Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado

Jefferson Fontinele da Silva

Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado

Jefferson Fontinele da Silva

Orientador: ***Prof. Dr. João Luís Garcia Rosa***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*.

USP – São Carlos
Julho/2011

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

SS586rr

Silva, Jefferson Fontinele da
Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado / Jefferson Fontinele da Silva; orientador João Luis Garcia Rosa -- São Carlos, 2011.
120 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011.

1. Processamento de Lingua Natural. 2. Correferência (Linguística). 3. Aprendizado de máquina. I. Rosa, João Luis Garcia , orient. II. Título.

Aos meus pais, com amor, pelo incansável apoio ao longo de todo período dos meus estudos.

Não vá aonde o caminho possa levá-lo.
Ao invés, vá aonde não ha caminho e deixe um rastro.
(Ralph Waldo Emerson)

Agradecimentos

Aos meus pais José Ribamar e Maria Inês por terem me guiado por todo esse caminho.

As minhas avós Francisca (*in memoriam*) e Cecília (*in memoriam*) por sempre acreditarem no neto.

As minhas irmãs Jéssica, Cláudia, e Lidinalva pelo apoio.

A minha namorada Cristina, por todo seu apoio e compreensão durante essa jornada.

Ao João Luís, meu orientador, pelo grande otimismo e paciência com que me orientou durante esses anos, e pela amizade.

Aos colegas do NILC pelas amizade e por compartilharem esses anos de estudo. Em especial a Carol, Claudinha, Erick, Fernando, Jean, Lúcia, Marcelo e Paula pela ajuda no desenvolvimento deste trabalho.

A Capes pelo apoio financeiro e a USP e NILC pelas instalações.

A todos que de alguma forma influenciaram no caminho para chegar ao fim deste trabalho.

Muito obrigado.

Resumo

Um dos problemas encontrados em sistemas de Processamento de Línguas Naturais (PLN) é a dificuldade de se identificar que elementos textuais referem-se à mesma entidade. Esse fenômeno, no qual o conjunto de elementos textuais remete a uma mesma entidade, é denominado de correferência. Sistemas de resolução de correferência podem melhorar o desempenho de diversas aplicações do PLN, como: sumarização, extração de informação, sistemas de perguntas e respostas. Recentemente, pesquisas em PLN têm explorado a possibilidade de identificar os elementos correferentes em múltiplos documentos. Neste contexto, este trabalho tem como foco o desenvolvimento de um método aprendizado não supervisionado para resolução de correferência em múltiplos documentos, utilizando como língua-alvo o português. Não se conhece, até o momento, nenhum sistema com essa finalidade para o português. Os resultados dos experimentos feitos com o sistema sugerem que o método desenvolvido é superior a métodos baseados em concordância de cadeias de caracteres.

Palavras-chave: Processamento de Línguas Naturais, correferência, múltiplos documentos, aprendizado não supervisionado.

Abstract

One of the problems found in Natural Language Processing (NLP) systems is the difficulty of identifying textual elements that refer to the same entity. This phenomenon, in which the set of textual elements refers to a single entity, is called coreference. Coreference resolution systems can improve the performance of various NLP applications, such as automatic summarization, information extraction systems, question answering systems. Recently, research in NLP has explored the possibility of identifying the coreferent elements in multiple documents. In this context, this work focuses on the development of an unsupervised method for coreference resolution in multiple documents, using Portuguese as the target language. Until now, it is not known any system for this purpose for the Portuguese. The results of the experiments with the system suggest that the developed method is superior to methods based on string matching.

Keywords: Natural Language Processing, coreference, multiple documents, unsupervised learning.

Sumário

Lista de Abreviaturas

AM Aprendizado de Máquina

IDC *Information Data Center*

MUC *Message Understanding Conference*

PLN Processamento de Línguas Naturais

IA Inteligência Artificial

VSM *Vector Space Model*

tf-idf *term frequency – inverse document frequency*

XML *Extensible Markup Language*

SVM *Support Vector Machine*

RST *Rhetorical Structure Theory*

Lista de Figuras

Lista de Tabelas

CAPÍTULO

1

Introdução

Existe uma grande disponibilidade e quantidade de informação que a sociedade moderna produz. Segundo o *Information Data Center* (IDC) (Gantz et al., 2008), é estimado que no ano 2011 o volume de dados produzido chegue a 1.800 *exabytes*, representando um aumento de 10 vezes se comparado ao volume de informação do ano 2006. A necessidade de se ter acesso rápido e eficiente a esse volume de conteúdo, bem como a urgência de identificação e processamento das informações têm gerado um ambiente adequado para o desenvolvimento de aplicações do Processamento de Línguas Naturais (PLN).

O PLN é uma subárea da Inteligência Artificial (IA), que compreende técnicas e recursos para tratar a língua natural automaticamente. As pesquisas em PLN têm produzido diversas técnicas com o objetivo de encontrar soluções para os vários problemas que surgem das aplicações. Essas aplicações facilitam o processamento do volume de informação disponível. Entretanto, vários desafios estão envolvidos na construção de aplicações de PLN que sejam capazes de processar essas informações satisfatoriamente. Um dos principais problemas é a resolução de correferência.

Correferência é um fenômeno que ocorre quando duas ou mais menções no texto referem-se a uma mesma entidade no mundo real (Mitkov, 2002). O conjunto das menções a uma mesma entidade no texto é denominado de cadeia de correferência. A identificação das cadeias de correferência pode melhorar o desempenho de várias aplicações, como: extração de informação,

tradução automática, sumarização automática e sistemas de perguntas e respostas (Baron e Freedman, 2008). Apresentam-se, no exemplo 1.1, as seguintes sentenças:

(1.1) Mário ganhou mais uma corrida de *kart*. O piloto foi o maior campeão de todos os tempos.

Observa-se no exemplo 1.1 que a informação contida na segunda sentença, ou seja, "Mário foi o maior campeão de todos os tempos" só é possível de ser obtida se o leitor compreender que "o piloto" na segunda sentença também refere-se a "Mário". Por exemplo, em um sistema de perguntas e respostas, um pergunta do tipo: "Quem foi o maior campeão de todos os tempos ?" só seria respondida corretamente caso houvesse conhecimento da relação de correferência entre "Mário" e "o piloto". Nessas sentenças, a relação de correferência ocorre entre elementos linguísticos que pertencem ao mesmo texto, sendo denominado de resolução de correferência em mono documento. Porém, também existe a necessidade de se encontrar elementos correferentes entre textos distintos. Como nos trechos de textos em 1.2.

(1.2) O presidente Luiz Inácio Lula da Silva afirmou hoje que o País baterá este mês um novo recorde de geração de empregos formais, acumulando 1,3 milhão de novas vagas em 2009. Fonte: O Estado de São Paulo

Lula disse que o Brasil terá mais um recorde na criação de vagas formais de emprego e ainda projetou para 2010 mais perspectivas de ampliação do mercado de trabalho. Fonte: Terra Economia

Nos textos apresentados no exemplo 1.2 é possível identificar que os elementos sublinhados "O presidente Luiz Inácio Lula da Silva" e "Lula" referem-se a uma mesma pessoa, ou seja, são correferentes. A identificação adequada da correferência entre os documentos pode facilitar a busca de informação sobre uma mesma entidade. Observa-se que no exemplo 1.2, ao se realizar uma pergunta como, "Qual a quantidade de novas vagas que Lula criou em 2009 ?", só se pode obter essa informação se o sistema souber que "Lula" e "O presidente Luiz Inácio Lula da Silva" referem-se a uma mesma pessoa. Isso em se tratando de um sistema automático, pois um humano encontraria a

resposta para a pergunta apenas utilizando o conhecimento de senso comum. O relacionamento entre “Lula” e “O presidente Luiz Inácio Lula da Silva” pode ser obtido através de um sistema de resolução automática de correferência em múltiplos documentos.

Vários trabalhos tratam da tarefa de resolução de correferência em múltiplos documentos como os de Bagga e Baldwin (1998b); Baron e Freedman (2008). O trabalho de Bagga e Baldwin (1998b) foi o primeiro a criar um método capaz de identificar as cadeias de correferência em múltiplos documentos. Em seu trabalho, Bagga e Baldwin identificam cadeias de correferência considerando apenas as diferentes entidades que possuíam o nome *John Smith* e variações com o nome do meio. Bagga e Baldwin justificam o desenvolvimento do trabalho argumentando que a tarefa de resolução de correferência em múltiplos documentos é diferente da em mono documento, pois na primeira não é possível utilizar alguns tipos de conhecimentos linguístico que são dependente da estrutura textual, como a árvore sintática.

Trabalhos como o de Baron e Freedman (2008) têm mostrado que é possível realizar a identificação de expressões correferentes em múltiplos documentos, entre entidades nomeadas do tipo pessoa e organização. Seu trabalho difere do de Bagga e Baldwin (1998b), pois utiliza o conjunto das entidades encontradas nos textos. Porém, apesar do sistema de Baron e Freedman (2008) ser mais completo que o de Bagga e Baldwin (1998b), ele ainda não trata todas as entidades nos textos, o que poderia ser útil para um sistema de perguntas e respostas.

Os métodos desenvolvidos tanto por Bagga e Baldwin (1998b) como por Baron e Freedman (2008) são baseados em algoritmos de aprendizado não supervisionado. No entanto, esses algoritmos utilizam um limiar que define quando o método de agrupamento aglomerativo deve parar. Esse limiar deve ser ajustado e seu valor pode ser dependente de cada conjunto de textos, para se obter a quantidade de cadeias de correferência no grupo de textos. Esse tipo de ajuste não é ideal, pois torna o método muito sensível aos ajustes de parâmetros do algoritmo de aprendizado. O ideal seria, portanto, a descoberta da quantidade de cadeias automaticamente.

Com relação às línguas nas quais os métodos de resolução de correferência em múltiplos documentos já foram utilizados, a língua mais explorada é o inglês, como nos trabalhos de Bagga e Baldwin (1998b); Phan et al. (2006); Saggion (2007); Wan (2008). No entanto, existem métodos para outras línguas como nos trabalhos de Baron e Freedman (2008) que, além do inglês, também

lida com o Árabe. Porém, para o português, até a escrita desta dissertação, não se conhece nenhum método para tratar do fenômeno de correferência em múltiplos documentos.

Nesse contexto, diante alguns dos problemas apresentados pelos métodos de resolução de correferência em múltiplos documentos, que são: (1) tratamento dos diversos tipos de entidade dos textos, (2) identificação automática da quantidade de cadeias de correferência e a (3) falta de métodos para línguas que não o inglês, esta dissertação estabelece como objetivo principal o desenvolvimento de um método que seja capaz de lidar com esses problemas dos sistemas de resolução de correferência em múltiplos documentos e que trate com textos da língua portuguesa.

Com base no objetivo dessa pesquisa, a hipótese deste trabalho é que, com a utilização de algoritmos não supervisionados é possível resolver cadeias de correferência em múltiplos documentos para o português, considerando os diversos tipos de entidade dos textos e sem haver a necessidade de informar o limiar necessário nos métodos aglomerativos para definir a quantidade de cadeias de correferência. Os resultados devem ser superiores aos métodos simples que, no contexto desta pesquisa, podem ser definidos com os que são baseados na concordância em cadeia de caracteres para construir as cadeias de correferência.

Para realizar o objetivo proposto e validar a hipótese de pesquisa, este estudo foi subdividido em diversas etapas: a) investigação dos métodos para resolução de correferência, com foco nos algoritmos não supervisionados, b) construção de um corpus com anotações de correferência em múltiplos documentos para ser utilizado na avaliação, c) definição de um método para resolução de correferência, d) implementação de um protótipo e e) avaliação do protótipo para verificar a validade da hipótese.

A organização dos próximos capítulos do trabalho são como segue.

No Capítulo 2 serão abordados os conceitos linguísticos relacionados ao fenômeno de correferência.

O Capítulo 3 trata dos métodos de resolução automático de correferência com o foco em algoritmos de aprendizado de máquina.

No Capítulo 4 apresenta os principais trabalhos relacionados à pesquisa desenvolvida no escopo desta dissertação.

O Capítulo 5 é apresentado o método para resolução de correferência desenvolvido no âmbito desta dissertação.

A avaliação do método proposto é apresentada no Capítulo 6.

As conclusões e os trabalhos futuros são apresentados no Capítulo 7.

CAPÍTULO

2

Conceitos linguísticos de correferência

Neste capítulo, são apresentados os conceitos linguísticos subjacentes ao estudo do fenômeno de correferência. Para introduzir esses conceitos, são apresentadas, inicialmente, as definições de coesão textual e referencial. Dentro desse contexto, este capítulo se aprofunda em um aspecto importante para esse estudo: as cadeias de correferência, particularmente os casos onde seus constituintes são sintagmas nominais. Outro ponto importante abordado é o fenômeno de correferência em múltiplos documentos, foco desta dissertação. O objetivo é elucidar conceitos e esclarecer o subconjunto do fenômeno que este trabalho visa abordar, do ponto de vista linguístico-computacional.

2.1 Coesão Textual

As pessoas quando se comunicam utilizando a língua (falada ou escrita) normalmente estabelecem conexões, ligações entre as diversas partes do texto. Um texto não é uma sequência de frases isoladas. Observe-se o seguinte trecho de um texto:

(2.1) Era uma vez... numa terra muito distante...uma princesa linda, independente e cheia de auto-estima.
Ela se deparou com uma rã enquanto contemplava a natureza e pensava em como o maravilhoso lago do seu castelo era relaxante e ecológico...
Então, a rã pulou para o seu colo e disse: linda princesa, eu já fui um príncipe muito bonito.
Uma bruxa má lançou-me um encanto e transformei-me nesta rã asquerosa.
Um beijo teu, no entanto, há de me transformar de novo num belo príncipe e poderemos casar e constituir lar feliz no teu lindo castelo...(Luís Fernando Veríssimo)

As expressões sublinhadas representam exemplos de elementos que dão ao texto a propriedade de unidade e não apenas de um conjunto de frases isoladas. Essa unidade é conseguida através das relações que essas expressões estabelecem no texto. As relações textuais fazem com que os elementos do texto (palavras, sintagmas, sentenças e parágrafos) estejam entrelaçados. O escritor utiliza os recursos de coesão textual para estabelecer relações textuais. Segundo Koch (1998), a coesão textual ocorre quando a interpretação de algum elemento no texto depende da de outro. A coesão textual garante, portanto, a conexão sequencial do texto.

No Texto 2.1, apresentado anteriormente, para a interpretação do pronome “Ela” é necessário que o leitor retorne aos elementos antes citados no texto e identifique que o pronome é uma expressão que retoma uma entidade já mencionada, no caso “uma princesa linda”. Já a expressão “o seu colo”, apesar de não retomar a entidade “uma princesa linda”, estabelece uma relação com a mesma. Existe uma dependência entre esses elementos no texto, pois a interpretação da primeira depende da segunda. Para estabelecer essas relações o escritor pode utilizar um conjunto de mecanismos de coesão, como: repetição, sinonímia, hiperonímia, elipse, substituição, uso de nomes genéricos e conjunções. A coesão textual estabelece por meio desses mecanismos um conjunto de relações no texto, constituindo uma verdadeira rede de ligações entre seus constituintes.

Segundo Koch (1998), a coesão textual é dividida em coesão referencial e coesão sequencial. A coesão referencial ocorre quando um elemento do texto retoma outro elemento do universo textual. Considere-se o trecho de texto a seguir:

(2.2) Minha mulher e eu temos o segredo para fazer um casamento durar: Duas vezes por semana, vamos a um ótimo restaurante, com uma comida gostosa, uma boa bebida e um bom companheirismo. Ela vai às terças-feiras e eu, às quintas.(Luís Fernando Verissímo)

No texto 2.2, é possível observar que o elemento linguístico "Ela" faz remissão ao componente do texto "Minha mulher". Nesse texto, o escritor utilizou o mecanismo de substituição para retomar o elemento anteriormente citado. Como já foi dito, esses mecanismos podem ser diversos. A seguir, são apresentados alguns exemplos que demonstram a utilização de alguns desses mecanismos.

Sinônimos:

(2.3) O avião já voava sobre São Paulo. O tempo de chegada da aeronave é de 1h.

Hiperônimos:

(2.4) Existe uma grande variedade de insetos. Esses animais estão presentes em boa parte do mundo.

Nomes genéricos:

(2.5) Um carro de corrida passou perto de mim. Essa foi a coisa mais rápida que eu já vir correr.

