero = true: correferente (186.0/33.0)
MNucleo = true: correferente (2124.0/30.0)

Number of Leaves  :       9
Size of the tree :        17

Time taken to build model: 1.26 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          36               44.4444 %
Incorrectly Classified Instances        45               55.5556 %
Kappa statistic                          -0.1233
Mean absolute error                       0.4834
Root mean squared error                   0.5163
Relative absolute error                  96.7169 %
Root relative squared error             103.2411 %
Total Number of Instances                81

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
 0          0.122       0          0          0          0.579     correferente
 0.878      1           0.474      0.878      0.615      0.579     nao_correferente

=== Confusion Matrix ===

  a  b   <-- classified as
  0 40 |  a = correferente
  5 36 |  b = nao_correferente
```

Figura 25 – Saída do Weka para o experimento utilizando a categoria Acontecimento.

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Correferencia
Instances:    14232
Attributes:   6
              MNucleo_Semantico
              MGenero
              MNumero
              MNucleo
              MConj_Semantico
              Classe
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------
MNucleo = false
|   MGenero = false
|   |   MConj_Semantico = false: nao_correferente (6377.0/1716.0)
|   |   MConj_Semantico = true: correferente (200.0/50.0)
|   MGenero = true
|   |   MConj_Semantico = false
|   |   |   MNumero = false
|   |   |   |   MNucleo_Semantico = false: correferente (2226.0/966.0)
|   |   |   |   MNucleo_Semantico = true: nao_correferente (132.0/48.0)
|   |   |   MNumero = true
|   |   |   |   MNucleo_Semantico = false: nao_correferente (2619.0/1218.0)
|   |   |   |   MNucleo_Semantico = true: correferente (351.0/144.0)
|   |   MConj_Semantico = true
|   |   |   MNumero = false: nao_correferente (17.0/3.0)
|   |   |   MNumero = true: correferente (186.0/33.0)
MNucleo = true: correferente (2124.0/30.0)

Number of Leaves  :       9
Size of the tree :        17

Time taken to build model: 1.34 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         175               98.3146 %
Incorrectly Classified Instances         3                1.6854 %
Kappa statistic                          0.9663
Mean absolute error                      0.224
Root mean squared error                  0.2614
Relative absolute error                 44.826  %
Root relative squared error             52.2639 %
Total Number of Instances              178

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
               1          0.033      0.967         1          0.983      0.991    correferente
               0.967          0          1          0.967      0.983      0.991    nao_correferente

=== Confusion Matrix ===

  a  b   <-- classified as
 88  0 |  a = correferente
  3 87 |  b = nao_correferente
```

Figura 26 – Saída do Weka para o experimento utilizando a categoria Local.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Correferencia
Instances:    14232
Attributes:   6
              MNucleo_Semantico
              MGenero
              MNumero
              MNucleo
              MConj_Semantico
              Classe
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------
MNucleo = false
|   MGenero = false
|   |   MConj_Semantico = false: nao_correferente (6377.0/1716.0)
|   |   MConj_Semantico = true: correferente (200.0/50.0)
|   MGenero = true
|   |   MConj_Semantico = false
|   |   |   MNumero = false
|   |   |   |   MNucleo_Semantico = false: correferente (2226.0/966.0)
|   |   |   |   MNucleo_Semantico = true: nao_correferente (132.0/48.0)
|   |   |   MNumero = true
|   |   |   |   MNucleo_Semantico = false: nao_correferente (2619.0/1218.0)
|   |   |   |   MNucleo_Semantico = true: correferente (351.0/144.0)
|   |   MConj_Semantico = true
|   |   |   MNumero = false: nao_correferente (17.0/3.0)
|   |   |   MNumero = true: correferente (186.0/33.0)
MNucleo = true: correferente (2124.0/30.0)

Number of Leaves  :       9
Size of the tree :        17

Time taken to build model: 1.26 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          56               49.1228 %
Incorrectly Classified Instances        58               50.8772 %
Kappa statistic                          -0.0699
Mean absolute error                       0.5291
Root mean squared error                   0.5676
Relative absolute error                 106.027  %
Root relative squared error             113.6625 %
Total Number of Instances               114