Elipse:

(2.6) Asse o frango até ficar dourado. Coloque Ø[a] em uma travessa enfeitada com pêssegos e rodelas de abacaxi.

---

[a] O símbolo Ø representa o elemento omitido da elipse, nesse caso "o frango".

Já a coesão sequencial diz respeito aos mecanismos que tornam as partes de um texto interdependentes. Essa interdependência dá ao texto a ideia de sequencialidade e continuidade. Os mecanismos de coesão sequencial estabelecem entre diversos segmentos dos textos (enunciados, parágrafos e

sequências textuais) vários tipos de relações semânticas e/ou pragmáticas. Observe-se o trecho de texto2.7 retirado de uma notícia do portal *online* Globo Esporte.com.

(2.7) A Fifa considera que o goleiro Rogério Ceni tem 94 gols na carreira. No entanto, nas contas do São Paulo, o artilheiro já marcou 96 vezes, ...
A partir disso, parece lógico que se utilize os mesmos critérios para a contagem dos gols. Critérios estes utilizados historicamente não apenas pelo clube, mas pelos mais diversos veículos de ...(Fonte: Globo Esporte.com . Dispónivel em http: //globoesporte.globo.com/futebol/times/sao-paulo/noticia/2011/01/ em-nota-oficial-sao-paulo-explica-contagem-de-gols-de-rogerio-ceni.html)

O escritor utiliza as expressões “No entanto” e “A partir disso” para dar ao texto uma noção de sequencialidade e desenvolvimento da ideia principal do texto, ao mesmo tempo em que atribui significado na relação entre as sentenças, neste caso, uma relação de contraste.

Os mecanismos de coesão referencial e sequencial dão ao texto a noção de progressão da ideia central do texto. A coesão permite, portanto, o encadeamento das relações entre os constituintes do texto.

No caso da coesão referencial, o encadeamento, como pode ser visto nos exemplos 2.2, 2.3, 2.4 e 2.5, representa a remissão de uma entidade já mencionada no texto. Esse encadeamento é possível pois os elementos entre si estabelecem uma relação denominada de **correferência**. O conjunto desses elementos que estabelecem entre si uma relação de correferência forma um encadeamento denominado de **cadeia de correferência**.

No exemplo 2.2 é possível identificar que os elementos linguísticos “Minha mulher” e “Ela” estão encadeados, portanto constituindo uma cadeia de correferência.

A relação de correferência e os constituintes da cadeia de correferência são os assuntos detalhados na próxima seção.

2.2 Correferência

O fenômeno de correferência é definido segundo Mitkov (2002) como expressões linguísticas, menções a uma entidade, que se referem a uma mesma entidade no mundo real. Nesse contexto, um termo que deve ser mais bem conceituado é o de cadeia de correferência, que foi mencionado na seção anterior. Neste trabalho, então, define-se cadeia de correferência como o conjunto de todas as menções a uma determinada entidade no texto (Mitkov, 2002). Para ilustrar esses conceitos considerem-se os exemplos 2.8 e 2.9:

(2.8) O time comandado pelo treinador Bernardinho só encontrou um pouco mais de dificuldades no segundo set. No terceiro, mesmo com vários reservas como o levantador Marcelinho e Samuel, os brasileiros conseguiram fechar a partida com tranquilidade. (Fonte: Jornal de Brasília)

(2.9) Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes. (Fonte: Folha de São Paulo)

Nos exemplos 2.8 e 2.9, os trechos sublinhados formam cadeias de correferência. No primeiro, os itens “O time comandado pelo treinador Bernardinho” e “os brasileiros” formam a cadeia. Já no segundo, são os elementos “o avião” e “O avião” que formam a cadeia de correferência.

Uma observação importante feita no trabalho de Koch (1998) é que entre os elementos pertencentes a uma cadeia de correferência estabelece-se uma relação de identidade. Apesar de, no primeiro exemplo, os itens lexicais não serem correspondentes existe uma relação semântica de identidade entre as duas menções.

Também no trabalho de Koch (1998), são apresentados outros autores que não consideram as referências no exemplo 2.8 idênticas. No trabalho de Halliday e Hasan (1976) os autores consideram que uma nova menção acrescenta uma nova especificação á entidade, ou seja, mais detalhes que antes não havia sido fornecido pelas menções a anteriores. No entanto, nesta dissertação é considerada a visão de Koch, na qual as menções em uma cadeia

de correferência têm entre si uma relação de identidade, premissa em que são baseados os estudos sobre coesão referencial.

As cadeias de correferência são constituídas através de uma relação de dependência. Essa relação geralmente se estabelece com um elemento linguístico anterior (anáfora), mas também pode ocorre com um elemento posterior (catáfora) (Koch, 1998). Abaixo são apresentados exemplos de anáfora e catáfora, respectivamente.

(2.10) A Dilma foi eleita a presidente do Brasil. Ela é a primeira mulher a exercer o cargo.

(2.11) O passáro seguia-o pelo caminho, reparou o moço.

Na anáfora apresentada no exemplo anterior, o pronome “Ela” tem uma relação anafórica com o sintagma nominal “A Dilma”. No segundo exemplo, o sentido da relação é oposto. O pronome “o” faz referência a uma menção à entidade que será apresentada posteriormente no texto, no caso “o moço”. Esse fenômeno caracteriza uma catáfora.

Para melhor entendimento do fenômeno de correferência nesta dissertação é realizada tanto a classificação dos tipos de componentes das cadeias de correferência, como dos tipos de relações de correferência que são geralmente encontrados. Essa classificação é feita baseada nos trabalhos de Vieira et al. (2008).

Na subseção 2.2.1 são apresentados os tipos de constituintes de uma cadeia de correferência. Já na subseção 2.2.2 explicita-se a definição de correferência em múltiplos documentos. Por fim, apresenta-se na subseção 2.2.3 os mecanismos linguísticos utilizados para realizar uma retomada a uma entidade já mencionada.

2.2.1 Constituintes das cadeias de correferência

Os constituintes das cadeias de correferência geralmente são classificados por meio da distinção entre os sintagmas nominais (SN) com núcleo nome e com núcleo pronome, apesar de os constituintes das cadeias de correferência não se resumirem a SNs, como pode ser observado no exemplo 2.12:

(2.12) A polícia federal realizou ontem uma busca na casa dos prefeitos suspeitos. A operação resultou em 5 prisões.

No exemplo 2.12, o SN "A operação" faz remissão à sentença "A polícia federal realizou ontem uma busca na casa dos prefeitos suspeitos". No entanto, esses casos em que os participantes da cadeia de correferência ultrapassam o tamanho de um SN são menos frequentes. A maior parte dos estudos linguístico-computacionais concentra-se nas cadeias de correferência compostas por SNs, como a cadeia apresentada no exemplo 2.13.

(2.13) O presidente Luiz Inácio Lula da Silva ironizou na quarta-feira, 17, sem citar nomes, os grandes empresários que têm de ser socorridos por causa de perdas bilionárias. Ele enfatizou a importância de emprestar recursos para os mais pobres.(Fonte: O Estado de São Paulo)

Nesse exemplo, a cadeia de correferência é constituída pelos SNs destacados o "O presidente Luiz Inácio Lula da Silva", cujo núcleo é o nome "presidente", e o pronome "Ele", no qual o núcleo é ele mesmo.

Uma subdivisão dos grupos em SNs com núcleo nome ou pronome é frequentemente realizada para um melhor entendimento da variabilidade do fenômeno da correferência. Para os SNs com núcleo nominal, ainda há a divisão dos SNs em dois grupos: os sintagmas com ou sem modificadores. Artigos, adjetivos e pronomes, todos são possíveis modificadores dos SNs com núcleo nominal, tanto para SNs que têm como núcleo um nome comum ou um nome próprio.

Segue uma lista apresentando a subclassificação que é considerada neste trabalho, seguindo a classificação proposta no trabalho de Carbonel (2007).

I. SN com núcleo nominal sem modificadores

   (a) O núcleo do SN é um nome comum (substantivo simples) sem modificadores. Observe-se, no exemplo 2.14, o SN "Pesquisas":

(2.14) "*Pesquisas* foram realizadas."

(b) O núcleo do SN é um nome próprio. No exemplo 2.15 há o SN “Lula”.

(2.15) Lula disse que irá definir quais áreas serão destinadas à produção de etanol no País. O presidente afirmou que o Brasil vai sediar, . . . (Fonte: O Globo)

II. SN com núcleo nominal com modificadores

(a) O núcleo do SN é um nome comum ou um nome próprio antecedido por um artigo definido. No exemplo 2.16, os SNs “o avião” e o “A aeronave” mostram este caso.

(2.16) Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. A aeronave se chocou com uma montanha e caiu.

(b) O núcleo do SN é um nome comum antecedido por um artigo indefinido. Como é apresentado no exemplo 2.17 o SN “Um acidente”.

(2.17) Um acidente aéreo matou 17 pessoas. As vítimas do acidente foram 14 passageiros e três membros da tripulação.

(c) O núcleo do SN é um nome comum antecedido por um pronome demonstrativo. “Esses colegiados”, no exemplo 2.18.

(2.18) Para cumprir uma dessas finalidades, funcionam no País o Tribunal de Contas da União, 27 Tribunais de Contas dos Estados e do Distrito Federal e três Tribunais de Contas dos Municípios. Esses colegiados são órgãos de assessoramento . . . (Fonte: direitoce.com.br. Disponível no endereço http://www.direitoce.com.br/noticias/46187/.html)

(d) O núcleo do SN é um nome comum antecedido por um pronome possessivo. Como exemplo são apresentados duas ocorrência de “seu trabalho” no trecho 2.19:

(2.19) Marcos irá terminar seu trabalho a tempo. Seu trabalho é muito difícil de realizar.

(e) O núcleo do SN é um nome comum ou próprio antecedido por um pronome interrogativo. No exemplo 2.20, o SN “Quantas vezes” demostra esse caso.

(2.20) Quantas vezes você vai ao cinema?

(f) Núcleo do SN é um nome comum ou próprio antecedido por qualificadores, geralmente pronomes indefinidos – “Várias pessoas” e “Esses ganhadores”, no exemplo 2.21.

(2.21) Várias pessoas já ganharam na loteria. Esses ganhadores tiveram muita sorte.

(g) Núcleo do SN é um nome comum ou próprio antecedido por um pronome numeral. O SN “O primeiro empreendimento da imobiliária na cidade” em 2.22 é um exemplo.

(2.22) O primeiro empreendimento da imobiliária na cidade foi um sucesso. Esse condomínio vai vender muito.

III. SN com núcleo pronominal

(a) O SN é formado apenas por um pronome demonstrativo, como o pronome “isso” no exemplo 2.23.

(2.23) Tenho a convicção de que mereço ganhar, mas não tenho a sensação de que isso vá acontecer.

(b) O SN é formado apenas por um pronome pessoal. No exemplo 2.24, o pronome “ele” demonstra isso.

(2.24) O Felipe Anderson tem somente 17 anos e a tendência é que ele vá melhorando a cada jogo.

(c) O SN é formado apenas por pronome indefinido. No exemplo 2.25, o pronome “Alguém” demonstra esse caso.

(2.25) Alguém sabe onde tem um bom lugar para comer ?

(d) O SN é formado apenas por um pronome possessivo, como pronome “Meu” no exemplo 2.26.

(2.26) “Esse carro está na frente da minha garagem. De quem é esse carro?”, falou o motorista enfurecido. “Meu”, respondeu a mulher.

(e) O SN é formado apenas por um pronome interrogativo, como o pronome “quando” no exemplo 2.27.

(2.27) Você volta da festa quando ?

Classificação quanto ao estado em que os SNs aparecem no discurso

Como relação ao estado em que os SNs aparecem no discurso, esses podem ser classificados como segue:

- *Elementos novos no discurso*: o SN introduz um novo referente no discurso sem apresentar parte de seu sentido ancorado em uma expressão anterior. Observa-se no exemplo 2.28 que a interpretação do SN “A presidente Dilma Rousseff” não depende da de outro.

(2.28) A presidente Dilma Rousseff deve promover a primeira inauguração de seu governo na próxima sexta-feira (28). Ela é esperada no Rio Grande do Sul, onde está a usina termelétrica Candiota 3, instalada no município de mesmo nome. (Fonte: O Globo)

- *Elementos já mencionados no discurso*: o SN retoma uma entidade já mencionada por outro elemento no discurso. No exemplo, 2.28 o pronome “Ela” retoma o SN “A presidente Dilma Rousseff”.

- *Elementos associativos*: introduzem uma nova entidade no discurso cujo sentido é ligado a outra entidade anteriormente mencionada. No exemplo 2.29, o SN "O seu motor" depende do SN "O carro", no entanto a relação que se estabelece não é de identidade e sim de parte/todo, não sendo, portanto uma relação de correferência.

(2.29) O carro que ganhou a corrida era muito rápido. O seu motor era muito bom.

- *Elementos dêiticos*: A referência do elemento linguístico não é encontrada no texto, mas é determinada pelo contexto. No exemplo 2.30, pronome "Eu" refere-se a uma entidade externa ao texto.

(2.30) Eu não posso ficar aqui sozinho.

Classificação dos relacionamentos dos SNs anafóricos

Os SNs novos no discurso e os já mencionados estabelecem diferentes tipos de relações. O estudo desses tipos de relações contribui para a análise do fenômeno de correferência como um todo. Segundo Vieira et al. (2008), os tipos de relações que são estabelecidas entre os SNs anafóricos são os seguintes:

- *Direta*: a expressão anafórica tem um antecedente com o núcleo idêntico. No trecho 2.31 é apresentado um exemplo com os SNs "O avião de Santos Dumont" e "O avião".

(2.31) O avião de Santos Dumont foi o primeiro a voar por Paris. O avião se chamava 14 bis.

- *Indireta*: a expressão anafórica tem um antecedente com o núcleo diferente, como exemplo os SNs "O novo carro" e "O veículo" no trecho 2.32.

(2.32) O novo carro foi vendido muito rápido. O veículo era muito bom.

- *Encapsulamento*: a expressão anafórica retoma um trecho de texto maior que um SN. Segue um exemplo com esse tipo de relação anafórica entre os SNs “A nossa ideia de marketing é conquistar os consumidores pelo visual” e “Essa proposta”.

(2.33) A nossa ideia de marketing é conquistar os consumidores pelo visual. Essa proposta vai dar muito certo.

2.2.2 Correferência em mono documento e múltiplos documentos

Segundo Bagga e Baldwin (1998b), a correferência pode ser classificada em dois tipos: a que ocorre entre menções de um documento (mono documento) e a que ocorre quando uma mesma menção é tratada em vários documentos (múltiplos documentos). Nos exemplos anteriores foram apresentadas apenas ocorrências em mono documento. Esse tipo de correferência é o mais abordado na literatura. Como exemplo de correferência multi-documentos apresentam os trechos em 2.34

(2.34) O presidente Luiz Inácio Lula da Silva afirmou hoje que o país baterá este mês um novo recorde de geração de empregos formais, acumulando 1,3 milhão de novas vagas em 2009. (Fonte: O Estado de São Paulo)

Lula disse que o Brasil terá mais um recorde na criação de vagas formais de emprego e ainda projetou para 2010 mais perspectivas de ampliação do mercado de trabalho.(Fonte: Terra Economia)

No exemplo 2.34, são apresentados dois trechos de textos de fontes diferentes. Porém, é possível estabelecer uma relação entre os sintagmas nominais “O presidente Luiz Inácio Lula da Silva” e “Lula”, apesar de os autores não estabelecerem essa relação intencionalmente nos textos. Observa-se que os sintagmas se referem a uma mesma entidade. Entre esses sintagmas, portanto, existe uma relação de correferência.

O fenômeno de correferência em múltiplos documentos pode ocorrer através da relação entre diversos tipos de textos que podem se diferenciar ou igualar quanto a: a) contextos, b) assuntos, c) focos, d) intenções e e) gêneros.

Para esclarecer qual a fração do fenômeno de correferência em múltiplos documentos que esta dissertação aborda é necessário elencar algumas variações desse fenômeno. Apresenta-se o exemplo 2.35:

(2.35)

O presidente do Brasil, Luiz Inacio Lula da Silva, afirmou ontem (17) que não pretende ser secretário-geral da ONU. (Fonte: Jornal Clarim)	A presidente eleita do Brasil admitiu, este sábado, a responsabilidade que acarreta a sua sucessão a Lula da Silva e apelou à união de todos os brasileiros. (Fonte: O Estado de São Paulo)

No exemplo 2.35, os textos tratam de assuntos diferentes. No entanto é possível identificar que os elementos marcados “O presidente do Brasil, Luiz Inácio Lula da Silva” e “Lula da Silva” referem-se à mesma entidade. Essa possibilidade de elementos correferentes ocorrerem em diversos tipos de textos torna o fenômeno de correferência em múltiplos documentos difícil de ser identificado. Segue outro exemplo apresentado no qual há variação de gênero dos textos.

(2.36)

— Bom dia João. — Bom dia. — Agora temos uma nova presidente do Brasil. — A Dilma será uma boa presidente.	A presidente eleita do Brasil admitiu, este sábado, a responsabilidade que acarreta a sua sucessão a Lula da Silva e apelou à união de todos os brasileiros. (Fonte: O Estado de São Paulo)

No exemplo 2.36, o primeiro texto é um diálogo, enquanto que o segundo é um monólogo (um texto jornalístico). Porém, é possível identificar que os SNs “A Dilma” no primeiro texto e “A presidente eleita do Brasil” são as mesmas pessoas.

Nesses exemplos, são apresentados possível verificar que o fenômeno de correferência em múltiplos documentos pode ocorrer entre diversos tipos de textos. No entanto, o presente trabalho tem como foco textos jornalísticos que

tratam sobre o mesmo evento[1], como os apresentados no corpus CST-News (Maziero et al., 2010).

Quanto aos constituintes das cadeias de correferência, esta pesquisa restringe-se aos SN com núcleo nominal, limitando-se a referências diretas e indiretas entre os SNs. Esse trabalho visa identificar as cadeias de correferência que ocorrem em mono-documento e as relações de correferência que ocorrem em múltiplos documentos, atendo-se à fração do fenômeno que obedece às restrições acima descritas.

2.2.3 Mecanismos linguísticos utilizados na correferência

Na construção de um texto, o escritor, na tentativa de deixar explícitas as ligações de correferência entre as diversas menções no texto, utiliza-se de alguns recursos linguísticos. Esses instrumentos podem ser: lexicais, sintáticos, semânticos, discursivos e pragmáticos. A utilização desses recursos facilita a identificação por parte do leitor das relações entre as menções às entidades no texto. Com a identificação das relações entre as menções no texto é possível reconhecer os elementos correferentes. Seguem os exemplos 2.37 e 2.38:

(2.37) A vitória de Dilma era certa. Dilma será a primeira presidente do Brasil.

(2.38) Paris hoje amanheceu linda. A cidade luz é mesmo maravilhosa.

No exemplo 2.37, é apresentada uma anáfora direta, pois os núcleos dos SNs são idênticos. Para construir esse tipo de relação, o escritor apenas utiliza a repetição lexical entre as menções, nesse caso, repetindo o SN “Dilma”. Já no exemplo 2.38, para identificar a relação entre as duas menções, o leitor precisa utilizar outros tipos de conhecimentos, como semântico e/ou pragmático. Como já apresentado na subseção 2.2.1, essa relação caracteriza-se como uma anáfora indireta. A anáfora indireta demanda processos cognitivos mais complexos que a anáfora direta (Vieira et al., 2008). Por causa dessa

[1]Evento no contexto desta dissertação, é um acontecimento ou uma ação; em particular, narrado por um texto jornalístico ou por vários. Por exemplo, textos que tratem de um mesmo jogo de futebol.

complexidade e da utilização de diversos tipos de conhecimento, a anáfora indireta pode ser expressa de várias formas com é apresentado no trabalho de Vieira et al. (2008). A seguir, são apresentadas algumas das principais formas:

- *Relação entre nome próprio e nome comum*:

(2.39) A Petrobrás desistiu de comprar a participação da italiana Eni na Galp. Ontem a companhia se negou a comentar as informações publicadas na imprensa portuguesa sobre as negociações, mas o Valor apurou que elas foram encerradas. Fonte: Revista portosenavios. Disponível em http://portosenavios.com.br/site/noticiario/geral/7952-petrobras-desiste-de-comprar-fatia-da-galp

- *Relação de sinonímia*:

(2.40) Os novos carros têm maior segurança. Esses veículos já trazem vários itens de série.

- *Nominalização de verbos*:

(2.41) Cuba propôs aos EUA "telefone vermelho"para tratar de disputas. Proposta foi feita no ano passado por Raul Castro à secretária de Estado dos EUA, Hillary Clinton, segundo documentos secretos revelados pelo WikiLeaks. (Fonte: Portal Exame.com)

- *Hiponímia/hiperonímia*:

(2.42) O cachorro entrou na casa de Maria. O animal estava com muita fome.

Além dos casos já mencionados anteriormente, existem casos em que o processo de identificação da relação anafórica exige por parte do leitor a utilização de conhecimento de mundo, como no exemplo 2.43.

(2.43) Ronaldo é um jogador muito versátil. O Fenômeno já foi 2 vezes o melhor jogador do mundo.

Nesse exemplo, para resolver a anáfora entre Ronaldo e O Fenômeno o leitor utiliza um conjunto de títulos ou codinomes. A necessidade de utilização de

diversos níveis do conhecimento para explicitar as relações de correferência torna esse fenômeno muito complexo, tanto para tratamento computacional, como para humanos (Vieira et al., 2008).

Observa-se que no trecho de texto abaixo, é ambígua a identificação do antecedente correto do pronome “ele”, pois existem dois SNs igualmente prováveis “João” e “Mário”.

(2.44) “João e Mário passeavam na rua, quando ele ao ver o policial evadiu-se do local”, afirmou a testemunha.

Observa-se que nesse trecho apenas o contexto poderia definir qual seria o antecedente correto. Esse tipo de ambiguidade pode ocorrer dificultando o entendimento por parte do leitor da mensagem que está sendo transmitida. No entanto, existem casos em que apesar de não haver ambiguidade no entendimento humano, pode ser um desafio para a identificação automática. Há um o exemplo disso em 2.45:

(2.45) Para bater o recorde o piloto chegou com o carro a 500km/h. O carro era um ótimo veículo, pois ele tinha mais de 2000cv.

No exemplo 2.45, apesar de ser claro para o leitor a quem o pronome ele se refere, para a identificação automática seria um caso de difícil resolução, pois existem vários SNs prováveis (“o recorde”, “o piloto”, “o carro” e “um ótimo veículo”). A identificação correta do antecedente por parte de um sistema automático depende da qualidade e diversidade do conhecimento utilizado pelo sistema. Modelar as estruturas complexas da língua é um grande desafio para a PLN. Esse problema torna a tarefa de resolução de correferência complexa para essa área de pesquisa.

2.3 Considerações finais

Neste capítulo foram abordados alguns conceitos para a compreensão do fenômeno da correferência. Foi introduzido o conceito de coesão textual e, a partir desse, apresentado o conceito de coesão referencial. Com base nela, o fenômeno da correferência foi demonstrado como ocorre e como foi definido. A

classificação para correferência que é utilizada neste trabalho foi apresentada. Com base nessa classificação foi definido o foco desta dissertação, que são as correferências entre SNs com núcleo nominal que ocorre em mono e em múltiplos documentos.

No próximo capítulo são abordados os métodos computacionais utilizados para identificar as cadeias de correferência automaticamente, os conhecimentos linguísticos utilizados nessa resolução e a forma como foi feita a avaliação.

CAPÍTULO

3

O processo de resolução automática de correferência

Neste capítulo, serão apresentados os passos para o desenvolvimento de um algoritmo de resolução automática de correferência. Inicialmente, é demonstrado como é feita a identificação dos elementos que podem participar das cadeias de correferência. Depois, são apresentados quais os tipos de conhecimentos linguísticos utilizados por sistemas que tratam dessa tarefa. Em seguida, são mostrados os algoritmos e as arquiteturas utilizadas para a construção das cadeias de correferência. Uma atenção especial é dada aos algoritmos baseados em Aprendizado de Máquina (AM), pois é esse tipo de algoritmo que é empregado no protótipo desenvolvido para validar o método apresentado nessa dissertação. Por fim, são apresentadas as formas de avaliação desses algoritmos, com foco na definição das medidas de avaliação.

3.1 Formas de obtenção dos sintagmas nominais

O primeiro passo de um algoritmo de correferência é obter o conjunto dos elementos do texto que podem participar das cadeias de correferência. Geralmente, os trabalhos sobre correferência delimitam esse conjunto apenas nos dos SNs. Na literatura são encontrados tipicamente três métodos para obter os SNs de um texto: (1) obtenção dos SNs automaticamente através de

um analisador sintático (*parser* sintático), (2) extração direta de um corpus anotado manualmente ou (3) utilização dos SNs extraídos de um corpus anotado manualmente na realização do treinamento do sistema, no caso de sistemas de aprendizado supervisionado, e na fase de teste os extrai automaticamente. Deve-se notar que no último caso, o número de menções obtidas automaticamente pode ser diferente das anotações manuais. Esse fato pode dificultar a avaliação do sistema que utiliza esse método.

Os três métodos podem levar a avaliações diferentes para um mesmo sistema, pois a quantidade de SNs obtida por cada um deles pode ser diferente.

Alguns pesquisadores argumentam que os resultados de uma avaliação obtida de um sistema que extrai os SNs de um corpus anotado reflete o verdadeiro desempenho do algoritmo de resolução, pois não seria inserido o erro do analisador sintático. No entanto, há trabalhos como o de Stoyanov et al. (2010), no qual seus autores discordam dessa afirmação, argumentando que esse tipo de avaliação não é realista, visto que um sistema automático real deveria resolver todas as subtarefas necessárias à resolução das correferências.

Neste trabalho, como será visto no Capítulo 4, é realizada a extração dos SNs automaticamente, realizando, portanto, uma avaliação do sistema segundo o argumento de Stoyanov et al. (2010).

3.2 Fontes de conhecimento para a resolução de correferência

Tipicamente, várias fontes de conhecimento são utilizadas na tentativa de melhor definir quais os elementos da cadeia de correferência. Um conjunto de traços ou características linguísticas é explorado pelos trabalhos desenvolvidos para resolver correferência. Essas características linguísticas podem variar quanto ao tipo de conhecimento utilizado, que pode ser superficial e/ou profundo da língua.

Nas subseções a seguir, são detalhadas as características linguísticas frequentemente utilizadas para resolver correferência.

3.2.1 Concordância em cadeia de caracteres

A maioria dos sistemas de correferência determina padrões de concordância em cadeia de caracteres (*string matching*). Esses padrões, que são fáceis de computar, contribuem para melhorar o desempenho desses sistemas. Apresenta-se o exemplo 3.1:

(3.1) A viagem para São Paulo vai ser no fim do ano. Essa viagem vai ser longa.

No exemplo 3.1, os SNs "A viagem para São Paulo" e "Essa viagem" têm em comum a cadeia de caracteres "viagem". É frequente esse tipo de ocorrência entre os constituintes da cadeia de correferência. No trabalho de Soon et al. (2001) é apresentada uma evidência de que a concordância em cadeia de caracteres é um importante traço que deve ser considerado quando da construção de um sistema automático.

Nesse trabalho, é definido um *baseline* somente com essa característica, obtendo-se um desempenho apenas 10% menor de que um sistema utilizando um conjunto de outros 11 traços.

Os tipos de padrões de concordância frequentemente utilizados são: concordância total, concordância parcial (*substring matching*) e concordância como o núcleo do SN. No entanto, formas mais sofisticadas foram abordadas na literatura, como a distância mínima de edição (Strube, 2002) e a mais longa cadeia de caracteres em comum (Castaño et al., 2002).

Outros trabalhos realizaram cálculos de medidas de similaridade utilizadas nas áreas de mineração de texto e extração da informação. Por exemplo, no trabalho Yang e Zhou (2004) é utilizada a medida tf-idf (*term frequency – inverse document frequency*) entre dois sintagmas nominais.

3.2.2 Características da árvore sintática

Vários trabalhos utilizam as árvores sintáticas na identificação dos elementos correferentes. As árvores sintáticas são exploradas principalmente para definir o antecedente de uma anáfora pronominal. Um dos primeiros trabalhos a utilizar a árvore sintática na resolução pronominal é o trabalho de

Hobbs (1977), desenvolvido para a língua inglesa. Hobbs apresenta um algoritmo capaz de resolver anáfora pronominal realizando uma busca em largura na árvore sintática, procurando por SNs com o mesmo gênero e número. O algoritmo de Hobbs foi adaptado para outras línguas como no trabalho de Santos (2008) para o português.

Geralmente, os algoritmos que utilizam a árvore sintática definem um conjunto de heurísticas para serem utilizadas na determinação do antecedente do pronome. Todavia, existem trabalhos como de Yang et al. (2006b) em que é desenvolvido um método no qual os padrões da árvore sintática são definidos automaticamente a partir de um corpus anotado, utilizando para treinamento um *Support Vector Machine* (SVM) (Vapnik, 1995).

No contexto desta dissertação, a árvore sintática será considerada quando tratamos da estrutura do SN, pois essa estrutura é um subconjunto da árvore sintática. Já as características que pode ser obtidas dá árvore sintática não será utilizadas na resolução de correferência em múltiplos documentos, pois não existe está estrutura entre os documentos. Essas características pode ser consideradas quando na resolução em mono-documentos das correferência.

3.2.3 Características Gramaticais

As características gramaticais são fortemente utilizadas na maioria dos trabalhos que tentam identificar as cadeias de correferência. No trabalho de Ng e Cardie (2002), por exemplo, é utilizado um conjunto de 34 características gramáticas. Além deste, vários outros trabalhos exploram essas características, algumas das mais frequentemente utilizadas são: a função gramatical (sujeito, objeto), definição do tipo de sintagma (definido, indefinido, demonstrativo, nominal ou preposicional), gênero e número. Outras características como a verificação do SN como aposto de outro SN também são utilizadas.

O uso de atributos gramaticais isoladamente não determina que elementos são correferentes. Todavia, como pode ser visto no trabalho de Ng e Cardie, a combinação de diversas dessas características pode representar uma melhoria no desempenho do sistema.

3.2.4 Características semânticas

Outro nível de conhecimento muito utilizado é o semântico, pois como foi visto no Capítulo 2, vários tipos de relações de correferência utilizam mecanismos linguísticos que podem ser classificados nesse nível, como a sinonímia

e a hiperonímia. Para adquirir esse tipo de informação, comumente, são utilizados repositórios de informações semânticas como a WordNet (Fellbaum, 1998).

Nos trabalhos de Soon et al. (2001) e Vieira e Poesio (2000) é utilizada a WordNet com o objetivo de identificar sinônimos entre os SNs. Um dos problemas de utilizar uma base de dados como a WordNet é determinar qual o sentido da palavra que deve ser utilizada.

No trabalho de Soon et al. é utilizado o primeiro sentido que a base de dados retorna, essa escolha simplifica seu algoritmo. No de Vieira e Poesio é verificado se os SNs pertencem ao mesmo *synset*[1] . Já outros trabalhos como o de Ponzetto e Strube (2006) utilizam todos os sentidos possíveis e desenvolvem uma medida de similaridade para calcular a proximidade semântica entre dois SNs.

Alguns trabalhos utilizam conhecimentos de bases como a Wikipedia[2]. Esses trabalhos desenvolvem um conjunto de heurísticas para tornar possível a extração simplificada de informação desse tipo de base. No trabalho de Ponzetto e Strube (2006), por exemplo, é desenvolvida uma heurística que extrai um conjunto de características realizando buscas pelo núcleo dos SNs nos títulos das páginas e nas categorias.

3.2.5 Características do discurso

Trabalhos que exploram características no nível do discurso são menos frequentes na literatura, dada a própria complexidade do tratamento desse nível e as poucas ferramentas disponíveis em comparação a outros níveis.

Contudo, existem trabalhos que adotam características desse nível como Rino e Seno (2006) que exploram a *Rhetorical Structure Theory* (RST) (Mann e Thompson, 1987) e a Teoria das Veias (Ide e Cristea, 2000). Rino e Seno exploram a importância do tratamento das cadeias de correferência na tarefa de sumarização automática. As autoras utilizam o conhecimento disponível nas árvores RST para evitar a quebra da continuidade referencial nos sumários produzidos.