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.067       0          0          0          0.3      correferente
                 0.933     1           0.509      0.933      0.659      0.3      nao_correferente

=== Confusion Matrix ===

   a  b   <-- classified as
   0 54 |  a = correferente
   4 56 |  b = nao_correferente
```

Figura 27 – Saída do Weka para o experimento utilizando a categoria Organização.

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Correferencia
Instances:    14232
Attributes:   6
              MNucleo_Semantico
              MGenero
              MNumero
              MNucleo
              MConj_Semantico
              Classe
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------
MNucleo = false
|   MGenero = false
|   |   MConj_Semantico = false: nao_correferente (6377.0/1716.0)
|   |   MConj_Semantico = true: correferente (200.0/50.0)
|   MGenero = true
|   |   MConj_Semantico = false
|   |   |   MNumero = false
|   |   |   |   MNucleo_Semantico = false: correferente (2226.0/966.0)
|   |   |   |   MNucleo_Semantico = true: nao_correferente (132.0/48.0)
|   |   |   MNumero = true
|   |   |   |   MNucleo_Semantico = false: nao_correferente (2619.0/1218.0)
|   |   |   |   MNucleo_Semantico = true: correferente (351.0/144.0)
|   |   MConj_Semantico = true
|   |   |   MNumero = false: nao_correferente (17.0/3.0)
|   |   |   MNumero = true: correferente (186.0/33.0)
MNucleo = true: correferente (2124.0/30.0)

Number of Leaves  :     9
Size of the tree :      17

Time taken to build model: 1.26 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         541               70.2597 %
Incorrectly Classified Instances       229               29.7403 %
Kappa statistic                          0.4018
Mean absolute error                      0.3905
Root mean squared error                  0.4471
Relative absolute error                 78.1508 %
Root relative squared error             89.4177 %
Total Number of Instances              770

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.548    0.148      0.781     0.548     0.644      0.723    correferente
                 0.852    0.452      0.661     0.852     0.745      0.723    nao_correferente

=== Confusion Matrix ===

   a   b   <-- classified as
 207 171 |   a = correferente
  58 334 |   b = nao_correferente
```

Figura 28 – Saída do Weka para o experimento utilizando o conjunto de categorias escolhidas.

7. CONSIDERAÇÕES FINAIS

Apesar de ser uma tarefa importante para a área de PLN, ainda há muitas pesquisas a serem realizadas para a resolução de correferência, em especial, tratando da língua portuguesa.

Trabalhos recentes têm empregado cada vez mais recursos semânticos para alimentar os sistemas de resolução de correferência. Porém, enquanto hoje estão disponíveis para a língua inglesa recursos que ajudam na tarefa, como a WordNet, a mesma disponibilidade não existe para o português. Um recurso alternativo para inclusão de informação semântica para o português é através do uso do analisador PALAVRAS e sua anotação semântica.

Nesta dissertação exploramos a anotação semântica e as categorias semânticas como uma informação adicional ao processo de resolução de correferência. Como método de avaliação, optamos realizar experimentos com diferentes categorias semânticas.

Os resultados obtidos mostraram que o uso das categorias proporciona uma melhora no processo de classificação dos pares de expressões anafóricas. Isso demonstra a importância do conhecimento de mundo para realizar a tarefa de resolução de correferência com resultados mais robustos.

A partir da análise dos resultados, infere-se que o comportamento do analisador PALAVRAS na etiquetação das categorias Pessoa e Local foi mais estável, atribuindo etiquetas semânticas mais coesas nessas categorias. Já para as categorias Acontecimento e Organização, o mesmo não ocorreu. Além do conjunto de etiquetas ser mais diferenciado, o erro do analisador na classificação também ocorre e tem influência nos resultados. Um exemplo disso é a atribuição da etiqueta “hum” (Pessoa) para Nasa (Figura 22), que na verdade é uma Organização.

Verificando o contexto que a expressão se encontrava no texto, constatou-se que realmente a marcação da etiqueta “hum” foi equivocada. Como uma revisão na marcação semântica do *corpus* não foi realizada não há como mensurar o percentual total de marcações incorretas. Contudo, num levantamento parcial, esses erros não são predominantes.