[1] *Synset*: conjunto de sinônimos (Miller et al., 1990)

[2] Wikipédia é uma enciclopédia multilíngue *online* livre colaborativa, ou seja, escrita internacionalmente por várias pessoas comuns de diversas regiões do mundo, todas elas voluntárias. Disponível no endereço http://pt.wikipedia.org

Após a definição dos conhecimentos linguísticos que são tipicamente utilizados para ajudar a resolver as correferências, o próximo passo é definir os algoritmos que combinaram esses traços para identificar as cadeias de correferência. Na Seção 3.3 são apresentados alguns tipos de algoritmos frequentemente utilizados.

3.3 Algoritmos de resolução de correferência baseados em AM

Os métodos de resolução de correferência podem ser divididos em abordagens heurísticas e baseados em AM.

As abordagens heurísticas foram utilizadas principalmente nos primeiros trabalhos de resolução de correferência. No entanto, essa abordagem foi substituída por métodos baseados em algoritmos de AM, principalmente os que fazem uso de algoritmos supervisionados (Ng, 2010).

As técnicas de aprendizado de máquina são as mais abordadas na literatura, sendo os resultados obtidos por essas técnicas considerados como os melhores (Souza et al., 2008).

Esses trabalhos obtiveram um grande impulso graças à disponibilização de corpora, como os das competições MUC-7 (1997) e MUC-6 (1995). Outro fator importante dentro desse contexto é a própria evolução dos algoritmos de AM e a consequente disponibilização de ferramentas com eles. Esses fatores culminaram com uma crescente utilização dos algoritmos para resolução de correferência.

Os principais algoritmos de aprendizado de máquina para resolução de correferência são dois: algoritmos supervisionados e não supervisionados. Nas subseções 3.3.1 e 3.3.2 são apresentadas as arquiteturas desses tipos de sistemas e seus funcionamentos.

3.3.1 Abordagens supervisionadas

As abordagens supervisionadas são, dentre as que utilizam AM, as mais exploradas na literatura, como já citado anteriormente, em parte pela disponibilização dos corpora anotados com informação de correferência. Vários

autores como os Cardie e Wagstaf (1999); Soon et al. (2001); Souza et al. (2008); Yang et al. (2003, 2004, 2006a) utilizam essa abordagem.

As abordagens supervisionadas consistem da construção de um classificador que seja capaz de determinar quais são os SNs correferentes de dado conjunto de textos anotados com informações de correferência. Apresenta-se o exemplo 3.2:

(3.2) Monteiro Lobato foi um dos maiores escritores do Brasil. Monteiro é autor do famoso "O sítio de pica-pau amarelo". Ele também escreveu obras como "Jeca Tatuzinho" e "A caçada da onça".

O classificador empregado na abordagem supervisionada tenta encontrar quais os pares de sintagmas correferentes, utilizando um conjunto de características linguísticas. No Texto 3.2, um classificador utilizando um atributo de concordância de cadeias de caracteres poderia identificar que "Monteiro Lobato" e "Monteiro" são correferentes. Já a relação do pronome "Ele" com o SN "Monteiro" poderia feita por meio de uma combinação de caraterísticas como concordância em gênero, em número e o paralelismo sintático, já que os dois são sujeitos de suas respectivas sentenças. Com a identificação dos pares de SNs correferentes, é possível construir a cadeia de correferência. Assumindo uma transitividade entre os pares, é factível afirmar que se "Monteiro Lobato" e "Monteiro" são correferentes e por outro lado "Monteiro" e "Ele" também, então "Monteiro Lobato" e "Ele" são correferentes. Assim sendo, a cadeia de correferência identificada é "Monteiro Lobato", "Monteiro" e "Ele".

Uma visão melhor detalhada de um sistema supervisionado é apresentada na Figura 3.1.

Essa figura apresenta uma arquitetura de sistema genérico. O sistema é dividido em duas fases: a de treinamento, na qual é induzido um modelo de classificação, e a de testes, na qual é feita a utilização do classificador construído na fase anterior para separar os pares de SNs correferentes e não correferentes e, por fim, agrupar as cadeias.

A fase de treinamento tem como entrada um corpus com as cadeias de correferência anotadas. Essa fase é dividida em diversas etapas. São elas: a extração dos SNs (descrita na Seção 3.1); a definição dos prováveis pares de SNs; a extração dos atributos e o treinamento do classificador.

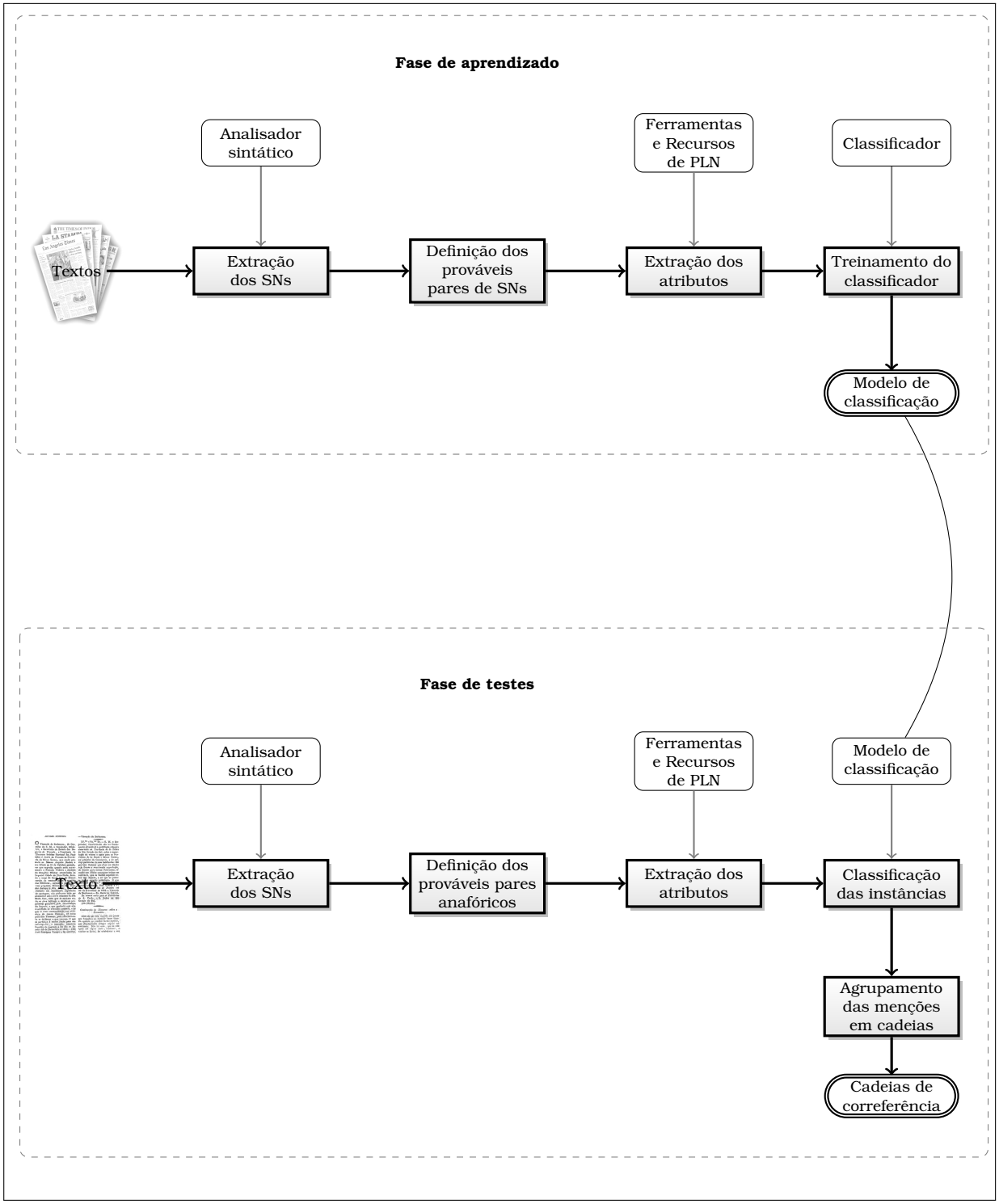

Figura 3.1: *Arquitetura de um sistema de resolução de correferência supervisionado*

Na etapa de definição dos pares de SN para treinar o classificador, o principal problema é o desbalanceamento das classes, pois em um texto a quantidade de pares de SNs correferentes é bem menor que a dos não correferentes. Por exemplo, no trabalho de Souza et al. (2008), desenvolvido para o português utilizando o corpus Summit (Collovini et al., 2007), na fase de treinamento, a quantidade dos pares de SNs não correferentes (instância negativa) foi de 6 vezes maior que a quantidade dos correferentes (instâncias positivas). Esse desbalanceamento dificulta, por parte do classificador, o aprendizado da classe dos SNs correferentes. Para isso, vários tipos de heurísticas foram desenvolvidas.

A mais utilizada foi a proposta por Soon et al. (2001). Dado um SN anafórico, SN_k, as instâncias positivas são criadas utilizando-se o SN antecedente anafórico (SN_j) de SN_k e o próprio. As instâncias negativas são criadas utilizando-se o SN_k combinado com os SNs encontrados no intervalo $[SN_{j+1}, SN_{k-1}]$.

Outros trabalhos desenvolveram mecanismos de filtro (Strube, 2002; Yang et al., 2003). Nesses mecanismos, alguns SNs são descartados, por exemplo, filtros que são geralmente utilizados para descartar aqueles que não concordam em gênero e número.

Com a definição das instâncias para treinamento, a próxima etapa é obter as características linguísticas. As características geralmente utilizadas na classificação são basicamente de dois tipos: os atributos que descrevem as menções e os que descrevem a relação do SN anafórico com seu antecedente.

Na Tabela 3.1 é apresentado o conjunto de atributos utilizados por Soon et al. (2001), na qual pode ser visto que existem atributos que descrevem a menção, como o i-Pronome e j-Pronome, que indicam se a menção é ou não um pronome. Há outros atributos que descrevem a relação entre as menções, como o que verifica se um SN é aposto do outro.

Para a extração dos atributos, como foi descrito na Seção 3.2, são utilizados diversos tipos de recursos como, por exemplo, analisadores sintáticos e semânticos, e ferramentas do PLN como tesauros.

Já a fase de teste tem como entrada um texto e o sistema terá como saída as cadeias de correferência anotadas desse texto. Essa fase, assim com a de treinamento, é divida em etapas. São elas: extração dos SNs, definição dos pares de SNs, extração dos atributos, classificação das instâncias e agrupamento das menções em cadeias.

Atributos		Descrição
Das Menções	Sintagma nominal definido	Recebe verdadeiro se j é um sintagma nominal definido ou falso caso contrário.
	Sintagma nominal demonstrativo	Recebe verdadeiro se j é um sintagma nominal demonstrativo ou falso caso contrário.
	i-Pronome	Recebe o valor verdadeiro se i for pronome ou falso caso contrário.
	j-Pronome	Recebe o valor verdadeiro se j for pronome ou falso caso contrário.
Da Relação	Distância	Número inteiro que mostra a distância em quantidade de sentenças entre i e j. Se i e j estão na mesma sentença o valor é 0, se i está na sentença anterior o valor é 1 e assim por diante.
	Cadeias de caracteres	Recebe verdadeiro se i e j têm concordância em cadeia de caracteres retirando os artigos e pronomes demonstrativos ou falso caso contrário.
	Número	Recebe verdadeiro se i e j concordam em número ou falso caso contrário.
	Gênero	Recebe verdadeiro se i e j concordam em gênero ou falso caso contrário.
	Classe semântica	Recebe verdadeiro se i e j pertencem a mesma classe semântica ou falso se os valores são desconhecidos ou não têm a mesma classe semântica.
	Nome próprio	Recebe verdadeiro se i e j são nomes próprios ou falso caso contrário.
	Pseudônimo	Recebe verdadeiro se i é o pseudônimo de j ou vice-versa, ou falso caso contrário. Nesse caso, essa característica é válida para entidades nomeadas (pessoas, organizações e data). Para cada tipo de entidade existem regras para verificar se os SNs são pseudônimos. Por exemplo, para a entidade do tipo organização é verificado se uma sigla é formada pelas iniciais de um nome próprio no texto. Um sigla como IBM seria identificada como pseudônimo de *International Business Machines* se ocorresse no texto.
	Aposto	Recebe verdadeiro se i é um aposto de j ou falso caso contrário. Nesse caso é verificado se os SNs ocorrem entre vírgulas como, por exemplo na sentença, "João, o pedreiro, foi a sua casa".

Tabela 3.1: *Atributos que descrevem a relação entre dois SNs i e j (Soon et al., 2001)*

As três primeiras etapas da fase de teste funcionam de forma praticamente idêntica à fase de aprendizado. Há porém, uma diferença na etapa de definição dos pares de SNs, pois nessa fase não há necessidade de balancear as classes.

Na fase de classificação das instâncias, é utilizado o modelo de classificação induzido para separar os pares de SNs correferentes dos não correferentes. Na última etapa, é realizado o agrupamento das menções referentes à mesma entidade. Vários tipos de algoritmos são utilizados nessa etapa. Por exemplo, no trabalho de Soon et al. (2001) é utilizado um método que escolhe a primeira cadeia para cada par de elementos definidos pelo classificador como correferentes. Ao finalizar esse último processo, o sistema tem as cadeias de correferência identificadas.

Os algoritmos supervisionados são largamente utilizados na literatura e seus resultados para a tarefa de correferência tem mostrado-se satisfatórios quanto às medidas de precisão e cobertura. No entanto, esses algoritmos apresentam alguns problemas. Um deles é a necessidade de corpora anotados, que muitas vezes não estão disponíveis. Outro problema que esse tipo de algoritmo apresenta é o fato de que o modelo de classificação é independente do modelo de agrupamento. Isso implica que a melhoria da precisão e cobertura da classificação não garante diretamente a melhoria do algoritmo como um todo (Ng, 2010). Esse é um grande problema dos métodos supervisionados, pois a tarefa de identificar as cadeias de correferência é tipicamente uma tarefa de agrupamento (Haghighi e Klein, 2007).

Com relação à utilização desse tipo de abordagem em múltiplos documentos, é possível elencar dois problemas. O primeiro é a disponibilidade de um corpus anotado que tenha um tamanho suficiente para o aprendizado. O segundo é o agravamento do problema de desbalanceamento de classe, pois em um cenário de múltiplos documentos, a quantidade de instâncias negativas será muitas vezes maior que das positivas, tornando a indução de um modelo de classificação mais difícil.

Nesse contexto, alguns trabalhos utilizaram algoritmos não supervisionados na tentativa de superar esses problemas. Na seção a seguir é apresentado o funcionamento desse tipo de algoritmo.

3.3.2 Abordagens não supervisionadas

As abordagens não supervisionadas para a resolução de correferência partem do pressuposto de que é possível considerar cada cadeia de correfe-

rência como uma classe alvo de um algoritmo de AM (Cardie e Wagstaf, 1999). Apresenta-se o seguinte exemplo:

(3.3) Palácio do Planalto divulgou por meio do seu blog nesta segunda-feira (1º) vídeo que mostra o presidente Luiz Inácio Lula da Silva$_1$ recebendo a presidente eleita$_2$, Dilma Rousseff$_2$, em uma festa no Palácio do Alvorada na noite de domingo (31). O vídeo tem 17 segundos. Ele mostra a chegada de Dilma$_2$ à residência oficial do presidente da República$_1$. Lula$_1$ abraça e beija a vencedora. O site oficial da Presidência da República também divulgou fotos da comemoração organizada por Lula$_1$ para Dilma$_2$.
A presidente eleita$_2$ chegou por volta de 22h40 da noite de domingo no Palácio da Alvorada. (Fonte: Portal G1. Disponível no endereço http://g1.globo.com/politica/noticia/2010/11/planalto-divulga-imagens-da-festa-de-lula-para-dilma-na-noite-de-domingo.html)

Os SNs sublinhados do Texto 3.3 representam duas cadeias de correferência no texto, diferenciadas pelos números subscritos. Na abordagem não supervisionada, cada cadeia de correferência é considerada uma classe. Portanto, a cadeia constituída pelos SNs “O presidente Luiz Inácio Lula da Silva”, “O presidente da República”, “Lula” e “Lula” é considerada uma classe, assim com a cadeia formada por “A presidente eleita”, “Dilma Rousseff”, “Dilma”, “Dilma” e “A presidente eleita”.

Assim, os algoritmos não supervisionados tentam descobrir o conjunto das menções que pertence a cada classe (cadeia de correferência). Essa forma de modelagem é mais natural, pois é intuitivo pensar que as cadeias de correferência formam grupos distintos de menções a entidades.

Na Figura 3.2 é apresentada uma arquitetura genérica de um sistema de resolução não supervisionado.

Observa-se que, em comparação com a arquitetura supervisionada, ela é bem mais simples. Essa arquitetura é dividida em poucas etapas. São elas: extração dos SNs, extração dos atributos, identificação das cadeias de correferência.

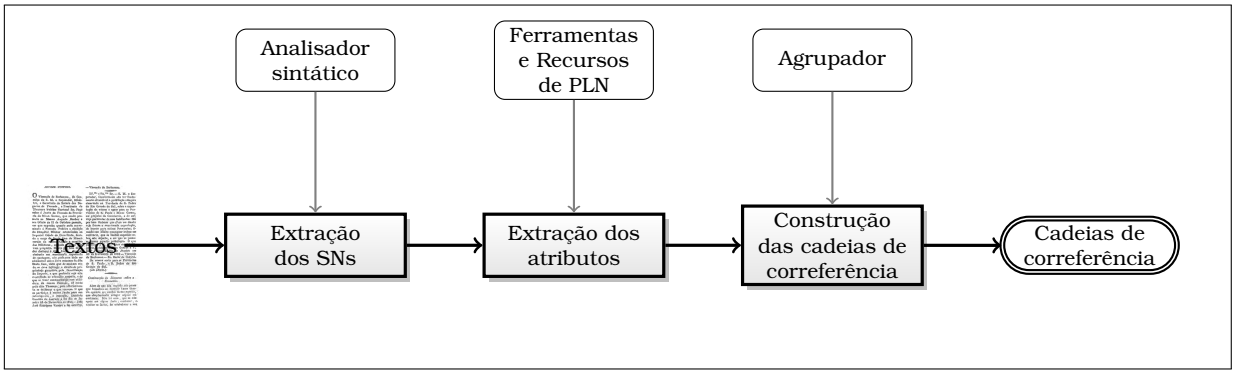

Figura 3.2: *Arquitetura de um sistema de resolução de correferência não supervisionado*

A primeira etapa dessa arquitetura é similar a todo sistema de correferência, no qual os SNs são extraídos utilizando-se um analisador sintático. Na segunda etapa, essa arquitetura diferencia-se de um sistema supervisionado, pois nessa etapa de extração de atributos definem-se apenas as características da menção. Por exemplo, as menções do Texto 3.3: “Dilma Rousseff”, “Dilma”, “O presidente Luiz Inácio Lula da Silva” e “Lula” podem ser descritas por um conjunto de atributos, como: gênero, número e palavras (palavras que constituem o SN retirando-se os artigos, as preposições e as conjunções, retando só as palavras com mais significado como os substantivos e adjetivos). Na Tabela 3.2 são apresentadas as instâncias formadas pela descrição dessas menções.

Sintagmas Nominais	Atributos		
	Gênero	Número	Palavras
“Dilma Rousseff”	Feminino	Singular	{Dilma, Rousseff}
“Dilma”	Feminino	Singular	{Dilma}
“O presidente Luiz Inácio Lula da Silva”	Masculino	Singular	{presidente, Luiz, Inácio, Lula, Silva }
“Lula”	Masculino	Singular	{Lula}

Tabela 3.2: *Descrição dos SNs por um conjunto de atributos utilizados em algoritmos não supervisionados para a resolução de correferência*

Apenas com a utilização dos atributos da tabela é possível constatar que as menções “Dilma Rousseff” e “Dilma” são mais parecidas do que as menções “O presidente Luiz Inácio Lula da Silva” e “Lula”. Todavia, do ponto de vista de um algoritmo, é preciso quantificar o quanto são parecidas ou diferentes essas menções. Então, é necessário determinar uma medida de distância entre as instâncias. No exemplo apresentado da Tabela 3.2, uma medida que poderia ser utilizada é a apresentada na Equação 3.1

$$dist(SN_i, SN_j) = \sum_{a \in F} igualdade_a(SN_i, SN_j) \tag{3.1}$$

Onde F representa o conjunto dos atributos que descrevem uma menção, enquanto a função $iqualdade_a$ verifica se para um determinado atributo, os sintagmas têm valores iguais. Nesse caso, a função retorna 0, caso contrário, retorna 1, se não for o atributo "palavras do SN". Nesse caso, retorna a quantidade de palavras diferentes entre os dois conjuntos.

Na Tabela 3.3 são apresentadas as distâncias entre as instâncias da Tabela 3.2. Nota-se que os elementos com menores distâncias entre si são os que pertencem à mesma cadeia de correferência.

	"Dilma Rousseff"	"Dilma"	"O presidente Luiz Inácio Lula da Silva"	"Lula"
"Dilma Rousseff"	0	1	8	5
"Dilma"	1	0	8	4
"O presidente Luiz Inácio Lula da Silva"	9	8	0	4
"Lula"	5	4	3	0

Tabela 3.3: *Diferenças entre as instâncias utilizando uma medida de distância utilizada em algoritmos não supervisionados*

A última etapa do algoritmo não supervisionado consiste da utilização de um algoritmo de agrupamento de dados que seja capaz de encontrar as classes corretas (cadeias de correferência), utilizando a distância entre os SNs. Vários trabalhos utilizam esse tipo de abordagem como Cardie e Wagstaf (1999), Haghighi e Klein (2007) e Poon e Domingos (2008), obtendo resultados comparáveis aos das abordagens supervisionadas.

Um ponto interessante dessa abordagem é que a sua aplicação para a resolução de correferência em múltiplos documentos é direta, pois a descrição da menção não depende do documento no qual ela se encontra. Esse tipo de abordagem, portanto, possibilita encontrar as cadeias de correferência em mono e múltiplos documentos utilizando a mesma arquitetura apresentada nessa seção.

Após a explanação do desenvolvimento dos sistemas de correferência, passa-se a apresentar a avaliação.

3.4 Avaliação dos sistemas de correferência

A resolução de correferência é vista como uma tarefa intermediária, que pode ser utilizada como parte de outros sistemas. Dessa forma, é possível avaliar a tarefa tanto intrínseca como extrinsecamente.

A avaliação intrínseca é feita através da comparação das cadeias obtidas por um sistema automático com as cadeias anotadas automaticamente. Já a avaliação extrínseca é realizada por meio da utilização de outros sistemas, verificando-se qual é a variação no desempenho de um sistema final com a adição do processo de resolução de correferências.

Os sistemas de correferência podem ser avaliados utilizando-se, por exemplo, sistemas de sumarização, de perguntas e respostas e tradução automática. Esse tipo de avaliação deve considerar a forma como as cadeias de correferência serão utilizadas para melhorar esses sistemas, pois é necessária a definição de como utilizar os resultados do sistema de correferência.

Os trabalhos sobre sistemas de correferência concentram seus esforços na avaliação intrínseca, pois é mais fácil de ser realizada e de se obter resultados que sejam comparáveis e reprodutíveis. Essa comparação é feita através de medidas de avaliação que visam quantificar o desempenho de um sistema de correferência. Dentre essas medidas, a mais utilizada na avaliação dos sistemas de correferência é a proposta por Vilain et al. (1995). Essa medida foi utilizada na competição *Message Understanding Conference* (MUC) (Grishman, 1994), e informa valores de precisão e cobertura para sistemas de correferência. Apresenta-se uma cadeia de correferência no texto 3.4.

(3.4) A casa não é tão bonita. Ela é apenas grande. Mesmo assim o imóvel será vendido logo, pois tem uma boa valorização no bairro.

Para calcular as medidas de precisão e cobertura propostas por Vilain et al. (1995), considera-se que um sistema automático obteve como resultado a cadeia “A casa”, “Ela” e “ imóvel ”. Para avaliar o desempenho desse sistema, seu resultado será comparado com a cadeia anotada manualmente (“A casa”, “Ela” e “imóvel”). Considera-se que cada cadeia de correferência é o conjunto das ligações entre as menções e seu antecedente. Para a cadeia automática tem-se {“A casa”–“Ela”, “Ela”–“imóvel”} e para a cadeia de referência: {“A casa”–“Ela”,“Ela”–“imóvel”}. Para o cálculo da precisão e cobertura, há as seguintes fórmulas:

$$Precisão = \frac{Nº\_de\_ligações\_corretas}{Nº\_de\_ligações\_da\_cadeia\_de\_referência} \tag{3.2}$$

$$Cobertura = \frac{Nº\_de\_ligações\_corretas}{Nº\_de\_ligações\_da\_cadeia\_automática} \tag{3.3}$$

No exemplo 3.4, então, as medidas de precisão e cobertura são $1/2$, pois a quantidade ligações corretas nesse caso é 1, e tanto o número de ligações da cadeia de referência, quanto da automática é 2. As Fórmulas 3.2 e 3.3 são generalizadas por Vilain et al. para tratar do conjunto de todas as cadeias de um texto. Nas Equações 3.4 e 3.5 é apresentada essa generalização.

$$Precisão\_Total = \frac{\sum_{i=1}^{N} N^o\_de\_ligações\_corretas}{\sum_{i=1}^{N} N^o\_de\_ligações\_da\_cadeia\_de\_referência} \tag{3.4}$$

$$Cobertura\_Total = \frac{\sum_{i=1}^{N} N^o\_de\_ligações\_corretas}{\sum_{i=1}^{N} N^o\_de\_ligações\_da\_cadeia\_automática} \tag{3.5}$$

Onde N é o número de cadeias de correferência que estão sendo avaliadas.

Outras medidas de avaliação vêm sendo utilizadas, como a proposta por Bagga e Baldwin (1998b). Nesse trabalho, os autores apresentam a medida B-CUBED que é baseada na medida de Vilain et al.. A medida de Bagga e Baldwin foi particularmente desenvolvida no cenário de múltiplos documentos. A diferença entre as medidas é a adição de pesos às entidades. Apresentam-se as seguintes equações.

$$Precisão\_BCUBED = \sum_{i=1}^{N} w_i * Precisão_i \tag{3.6}$$

$$Cobertura\_BCUBED = \sum_{i=1}^{N} w_i * Cobertura_i \tag{3.7}$$

Onde N é o número de entidades no documento e $w_i = 1/N$ para todas as entidades. Essa duas medidas são as utilizadas na avaliação desse trabalho. Além delas, também é utilizada a $medida\_f$, que é uma média harmônica entre precisão e cobertura. A equação dessa medida é apresentada em 3.8

$$medida\_f = 2 * \frac{Precisão * Cobertura}{Precisão + Cobertura} \tag{3.8}$$

Outro ponto importante na avaliação de sistema de correferência é a definição dos métodos *baselines*. Um método *baseline* trivial utilizado na avaliação de sistemas é a concordância em núcleo dos SNs. Apesar de trivial, esse método obtém valores altos de precisão e cobertura (Yang e Zhou, 2004).

3.5 Considerações Finais

Salienta-se que nesse capítulo, foram apresentadas as principais características linguísticas utilizadas para tarefa de resolução de correferência. Também foram mostrados os principais tipos de algoritmos utilizando aprendizado de máquina para essa tarefa. É importante esclarecer que o algoritmo supervisionado descrito neste capítulo não é aplicável para a resolução inter documento, pois ele explora as características da relação do SN anafórico e seu antecedente, o que, em um cenário em múltiplos documentos, não é possível. Já a abordagem não supervisionada pode ser aplicada para esse tipo de tarefa, pois a descrição das características das menções pode ser realizada tanto em mono como em múltiplos documentos, já que as características não se baseiam na relação anafórica e sim na descrição da própria menção a entidade ocorrida no texto. Outro fato é que as menções tanto em mono como em múltiplos documentos devem compartilhar valores do vetor de atributos, quando pertencerem a mesma cadeia de correferência. Nesta dissertação, é explorada a abordagem não supervisionada para construir as cadeias tanto em inter como intra documento, pelos motivos antes apresentados.

No próximo capítulo, são apresentados detalhadamente os trabalhos nos quais se baseia esta proposta e os resultados que estão sendo alcançados por eles.

CAPÍTULO

4

Trabalhos relacionados

Neste capítulo são descritos alguns trabalhos que utilizam métodos não supervisionados para obter as cadeias de correferência tanto em mono como em múltiplos documentos. Também é apresentado o método de avaliação que esses trabalhos desenvolveram e os resultado que esses algoritmos têm obtidos.

O capítulo é dividido em duas seções: a primeira apresenta trabalhos que tentam resolver correferência apenas em mono documento e a segunda é composta por trabalhos que lidam com a correferência em múltiplos documentos.

4.1 Modelos para resolução de correferência em mono documento

4.1.1 Modelo de Cardie et al. (1999)

O trabalho de Cardie et al. (1999) descreve um dos primeiros algoritmos não supervisionados na área de resolução de correferência. Em seu modelo, eles tentam resolver correferência entre SNs, que são representados por

um conjunto de 11 características. As ligações de correferência são construídas utilizando-se um algoritmo de agrupamento definido especialmente para essa tarefa.

Na fase de obtenção dos SNs, é utilizado o analisador sintático apresentado no trabalhos de Cardie e Pierce (1998). Esse analisador obtém apenas os SNs simples, ou seja, os que na sua estrutura não contêm outros SNs. Apresenta-se o Texto 4.1.

(4.1) A casa do João fica perto da saída.

Nesse texto, o SN "A casa do João" é constituído por dois SNs simples "A casa" e "o João". O analisador sintático utilizado por Cardie et al. (1999), apenas identifica SN simples. Extraídos os SNs, o próximo passo é definir o conjunto de características de cada SN obtido. Na Tabela 4.1 é descrito o conjunto de características que foram utilizadas no modelo. Todos os valores das características são automaticamente extraídos, ou seja, a precisão da extração dessas características não é 100%, o que pode degradar o desempenho do algoritmo proposto.

Para determinar o quanto dois SNs podem ser correferentes utilizando-se o conjunto de características descritas na Tabela 4.1, o modelo usa uma medida de distância. A ideia é a de que quanto menor for a distância entre os SNs maior é a probabilidade de que eles sejam correferentes.

Dois SNs são considerados correferentes caso a distância entre eles seja menor que um limiar determinado empiricamente. A medida é definida como:

$$dist(SN_i, SN_j) = \sum_{f \in A} w_a * incompatibilidade_a(SN_i, SN_j) \quad (4.1)$$

Onde A é o conjunto das características do SN, w_a é o peso de cada característica e a $incompatibilidade_a$ é uma função que retorna um valor entre 0 e 1 inclusive, que indica o grau de incompatibilidade de uma característica a entre SN_i e SN_j. Na Tabela 4.2 é apresentada a função de incompatibilidade e os pesos correspondentes a cada característica.

Cardie et al. (1999) escolhem os pesos na tentativa de representar o conhecimento linguístico sobre a correferência. Termos com o valor de peso ∞

Características	Descrição
Palavras individuais	As palavras contidas no SN. Por exemplo, o SN "A casa do João", teria como valores [A, casa, do, João].
Núcleo do SN	A última palavra do SN é considerada a núcleo do SN (No caso do inglês)
Posição	Posição do SN no texto. Os SNs são numerados sequencialmente, começando do início do documento.
Tipo de Pronome	Armazena os tipos dos pronomes nominativo, acusativos, possessivos e ambíguos (No caso do inglês, língua para qual esse trabalho foi realizado, pronomes ambíguos são os pronome *you* e *it*).
Artigo	Verdadeiro se o tipo de SN é definido ou falso caso contrário.
Aposto	Verdadeiro se o SN é um aposto de outro SN e falso caso contrário.
Número	Número do núcleo do sintagma (plural ou singular)
Nome próprio	Verdadeiro se SN é um nome próprio e falso caso contrário.
Classe semântica	Define a classe semântica do núcleo do sintagma utilizando a WordNet. As classes utilizadas são: *time*, *city*, *animal*, *human*, *object* e *number*, *money*, e *company*
Gênero	Gênero (masculino, feminino ou neutro) do SN utilizando a WordNet.

Tabela 4.1: *Conjunto de características utilizadas no trabalho de Cardie e Wagstaf (1999)*

representam filtros que determinam se dois SNs têm os valores incompatíveis para alguma característica. Nesse caso, eles não podem ser correferentes.