7.1. Contribuições

Pode-se considerar como contribuições deste trabalho:

- Estudo da evolução das *features* referente à tarefa de resolução de correferência;
- Desenvolvimento de um protótipo para experimentação de resolução de correferência;
- Realização de experimentos considerando a proposta deste trabalho (o uso das categorias);
- Análise e discussão dos resultados;
- Produção de um artigo no CelSul (Círculo de Estudos Linguísticos do Sul) intitulado “A dificuldade da tarefa de resolução de correferência”.

7.2. Limitações

Um dos pontos limitadores encontrado é o tamanho do *corpus* utilizado nos experimentos deste trabalho. A limitação de recursos para a língua portuguesa dificulta o avanço das pesquisas na área.

Outra consideração é a não revisão da anotação do analisador sintático PALAVRAS no *corpus* escolhido. Uma revisão com o levantamento do que há de marcação incorreta auxiliaria na análise dos resultados obtidos, principalmente em relação ao uso das categorias. Com isso, uma avaliação mais precisa do sistema seria possível.

7.3. Trabalhos futuros

Após a realização deste trabalho identificou-se a possibilidade dos seguintes trabalhos futuros:

- Realizar o processo de identificação das ENs através de palavras vizinhas e/ou verbos, não se limitando ao analisador PALAVRAS;

- Acrescentar ao processo de resolução o uso de outros recursos, tais como dicionários e Wikipédia, por exemplo;
- Inclusão de novas *features* a partir da inserção dos novos recursos;
- Adaptação do protótipo para realizar a tarefa para a língua inglesa.

REFERÊNCIAS BIBLIOGRÁFICAS

[ACE08a] ACE. “Automatic Content Extraction 2008 Evaluation Plan (ACE08) - Assessment of detection and recognition of entities and relations within and across documents”. Capturado em: http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf, Novembro 2009.

[ACE08b] ACE. “ACE (Automatic Content Extraction) English Annotation Guidelines for Entities”. Capturado em: http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.1.pdf, Novembro 2009.

[ABR05] Abreu, S. C. de. “Análise de expressões referenciais em corpus anotado da Língua Portuguesa”, Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, Unisinos, 2005, 103p.

[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, 505p.

[BIC09] Bick, E. “Visl - Portuguese”. Capturado em: http://visl.sdu.dk/visl/pt/info/portsymbol.html, Dezembro 2009.

[COL07] Collovini, S. S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Rino L. H. M.; 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, pp. 1605-1614.

[CHA07] Chaves, A.; Rino, L. H. M. “A resolução de pronomes anafóricos do português com base em heurísticas que apontam o antecedente”. In: VI Congresso de Pós-Graduação da UFSCar. Proceedings of VI Congresso de Pós-Graduação da UFSCar, São Carlos, São Paulo, 2007, v. 2, pp. 1272-1273.

[CHA08] Chaves, A. R.; Rino, L. H. ”The mitkov algorithm for anaphora resolution in portuguese”. In: 8th International Conference on Computational Processing of the Portuguese Language. Proceedings of Propor, Portugal, 2008, pp. 51-60.

[COE05] Coelho, T.; Carvalho, A. “Uma adaptação do algoritmo de Lappin e Leass para resolução de anáforas em português”. In: III Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXV Congresso da SBC,São Leopoldo, 2005, pp. 2069-2078.

[HAR08] HAREM. “Reconhecimento de entidades mencionadas em português”. Capturado em: http://www.linguateca.pt/HAREM/, Dezembro 2009.

[KOC04] Koch, I. G. V. “Lingüística Aplicada ao Português: Sintaxe”. Ed. Cortez. 2004, 80p.

[MIT97] Mitchell, T. “Machine Learning”. Ed McGraw-Hill Companies, 1997, 414p.

[NG02] Ng, V.; Cardie, C.. “Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution”. In: 19th International Conference on Computational Linguistics. Proceedings of the 19th International Conference on Computational Linguistics, Taipei, 2002, v. 1, pp. 730-736.