Quando o peso de uma característica é $-\infty$ e os valores da característica dos SNs são compatíveis, então esses SNs são correferentes, não importando os valores das outras características. As características com o valor do peso igual a r são determinadas empiricamente durante a execução do algoritmo de agrupamento. O valor de r é utilizado para determinar até que distância dois SNs podem ser considerados correferentes para as características posição e artigo. Os outros pesos (Palavras, Núcleo) foram obtidos realizando-se uma análise de corpus.

Há casos em que é necessário computar a soma entre os valores de $-\infty$ e ∞. Essa abordagem assume que a soma é igual a ∞, pois dessa forma a medida de distância dá prioridade aos SNs não correferentes.

Características	Pesos	Função de Incompatibilidade
Palavras	10.0	(número de palavras diferentes) / (número de palavras do SN mais extenso)
Núcleo	1.0	1 se o núcleo do SN for diferente, caso contrário 0
Posição	r	1 se o SN_i é um pronome e SN_j não for, caso contrário 0
Artigo	r	1 se o SN_j é indefinido, caso contrário 0
Sub-cadeias	$-\infty$	1 se o SN_i contém a cadeia de caracteres de SN_j
Aposto	$-\infty$	1 se o SN_j é aposto do SN_i, caso contrário 0
Número	∞	1 se os dois SNs não concordam em número, caso contrário 0
Classe semântica	∞	1 se os dois SNs não concordam quanto à classe semântica, caso contrário 0
Gênero	∞	1 se os dois SNs não concordam em gênero, caso contrário 0
Animado	∞	1 se os dois SNs não concordam no atributo animado, caso contrário 0

Tabela 4.2: *Função de incompatibilidade e os pesos para cada termo na medida de distância utilizada no método de Cardie et al. (1999)*

Definida a medida de distância, o próximo passo é realizar o agrupamento dos SNs correferentes. Inicialmente o algoritmo considera que cada SN representa uma classe. Então, o algoritmo de agrupamento inicia-se a partir do fim do documento, comparando cada SN com todos os predecessores. Se a distância do SN for menor que o r, então é feita a junção das duas classes em uma só. Duas classes podem ser agrupadas se não houver nenhuma restrição de incompatibilidade entre elas.

Essa abordagem foi avaliada utilizando-se o corpus da competição MUC-6 (MUC-6, 1995) e as medidas propostas por Vilain et al. (1995). Para validar sua abordagem, Cardie et al. (1999) compararam seus resultados com três algoritmos *baselines*.

O primeiro *baseline* define todos os SNs como correferentes, ou seja, no documento existe apenas uma classe, obtendo-se o resultado de 44,8% de medida-f. Esse *baseline* é utilizado para definir o limite inferior da abordagem de Cardie et al. (1999). O segundo *baseline* considera correferentes os SNs que têm palavras em comum, e produz um resultado de 41,3%. Por fim, o terceiro

baseline considera correferentes os SNs que têm o mesmo núcleo, produzindo resultados de 45.7% de medida-f.

Para avaliar o sistema proposto, Cardie et al. (1999) utilizam diferentes valores de r em um intervalo de 1 a 10, em um corpus de teste para que depois, com o valor de r definido, realizar o teste no corpus da MUC-6. Os resultados dessa avaliação podem ser vistos na Tabela 4.3.

Observa-se que o valor $r = 4$ foi o que obteve o melhor resultado quanto à medida-f (52,8%). Para os dados de avaliação utilizando-se o valor de $r = 4$, o algoritmo obteve o desempenho de 54% de medida-f.

r	Cobertura	Precisão	Medida-f
1	34,6	69,3	46,1
2	44,7	61,4	51,7
3	47,3	58,5	52,3
4	**48,8**	**57,4**	**52,8**
5	49,1	56,8	52,7
6	49.8	55,0	52,3
7	50,3	53,8	52,0
8	50,7	53,0	51,8
9	50,9	52,5	51,7
10	50,9	52,1	51,5

Tabela 4.3: *Desempenho dos dados de teste para diferentes valores de r no trabalho de Cardie et al. (1999)*

Cardie et al. (1999) também compararam seus resultados com algoritmos supervisionados que participaram da competição MUC-6. O melhor sistema obteve valor de 65% de medida-f e o pior de 40%. Como pode ser observado, o sistema de Cardie et al. (1999) obteve um valor intermediário.

Apesar de os valores obtidos não serem os melhores da literatura, esta abordagem apresenta-se promissora, pois não utiliza dados para treinamento, ou seja, não requer corpus anotados para induzir o modelo. No entanto, como pode ser observado, este método exige os ajustes dos parâmetros do algoritmo, no caso, o valor de r. Para isso, é necessária a utilização de um corpus anotado. Outro ponto é o recorte que foi realizado quanto aos SNs extraídos, pois utiliza-se apenas os SNs simples, não tratando portanto apenas de um subconjunto dos SNs. Realizando, portanto, uma tarefa mais simples que a de resolução de correferência utilizando todo o conjunto de SNs.

4.1.2 Modelo de Haghighi e Klein (2007)

Haghighi e Klein (2007) propuseram uma abordagem que utiliza um modelo não supervisionado de aprendizado. Os resultados obtidos por esse método são próximos aos resultados das abordagens supervisionadas.

Essa abordagem segue as mesma etapas do método utilizado por Cardie et al. (1999), extraindo-se os SNs de forma automática e descrevendo-os segundo um conjunto de características, e por fim, utilizando um algoritmo de agrupamento para construir as cadeias de correferência.

O conjunto de características utilizados nesse trabalho é composto de: tipo da entidade (pessoa, local, organização e diverso), gênero (masculino, feminino e neutro), número (singular e plural) e núcleo do sintagma.

Como algoritmo de agrupamento, Haghighi e Klein utilizam o método de agrupamento de *Dirichlet* (Teh et al., 2006). Esse modelo baseia-se em assumir uma distribuição estatística a *priori* para o conjunto das menções e então utilizar métodos de inferência para obter a distribuição a *posteriori*. Nesse método de agrupamento não há a necessidade da definição do número de classes, o que torna sua utilização adequada em sistemas de correferência, já que não se conhece a *priori* o número de entidades mencionadas (quantidade de cadeias de correferência). Observa-se que no trabalho de Haghighi e Klein não se define uma medida de distância, pois assume-se que as menções descritas pelo vetor de características distribuem-se no espaço segundo uma distribuição β[1].

A avaliação desse sistema foi realizada com o corpus do MUC-6, obtendo-se 80,8 de precisão, 52,8 de cobertura e 63,9 de medida-f. Também foi utilizado o corpus ACE 2004 (Mitchell et al., 2004) para a avaliação. Os resultados foram: 66,7 de precisão, 62,3 de cobertura e 64,2 de medida-f. Dessa avaliação é possível verificar que os métodos de aprendizado não supervisionados têm evoluído. Comparando-se os resultados com os de Cardie et al. (1999) é verificado um aumento de 21% na medida-f. Observa-se que essa diferença pode ser acentuada se for levado em conta que os resultados são obtidos não só os utilizando SNs simples, mas também SNs que contêm outros SNs.

Um ponto importante do trabalho de Haghighi e Klein é que, além de tratar do fenômeno de correferência em mono-documento, também aborda

[1]Uma distribuição β é uma distribuição contínua definida no intervalo entre [0,1] e é parametrizada com dois valores. Essa distribuição é frequentemente utilizada quando se tenta descrever uma distribuição e não se conhece os valores das probabilidades.

o seu modelo de ponto de vista de correferência em múltiplos documentos, apesar de não realizar a avaliação desse modelo. O modelo discutido nesta dissertação é baseado nesse modelo proposto por Haghighi e Klein (2007).

4.2 Modelos de resolução de correferência em múltiplos documentos

4.2.1 Modelo de Bagga e Baldwin (1998b)

Um dos primeiros trabalhos relacionados à resolução de correferência em múltiplos documentos foi o trabalho de Bagga e Baldwin (1998b). O método proposto por Bagga e Baldwin (1998b) utiliza o modelo *bag of words* (Salton, 1989), para relacionar os textos que tratam de uma mesma entidade. O sistema desenvolvido por Bagga e Baldwin (1998b) é apresentado na Figura 4.1.

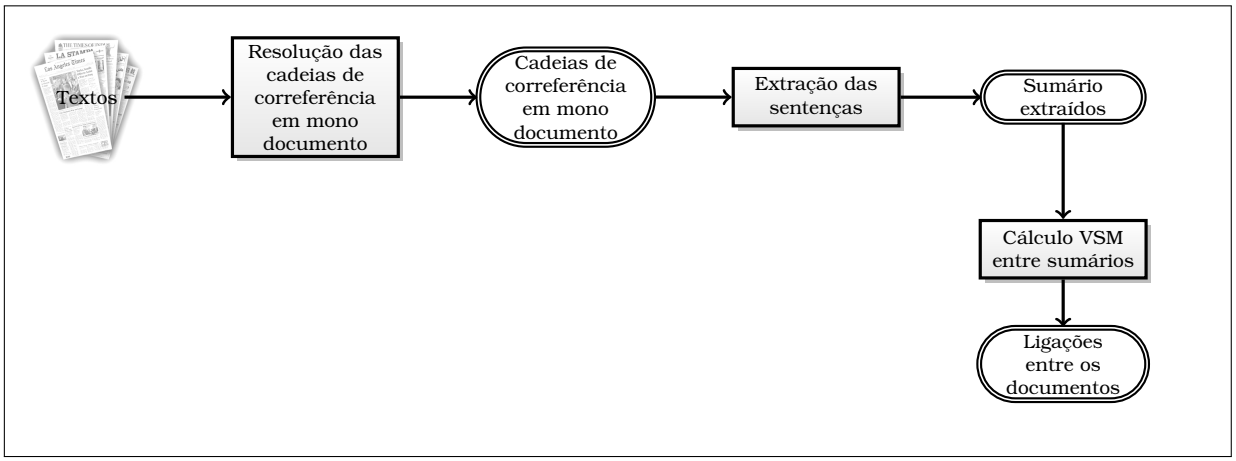

Figura 4.1: *Arquitetura do sistema de resolução de correferência em múltiplos documentos proposto por Bagga e Baldwin (1998b)*

Na Figura 4.1 percebe-se que o sistema recebe como entrada um conjunto de documentos. Cada documento é submetido a um sistema de resolução de correferência em mono-documento para a identificação das cadeias de correferência. No próximo passo, é realizada uma consulta por nome. Para cada documento que contém o nome procurado, é identificada a cadeia de correferência à qual pertence a menção, e então, é produzido um extrato com as sentenças que contêm as menções dessa cadeia de correferência. Com os extratos realizados, o próximo passo é a transformação desses em um vetor

de termos com os pesos definidos pela medida de frequência ponderada *tf-idf* (Equação 4.2):

$$tf\text{-}idf(t_j, d_i) = freq(t_j, d_i) \times \log \frac{N}{d(t_j)}, \tag{4.2}$$

onde (t_j, d_i) é o termo j no documento i, $freq(t_j, d_i)$ é a frequência desse termo no documento, N é o número de documentos e $d(t_j)$ é o número de vezes que o termo ocorre no documento j. Então é realizada a aplicação do algoritmo de agrupamento *Vector Space Model* (VSM) proposto por Salton (1989) para definir os extratos similares e assim obter as cadeias de correferência. O VSM obtém as cadeias de correferência incrementalmente. Inicialmente, uma cadeia unitária é criada com um vetor de termos de um extrato. Então é calculada a similaridade entre essa cadeia e o próximo vetor utilizando-se o cálculo de similaridade pelo ângulo do cosseno (Equação 4.3):

$$cosseno(d_1, d_2) = \frac{\sum_{i=1}^{n} w_{i,d_1} \times w_{i,d_2}}{\sqrt{\sum_{i=1}^{n} (w_{i,d_1})^2} \times \sqrt{\sum_{i=1}^{n} (w_{i,d_2})^2}}, \tag{4.3}$$

onde d_1 e d_2 representam os documentos, w_{i,d_j} o peso do termo i no documento d_j, e n o número de documentos. Se a similaridade for maior que um limiar, que é ajustado quando na execução do método, então as entidades são correferentes.

Para avaliação, Bagga e Baldwin (1998a) criaram um *corpus* com entidades ambíguas. O *corpus* foi constituído por 197 notícias do jornal *The New York Times*. Todos os textos têm o nome próprio *John Smith* ou uma variação com nomes do meio. As cadeias de correferência foram anotadas manualmente. No *corpus* foram encontrados 35 *John Smith* diferentes mencionados nas notícias, sendo que 24 deles foram mencionados apenas um vez. Bagga e Baldwin (1998a) assumem que cada documento trata apenas de um *John Smith*.

Os resultados obtidos por Bagga e Baldwin (1998a) são avaliados utilizando-se o algoritmo de Vilain et al. (1995) e o *B-CUBED*, sendo que este último foi criado por Bagga e Baldwin na tentativa de obter uma avaliação mais intuitiva de métodos para resolução de correferência em múltiplos documentos. Utilizando-se a medida do algoritmo de Vilain et al. (1995) o método de Bagga e Baldwin (1998a) obteve um valor de 83% de medida-f e para o algoritmo *B-CUBED* de 84,6%.

Deve-se, no entanto, atentar que os valores obtidos, apesar de altos, não representam valores que devem ser tidos como parâmetros para a tarefa

de resolução de correferência por dois motivos principais. O primeiro motivo é que o algoritmo é avaliado apenas com uma entidade em diferentes textos, o que pode caracterizar essa tarefa como desambiguação de nomes, uma tarefa mais simples do que resolver os conjuntos das cadeias de todos os documentos. O segundo é a sensibilidade do algoritmo ao limiar definido para realizar o corte no dendrograma do algoritmo de agrupamento aglomerativo e, assim, determinar a quantidade de cadeias de correferência existente. A definição desse limiar pode ser crítica e nesse trabalho é feita a definição no próprio conjunto de testes, o que inviabiliza a utilização desse limiar em outros grupos de textos.

4.2.2 Modelo de Baron e Freedman (2008)

O modelo apresentado em Baron e Freedman (2008) para resolução de correferência em múltiplos documentos é mais completo que o de Bagga e Baldwin, pois realiza a identificação das ligações de correferência entre os documentos, não se limitando a uma entidade, mas ao conjunto das entidades dos textos. Sua prosposta faz uso de métodos de extração da informação para obter um conjunto de características de cada menção, para serem submetidas a algoritmos de agrupamento. Na Figura 4.2 é apresentada a arquitetura do modelo.

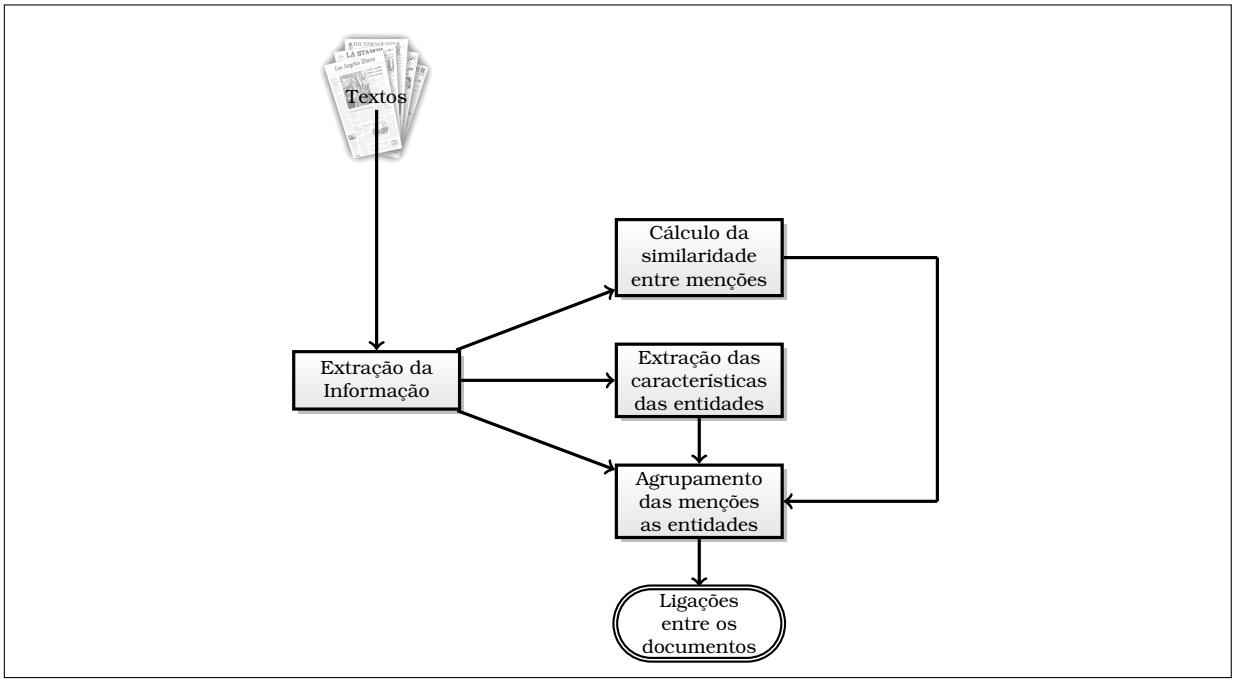

Figura 4.2: *Arquitetura do sistema de resolução de correferência em múltiplos documentos proposto por Baron e Freedman (2008)*

Como é apresentado na Figura 4.2, o método utiliza um sistema de extração da informação, o SERIF (Ramshaw et al., 2001). O SERIF extrai um conjunto de informações do texto que são utilizadas nas etapas seguintes. Esse conjunto é composto por: menções às entidades que ocorrem nos textos, tipo das menções (Pessoa, Organização ou Entidade Geo-Política), informações dos relacionamentos das menções e eventos nos textos. Na etapa do cálculo da similaridade entre menções é utilizado o conjunto de menções extraídas pelo SERIF. A similaridade é calculada utilizando-se a distância de edição de Levenshtein (1965). Já na etapa de extração das característica das menções o resultado do SERIF é utilizado para construir um vetor que descreva cada menção. Com essas etapas concluídas é realizada a construção das cadeias de correferência utilizando os dados fornecidos por essas três etapas. Nessa etapa é utilizado um algoritmo de agrupamento de dados aglomerativo. O algoritmo inicia considerando cada menção como sendo uma classe (cadeia de correferência). Então, é calculada a distância entre as menções. O par de menções que tiver o menor valor de distância é agrupado, ou seja, as duas menções agora pertencem à mesma classe. Esse processo continua até que as distâncias entre as classes sejam todas maiores que um determinado limiar. O limiar é determinado experimentalmente utilizando-se um corpus de testes.

Para a avaliação, Baron e Freedman utilizaram o corpus ACE 2005 (Walker et al., 2005). Sua avaliação quanto à medida B-CUBEB foi de 52,6% de medida-f. Outras avaliações são consideradas utilizando-se uma variação da medida B-CUBEB, considerando com peso maior as entidades que são pessoas. Utilizando essa medida, o sistema obteve para o melhor valor de medida-f 71,5. Esses resultados foram comparados com dois *baselines*. O primeiro *baseline* considera cada menção como sendo uma cadeia de correferência. O segundo considera duas menções correferentes se elas têm a concordância exata em cadeias de caracteres. Os resultados obtidos foram 50,0 e 65,44 de medida-f, respectivamente, utilizando-se o B-CUBEB alterado.

O modelo de resolução de correferência em múltiplos documentos proposto por Baron e Freedman explora pontos importantes da tarefa de resolução de correferência. O principal deles é a avaliação do sistema utilizando um conjunto de menções de entidades diferentes.

4.3 Considerações Finais

Neste capítulo foram apresentados métodos diferentes para resolver correferência. Esses métodos diferem-se quanto ao tipo de menção à entidade

que é tratada, aos tipos de atributos utilizados e aos métodos de agrupamentos. Quanto ao tipo de menção, é possível listar três tipos: os SNs simples (Cardie et al., 1999), os SNs de todos os níveis (Haghighi e Klein, 2007) e as entidades mencionadas (Bagga e Baldwin, 1998a; Baron e Freedman, 2008). Com relação aos atributos, são utilizados aqueles que descrevem a menção. Como método de agrupamento, são predominantemente utilizados os algoritmos de agrupamento hierárquico, que têm como principal problema a determinação do limiar, que é necessário para definir o corte no dendrograma com o intuito de identificar o número de cadeias de correferência. No entanto, o trabalho de Haghighi e Klein (2007) apresenta uma alternativa, o algoritmo de agrupamento baseado no processo de *Dirichlet*. Com esse algoritmo não há necessidade de informar o número de cadeias de correferência, ou seja, o número de classes dos dados para agrupamento.

Outro ponto importante é o desenvolvimento que vem sendo apresentado pela área de aprendizado não supervisionado para resolução de correferência, como pode ser visto quando é comparado os resultados, por exemplo de Cardie et al. (1999) e Haghighi e Klein (2007). A evolução dos resultados deve-se, principalmente, a dois fatores: melhoria das ferramentas que extraem as característica linguísticas e a própria utilização de algoritmos de AM mais robustos.

Nesse contexto, esta dissertação apresenta um método que explora o algoritmo de *Dirichlet* com objetivo de resolver correferência em múltiplos documento e em mono-documento, entre os SNs de todos níveis.

No próximo capítulo é apresentado o método que foi desenvolvido no âmbito deste mestrado.

CAPÍTULO

5

MemexLink - Um sistema de resolução de correferência em múltiplos documentos

Neste capítulo é apresentado o método implementado na construção de um protótipo para um sistema de resolução de correferência em mono e múltiplos documentos. Esse protótipo foi desenvolvido para a língua portuguesa, no entanto, essa é uma instanciação para a validação do método. O método baseia-se em algoritmos não supervisionados com a combinação de regras simbólicas. Na literatura, como foi visto no Capítulo 4 geralmente são utilizados apenas os algoritmos de agrupamento não supervisionado. A adição de regras a esse tipo de arquitetura visa obter melhores resultados do que os sistemas que utilizam apenas agrupamento. As regras utilizadas são baseadas na análise de corpus e na tentativa de solução dos problemas que o método de agrupamento revelou. As limitações do algoritmo de agrupamento devem-se principalmente ao fato de que o conhecimento utilizado por esse algoritmo não compreende parte das variações do fenômeno de correferência.

O método utilizado é dividido basicamente em duas fases: a de identificação das menções (SNs) e características e a de identificação das cadeias de correferência.

O protótipo desenvolvido, denominado de MemexLink[1], é descrito na Figura 5.1.

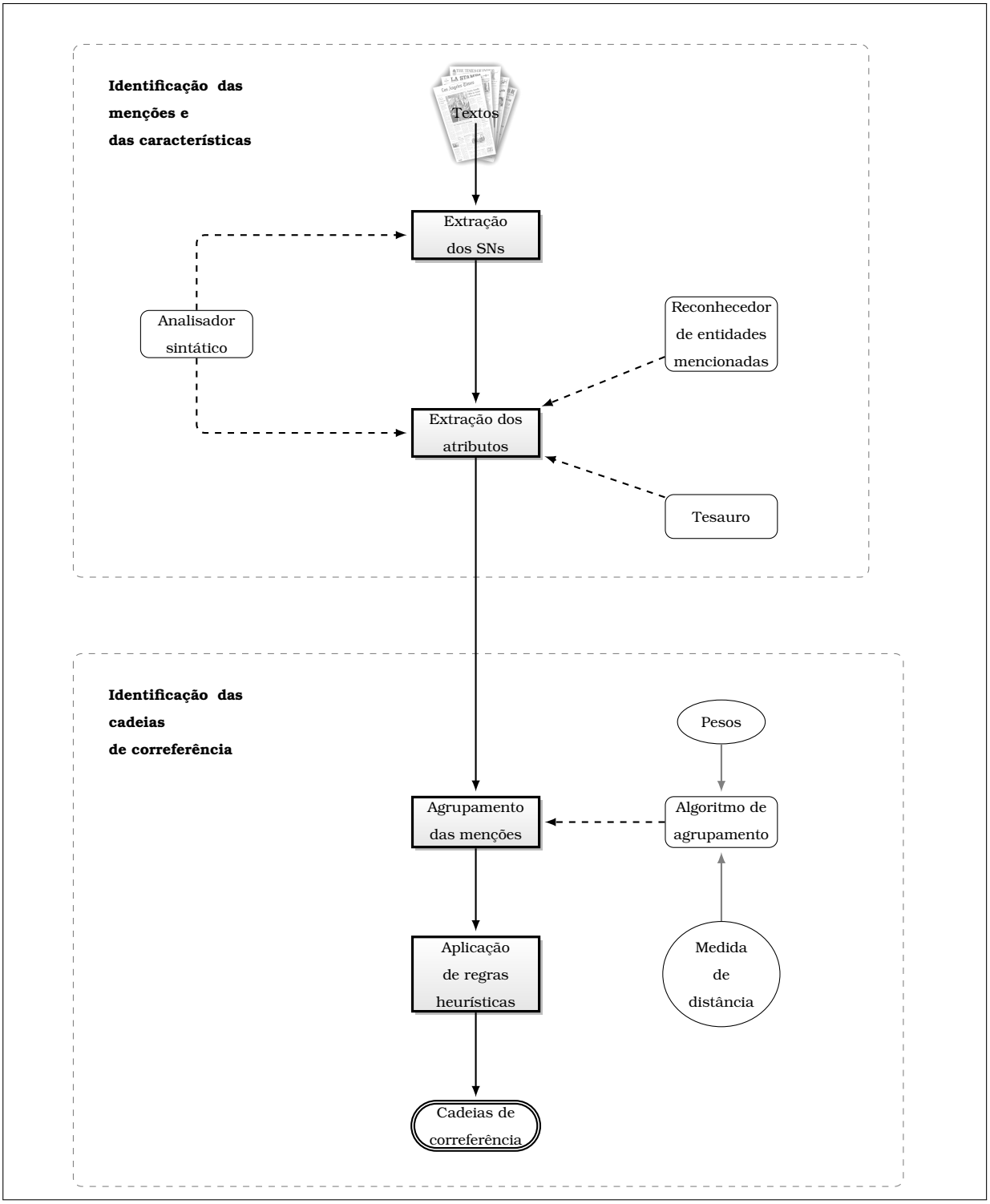

Figura 5.1: *Arquitetura do sistema de resolução de correferência em múltiplos documentos proposto nessa dissertação*

[1]O nome MemexLink foi dado ao sistema em homenagem e alusão à máquina Memex. A máquina foi descrita por Vannevar Bush em 1945 no ensaio *As We May Think*. Esse ensaio descreve uma máquina capaz de organizar grande quantidade de informação através de ligações e associações. Esse trabalho é considerado o precursor da ideia de hipertexto.

Na Figura 5.1, a primeira fase, a de identificação das menções (SNs) e características, tem como entrada um conjunto de textos. Esses são textos jornalísticos que tratam de um mesmo assunto, que foram previamente agrupados, já que este protótipo não é realizada essa etapa de agrupamento. É identificado o conjunto dos SNs do texto, utilizando-se um analisador sintático. Para extrair o conjunto de atributos são utilizados um analisador sintático, um reconhecedor de entidades mencionadas e um tesauro. Cada ferramenta ajuda na extração de uma ou mais características dos SNs. Ao fim dessa fase tem-se os SNs com as características extraídas.

A segunda etapa, a de identificação das cadeias de correferência, recebe como entrada a saída da primeira fase e com essa informação realiza o agrupamento das menções em cadeias. A fase inicia com a utilização do método não supervisionado de AM para um primeiro agrupamento. Nessa fase é utilizada a medida de distância desenvolvida e pesos para cada característica. Após o agrupamento, é realizada a aplicação das regras heurísticas para tentar melhorar a qualidade das cadeias geradas. Por fim, têm-se as cadeias de correferência identificadas.

Apresenta-se, em seguida, em detalhes cada processo do MemexLink. Também são descritos as ferramentas e recursos que foram utilizados no sistema.

5.1 Extração dos sintagmas nominais no MemexLink

Para a extração dos sintagmas nominais o sistema utiliza o analisador sintático Palavras (Bick, 2000). Através da anotação obtida pelo Palavras, o sistema identifica os sintagmas nominais em todos níveis. Observa-se o Texto 5.1.

(5.1) O avião explodiu e se incendiou, acrescentou o porta-voz de a ONU em Kinshasa , Jean-Tobias Okala. “Não houve sobreviventes” , disse Okala. (Fonte: Jornal de Brasília)

No texto 5.1 é apresentado os SNs que são anotados pelo Palavras. Um ponto que deve ser acrescentado é que, apesar do Palavras não anotar como SN os nomes próprios sem modificadores, nesse sistema eles são considerados SNs. Para identifica-los, o sistema utiliza a própria marcação de nomes próprios que

o Palavras define, e verifica se pertence a um SN maior. Se não, marca o nome próprio como SN. No texto, o nome próprio “Okala” foi marcado utilizando-se esse algoritmo. O nome próprio “ONU” não foi marcado separadamente, pois pertence ao SN “a ONU”.

Outra estratégia adotada é quanto aos SN que têm um aposto, geralmente separados por vírgulas. A anotação do Palavras define tudo como um SN, mas neste trabalho é considerado que o aposto e o SN que o contém são diferentes.

(5.2) O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.

No texto 5.2 os SNs sublinhados “uma companhia congolesa” e seu aposto “a Trasept Congo”, são marcados como SNs separadamente. Esse método é adotado, pois apesar das menções serem correferentes, elas apresentam características bem distintas do ponto de vista lexical, principalmente.

5.2 Características utilizadas no MemexLink

A definição das características utilizadas neste sistema foi baseada nos trabalhos de Cardie e Wagstaf (1999), Haghighi e Klein (2007) e Souza et al. (2008). Nesses trabalhos, como já foi apresentado no Capítulo 4, é utilizado um conjunto de atributos para descrever as menções que ocorrem nos textos. Os atributos utilizados pelo MemexLink são descritos na tabela 5.1.

Os atributos núcleo do SN, número, gênero e nome próprio são obtidos utilizando-se o analisador sintático Palavras. Para o atributo núcleo do SN também é utilizado o TeP[2] (Maziero et al., 2008) para se obter os sinônimos do núcleo quando ele é um nome comum. Essa estratégia é adotada com o intuito de descobrir se no conjunto de SNs existem núcleos que pertencem ao mesmo *synset* e, assim, utilizar um representante do *synset* com núcleo e não o próprio núcleo do SN. Observe o exemplo 5.3

[2]TeP (Maziero et al., 2008) é um tesauro eletrônico para o Português do Brasil que armazena conjunto de formas lexicais sinônimas e antônimas.

Características	Descrição
Núcleo do SN	Nome comum ou um nome próprio que é o núcleo do SN.
Número	Número do núcleo do sintagma (singular, plural ou neutro (a mesma forma da palavra tanto para singular e como para o plural)).
Gênero	Gênero do núcleo do sintagma (masculino, feminino ou neutro(a mesma forma da palavra tanto para feminino e como para masculino)).
Classe semântica	Classe semântica no SN utilizando as etiquetas definidas para competição Harem (Mota e Santos, 2008). Essas etiquetas tipificam as entidades nomeadas encontradas no texto. No Apêndice A é apresentado o conjunto dessa etiquetas.
Pseudônimo	Menor cadeia de caracteres possível de identificação do SN. Esse atributo só é utilizado para SN cujo núcleo é um nome próprio. E é definido com sendo o último nome do nome próprio, por exemplo, no SN "A República Democrática do Congo", cujo núcleo está sublinhado, o valor do pseudônimo é o nome "Congo".
Nome próprio	Verdadeiro se o núcleo do SN é um nome próprio e falso caso contrário.
Definido	Verdadeiro se SN é definido e falso caso contrário

Tabela 5.1: *Conjunto de características utilizadas pelo MemexLink*

(5.3) A casa era muito bonita. A residência poderia valer muito.

No exemplo, os SNs destacados que têm como núcleos "casa" e "residência", poderiam ser os dois representados apenas pelo núcleo "casa", pois no TeP existe um *synset* que contém os dois nomes. Esse tipo de estratégia reduz a variedade de núcleos e possibilita a descoberta de elementos correferentes como os apresentados no exemplo.

Já as classes semânticas são obtidades atráves do anotador de entidades mencionadas Rembrandt[3] (Cardoso, 2008). Um problema quanto a esse tipo de abordagem é que a relação entre entidades mencionadas e os SNs anotados pelo Palavras não é direta. Então, é verificado se os SNs sem os artigos foram anotados como entidades mencionadas automaticamente.

[3]O Rembrandt (**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em Relações e **AN**álise **D**etalhada do **T**exto) é um sistema de reconhecimento de entidades mencionadas (REM) e de detecção de relações entre entidades (DRE), projetado para reconhecer todo o tipo de entidades mencionadas (EM) em textos escritos em português.