[NG04] Ng, V. “Learning noun phrases anaphoricity to improve coreference resolution: Issues in representation and optimization”. In: 42nd Annual Meeting on Association for Computational Linguistics. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, 2004, pp. 151-158.

[NG05] Ng, V. “Supervised ranking for pronoun resolution: Some recent improvements”. In: AAAI Conference on Artificial Intelligence. Proceedings of the 20th National Conference on Artificial Intelligence, Pittsburgh, 2005, v. 3, pp. 1081-1086.

[NG07] Ng, V. “Semantic class induction and coreference resolution”. In: 45th Annual Meeting of the Association for Computional Linguistics. Proceedings of ACL, Prague, 2007, pp. 536-543.

[NG09] Ng, V. “Graph-Cut-Based anaphoricity determination for coreference resolution”. In: Human Language Techologies. Proceedings of Human Language Tecnologies: The 2009 Annual Conference of the North American Chapter of Association for Computional Lingustics, Boulder, 2009, pp. 575-583.

[SAN07] Santos, D., Cardoso, N.: “Directivas para a identificação e classificação semântica na colecção dourada do HAREM”. In: Reconhecimento de entidades mencionadas em português. Documentação e actas do HAREM, a primeira avaliação conjunta na área, 2007, v. 1, pp. 239-244.

[SOO01] Soon, W. M.; Ng, H. T.; Lin, D. C. Y. “A machine learning approach to coreference resolution of noun phrases”. In: Computational Linguistics, 2001, v. 27, number 4, pp. 521-544.

[SOU08] de Souza, J. G. C.; Nunes, P. G.; Vieira, R. “Learning coreference resolution for portuguese texts”. In: 8th International Conference on Computational Processing of the Portuguese Language. Proceedings of Propor, Portugal, 2008, pp. 153-162.

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

[WIT05] Witten, I. H.; Frank, E. “Data Mining: Practical machine learning tools and techniques”. Ed. Morgan, 2005, 525p.

ANEXO A - TABELA DE *FEATURES* DE NG E CARDIE[NG02]

Lexical	
Uppercase	se SNj for escrito em maiúsculo recebe o valor verdadeiro, senão o valor falso;
Conj	se SNj for conjunção recebe o valor verdadeiro, senão falso;
Gramatical (tipo de SN)	
Indefinitive	verdadeiro se SNj começar com “a” ou “na”, caso contrário falso;
quantified	se SNj começar por “every”, “some”, “all”, “most”, “many”, “much”, “few” ou “ none” recebe o valor verdadeiro, senão falso;
article	recebe o valor definido se o SNj for um SN de nido, quantitativo se for um SN quantidade ou indefinido
possessive	recebe o valor verdadeiro se SNj começar por um pronome possessivo ou sintagma nominal, caso contrário recebe o valor falso;
bare_singular	se SNj for singular e não começar por artigo atribui-se o valor verdadeiro, senão falso;
bare_plural	se SNj for plural e não começar por artigo atribui-se o valor verdadeiro, caso contrário falso;
Gramatical (relacionamento de SN)	
prednom	se SNj é o primeiro de dois SNs em uma construção de predicado nominal atribui-se o valor verdadeiro, caso contrário o valor falso;
modifier	se SNj for pré-modificador recebe o valor verdadeiro, senão falso;
postmodified	se SNj for pós-modificador por uma oração relativa recebe o valor verdadeiro, senão falso;
contains_pn	se SNj não for nome próprio mas contiver um nome próprio recebe o valor verdadeiro, caso contrário o valor falso;
special_nouns	Se o núcleo de SNj é comparativo ou SNj for pré-modificador de superlativo recebe o valor verdadeiro, senão falso;
Gramatical (padrão sintático)	
the_n	caso SNj começar por "the" seguido por um substantivo comum atribui-se o valor verdadeiro, senão falso;
the_2n	caso SNj começar por "the" seguido por dois substantivos comum atribui-se o valor verdadeiro, senão falso;
the_pn	caso SNj começar por "the" seguido por um nome próprio atribui-se o valor verdadeiro, caso contrário falso;
the_pn_n	caso SNj começar por "the" seguido por um nome próprio e um substantivo comum atribui-se o valor verdadeiro, ao contrário falso;
the_adj_n	caso SNj começar por "the" seguido por um adjetivo e um substantivo comum atribui-se o valor verdadeiro, ao contrário falso;
the_num_n	caso SNj começar por "the" seguido por um número cardinal e um substantivo comum atribui-se o valor verdadeiro, senão falso;
the_ne	caso SNj começar por "the" seguido por uma entidade nomeada atribui-se o valor verdadeiro, senão falso;
the_sing_n	caso SNj começar por "the" seguido por um sintagma nominal não contendo nome próprio atribui-se o valor verdadeiro, ao contrário falso;
Semântico	
post	se for post (não se encontrou referência sobre essa feature no trabalho) será atribuído o valor verdadeiro, senão falso;
subclass	recebe o valor verdadeiro se existir um SN anterior ao SNj e SNi e apresentem uma relação de antecessor-descendente na WordNet, senão falso;
title	verdadeiro se for um título de pessoa, senão falso;
Posicional	
rst_sent	se SNj for a primeira sentença do corpo do texto recebe o valor verdadeiro, senão falso;
rst_para	se SNj for o primeiro parágrafo do corpo do texto recebe o valor verdadeiro, senão falso;
header	caso SNj seja o cabeçalho do texto atribui-se o valor verdadeiro, senão falso.

APÊNDICE A - TABELA DAS CATEGORIAS POR CADEIA POR TEXTO

Texto	Categoria						
	Pessoa	Organização	Acontecimento	Local	Obra	Coisa	Outro
CIENCIA_2000_17082	2	1		3			4
CIENCIA_2000_17088	2		2	2			5
CIENCIA_2000_17101	2	1	2	2			10
CIENCIA_2000_17108							9
CIENCIA_2000_17109	2		1				9
CIENCIA_2000_17112	3	1		2			3
CIENCIA_2000_17113	1						15
CIENCIA_2000_6380	2		1				7
CIENCIA_2000_6381	1	2		1			7
CIENCIA_2000_6389	2	1		1			4
CIENCIA_2000_6391		2		1	1	1	1
CIENCIA_2001_19858	2		2	2			6
CIENCIA_2001_6406	2						5
CIENCIA_2001_6410	1						7
CIENCIA_2001_6414	1			2			5
CIENCIA_2001_6416	2						7
CIENCIA_2001_6423	1		1				1
CIENCIA_2002_22005	1			2			9
CIENCIA_2002_22010	1			3			6
CIENCIA_2002_22015	2		1				13
CIENCIA_2002_22023	2		1	1			8
CIENCIA_2002_22027	3		1				18
CIENCIA_2002_22029	3	4					12
CIENCIA_2002_6441							5
CIENCIA_2003_24212	2	3		2			13
CIENCIA_2003_24219	2	2	1	2			6
CIENCIA_2003_24226	2	1	1	1			10
CIENCIA_2003_6457		1	1	2			5
CIENCIA_2003_6465		1			2		8
CIENCIA_2003_6472	1		1				2
CIENCIA_2004_26415	2						4
CIENCIA_2004_26417	2						12
CIENCIA_2004_26423	3		2	2			17
CIENCIA_2004_26425	5	2	4				10
CIENCIA_2004_6480	2	1	1	1			5
CIENCIA_2004_6488			1				4
CIENCIA_2004_6494	2			3			3
CIENCIA_2005_28743	2						7
CIENCIA_2005_28747	3	1					7
CIENCIA_2005_28752	2	2	1	1	3		5
CIENCIA_2005_28754	3	1	1	3	2		5
CIENCIA_2005_28755	2	1		1			10
CIENCIA_2005_28756	2			1			10
CIENCIA_2005_28764	2						10
CIENCIA_2005_28766	2			1	2		18
CIENCIA_2005_28774	2	1	1	1			16
CIENCIA_2005_6507	1		1	1	1		2
CIENCIA_2005_6514			1	3	1		3
CIENCIA_2005_6515	1	1		3	1		5
CIENCIA_2005_6518	1	1					8
Total	**84**	**31**	**29**	**50**	**13**	**1**	**381**