5.3 Representação das características das menções no MemexLink

Após ser extraído o conjunto de características dos SNs, é necessário representar essas características de forma que seja possível o aprendizado. Existem questões de projeto de sistema baseados em AM que fazem parte dessa etapa, como o conjunto de características que deve ser representado de forma a possibilita o aprendizado das regularidades encontradas nos dados.

Observada a natureza do conjunto das características, é possível verificar que todos os atributos são categóricos, ou seja, os valores que eles podem assumir são nomes de classes diferentes (Tan et al., 2005). Assim sendo, as operações possíveis para esse valores são apenas de $=$ e $\neq$. Observa-se na Tabela 5.2 as características das menções encontradas no Texto 5.4.

(5.4) O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes (Fonte: Folha Online, 2009).

Características	Sintagmas Nominais			
	"O avião acidentado"	"Air Traset"	"14 passageiros"	"três tripulantes"
Núcleo do SN	avião	Air Traset	passageiros	tripulantes
Número	singular	singular	plural	plural
Gênero	masculino	feminino	masculino	masculino
Classe semântica	COISA	ORGANIZACAO	-	-
Pseudônimo	-	Traset	-	-
Nome próprio	NAO	SIM	NAO	NAO
Definido	SIM	SIM	SIM	SIM

Tabela 5.2: *Exemplo de um conjunto de características extraídas pelo MemexLink.*

Os valores que as características na Tabela 5.2 assumem são todos categóricos, alguns têm a quantidade de valores possíveis fixos como o gênero e número, já outros como o núcleo tem uma quantidade indeterminada a *priori*. Por exemplo, para a característica gênero existe apenas 3 valores possíveis (singular, plural e neutro), já para características como o núcleo do atributo esse valor será determinado pela quantidade de SNs com núcleos diferentes. Na Tabela 5.2 são 4 valores possíveis, já que existem 4 núcleos diferentes. Para tratar esse conjunto de atributos de forma que seja possível o aprendizado automático existem basicamente duas opções: (1) o algoritmo de AM pode tratar com dados categóricos ou (2) deve-se transformar os dados em

atributos numéricos. O algoritmo de agrupamento utilizado no MemexLink, como é detalhado na seção 5.4, é o *Dirichlet*. Esse algoritmo não trata atributos categóricos. Então, para possibilitar o aprendizado, os valores das características foram transformados em atributos numéricos. Foi utilizada a transformação em um vetor binário (0 ou 1). Essa transformação consiste em criar um vetor de n posições para cada característica, em que n é o número de valores que a característica pode assumir. Para cada posição do vetor representa-se, então, um possível valor para a característica. Apenas uma célula do vetor recebe 1, que caracteriza o valor que aquela instância assume, os outros devem ser 0. Na Tabela 5.3 é apresentado o vetor de cada característica para o SN "O avião acidentado" do exemplo 5.4. Observa-se que na Tabela 5.3 o núcleo do SN é representado por 4 *bits*, pois é a quantidade de valores que esse atributo pode assumir no trecho de texto 5.4 (veja tabela 5.2). Assim como, a quantidade de *bits* para as outras características é 2, pois cada uma delas, no exemplo 5.4, só assume 2 valores distintos.

"O avião acidentado"				
Núcleo do SN	1	0	0	0
Número	1	0		
Gênero	1	0		
Classe semântica	1	0		
Pseudônimo	1	0		
Nome próprio	1	0		
Definido	1	0		

Tabela 5.3: *Exemplo da forma de representação das características de uma menção no MemexLink*

A representação por um vetor binário é adequada, pois não adiciona correlação ou noção de ordem nos valores das características. No entanto, dependendo da quantidade de valores possíveis para atributos como o núcleo do SN, esse vetor pode tornar-se muito esparso. Um vetor esparso pode dificultar o aprendizado das classes (cadeias de correferência) por parte do algoritmo de agrupamento, no entanto, o algoritmo utilizado mostrou-se robusto a esse problema, como é mostrado no Capítulo 5.

Um caso especial de atributo é a classe semântica, pois as classes semânticas utilizadas nesse protótipo são baseados no Harem (Mota e Santos, 2008). Essa classes obedecem a uma hierarquia, como pode ser visto no Apêndice A. Então, para representar esse atributo foi utilizada uma representação que tenta manter a hierarquia ao mesmo tempo que facilita a transformação dos valores em vetores binários. Observa-se na Tabela 5.4 a hierarquia de Organização. Nesse protótipo, a forma utilizada é criar novas classes

com os valores correspondentes ao caminhamento da raiz para os níveis inferiores. No caso desse exemplo as classes seriam: organização-administração, organização-empresa, organização-instituição e organização-outro.

Organização	Administração
	Empresa
	Instituição
	Outro

Tabela 5.4: *Exemplo de categorias semânticas do Harem (Mota e Santos, 2008)*

As classes semânticas definidas pelo Harem têm no máximo três níveis de profundidade. Um anotador semântico pode definir uma entidade mencionada apenas com os níveis mais altos da hierarquia ou especificá-la. Com base nessa informação foi necessário então dividir o atributo classe semântica em três níveis numerados de 1 a 3. O primeiro nível são as classes mais gerais, as classes subsequentes são especificações.

5.4 Algoritmo de agrupamento utilizando no MemexLink

Após definir o conjunto de características e sua forma de representação, a próxima etapa é agrupar os SNs. Nessa etapa deve-se estabelecer qual o algoritmo de agrupamento pode ser utilizado para o problema de resolução de correferência. Na literatura os algoritmos mais utilizados são os aglomerativos, como pode ser visto nos trabalhos de Cardie et al. (1999), Bagga e Baldwin (1998b) e Baron e Freedman (2008). No entanto, esses algoritmos têm a necessidade de determinar um limiar, para que seja possível determinar a quantidade de grupos. Decidir o limiar também é conhecido como determinar o corte no dendrograma. Apresenta-se o dendrograma na Figura 5.2.

Considera-se que o limiar na Figura 5.2 seja representado pela linha pontilhada. Observa-se que a primeira linha de cima para baixo define os dados como sendo um conjunto de 3 classes, são elas: $\{A, B, C\}, \{D\} e \{E\}$. Já a segunda 4^a classe ($\{A, B\}, \{C\}, \{D\} e \{E\}$) e a última define as classes com apenas um elemento ($\{A\}, \{B\}, \{C\}, \{D\} e \{E\}$). Verifica-se que quanto maior o limiar, menor é o número de classes que serão obtidas pelo algoritmo de agrupamento. E quanto menor o limiar, maior é número de classe obtidas. A definição do valor do limiar em um algoritmo de agrupamento hierárquico é que define a quantidade de classe obtidas pelo método.

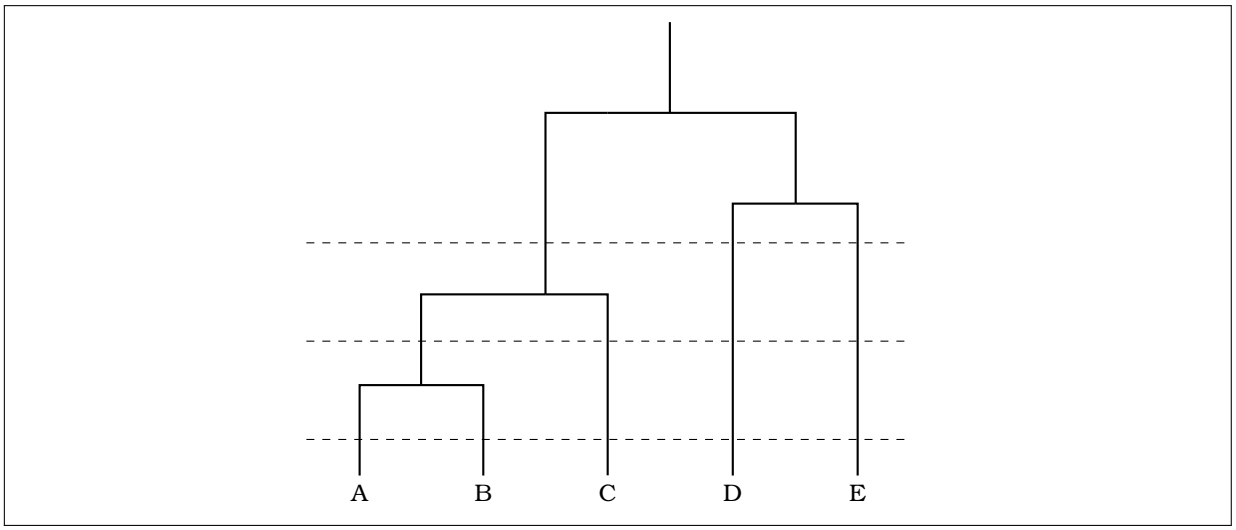

Figura 5.2: *Dendrograma exemplo para demonstrar a influência da escolha do limiar*

Uma solução para o problema da necessidade de determinação de um limiar para identificar a quantidade de cadeias de correferência foi apresentada no trabalho de Haghighi e Klein (2007) no qual é mostrada uma solução que utiliza o algoritmo de agrupamento baseado no processo *Dirichlet*. Esse algoritmo apresenta duas características que são muito úteis em problemas como a tarefa de resolução de correferência. São elas: a não necessidade de informar a *priori* o número de classes e de ser robusto a vetores esparsos.

Na Figura 5.3 é apresentado o resultado da execução de um algoritmo de *Dirichlet* encontrado a regularidade em dados distribuídos segundo uma distribuição normal. Os círculos mais claros representam as iterações do algoritmo, enquanto o círculos mais escuros são os grupos encontrados.

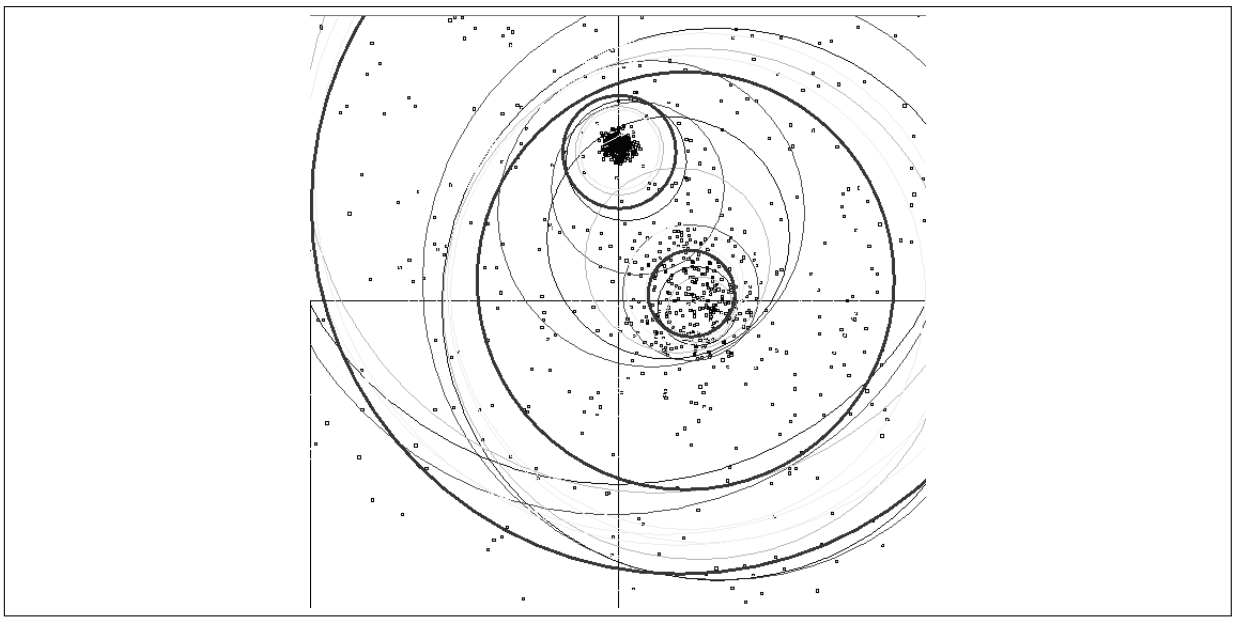

Figura 5.3: *Agrupador Dirichlet em distribuições normais. Extraído de Apache (2011)*

Para utilizar o *Dirichlet* é necessário definir como são distribuídos os dados pelo espaço, ou seja, determinar as medidas de distância entre os vetores.

Na subseção a seguir é apresentada a medida de distância utilizada no MemexLink.

5.4.1 Medida de distância

A medida de distância utilizada no MemexLink é a de Manhattan modificada para que possam ser atribuídos pesos às características. A medida distância é apresentada na equação 5.1.

$$d(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_i|x_j - y_j| \tag{5.1}$$

Onde x e y são os vetores de características dos SNs, n é o número de característica, m é o número de *bits* de cada característica e w_i é o peso atribuídos a cada característica. Na Tabela 5.5 é apresentado o conjunto de pesos utilizados para cada atributo.

Características	Pesos
Núcleo do SN	20
Número	1
Gênero	1
Classe semântica 1	5
Classe semântica 2	10
Classe semântica 3	20
Pseudônimo	10
Nome próprio	1
Definido	1

Tabela 5.5: *Pesos do conjunto de característica do MemexLink.*

Os pesos foram definidos empiricamente através da análise da execução do algoritmo de agrupamento para um corpus com anotações de correferência (veja, no Capítulo 6 a descrição do corpus) e da posterior verificação comparativamente do desempenho de cada atributo na tarefa. Os pesos então foram definidos proporcionalmente aos resultados obtidos para cada característica.

5.5 Aplicação de regras heurísticas

Após a execução do algoritmo de agrupamento, o sistema já tem o conjunto das menções agrupado. Então, por fim, são aplicadas regras para realizar a junção de cadeias ou desfazer os grupos de menções que não representa uma cadeia de correferência.

Essas regras foram criadas, após a análise dos resultados obtidos pelo algoritmo de agrupamento e baseadas nas próprias características do fenômeno de correferência. Por exemplo, a do aposto definido, que é baseado no fato que para maioria dos casos nos quais existe uma relação de aposto entre SNs, esses representam menções a uma mesma entidade. As regras utilizadas são três:

1. Aposto definido: essa regra define se uma menção tem um aposto definido. Se esse não pertence a cadeia de correferência, a cadeia do aposto e da menção devem ser combinadas;
2. Cadeia com sintagmas indeterminados: se uma cadeia só tem sintagmas nominais indeterminados esses sintagmas não podem ser correferentes. Então essa cadeia é desfeita.
3. Sintagmas com modificadores numerais: se um sintagma só tem modificadores numerais e esses têm valores diferentes, então eles pertencem a cadeias diferentes.

A primeira regra visa combinar cadeias que, pelos atributos, não seriam identificadas como correferentes. Observa-se o texto 5.5.

(5.5) João, o ferreiro, trabalhava na zona sul da cidade. O ferreiro não trabalhava dia de segunda.

No exemplo 5.5, o algoritmo de AM agruparia os dois SNs “o ferreiro” em uma só cadeia. No entanto, o SN “João” ficaria em uma cadeia diferente. Porém, utilizando essa regra, é possível colocar os SNs na mesma cadeia.

Já a segunda regra visa retirar do conjunto das entidades correferentes as cadeias de SNs que não estão se referindo a uma entidade específica, ou seja, apesar de haver um referente, ele não foi identificado nos textos.

A última regra é uma forma de corrigir cadeias de correferência que contêm sintagmas do tipo “2 gols” » “3 gols” » “4 gols”, que apesar dos SNs serem muitos parecidos, os numerais que os antecede determinam que eles são diferentes.

5.6 Ferramentas utilizadas no MemexLink

Para o desenvolvimento desse protótipo, foram utilizadas ferramentas de PLN e de aprendizado de máquina. São elas: o reconhecedor de entidades mencionadas Rembrandt (Cardoso, 2008), o etiquetador morfossintático para o português PALAVRAS (Bick, 2000), o tesauro TeP2.0 (Maziero et al., 2008), a ferramenta de AM com o algoritmo de agrupamento Mahout (Apache, 2011) e a ferramenta para visualização e anotação das cadeias de correferência MMAX (Mueller e Strube, 2001). Na Figura 5.4 são apresentadas as relações das ferramentas com as diversas etapas do MemexLink. Também são mostradas as entradas e saídas das ferramentas.

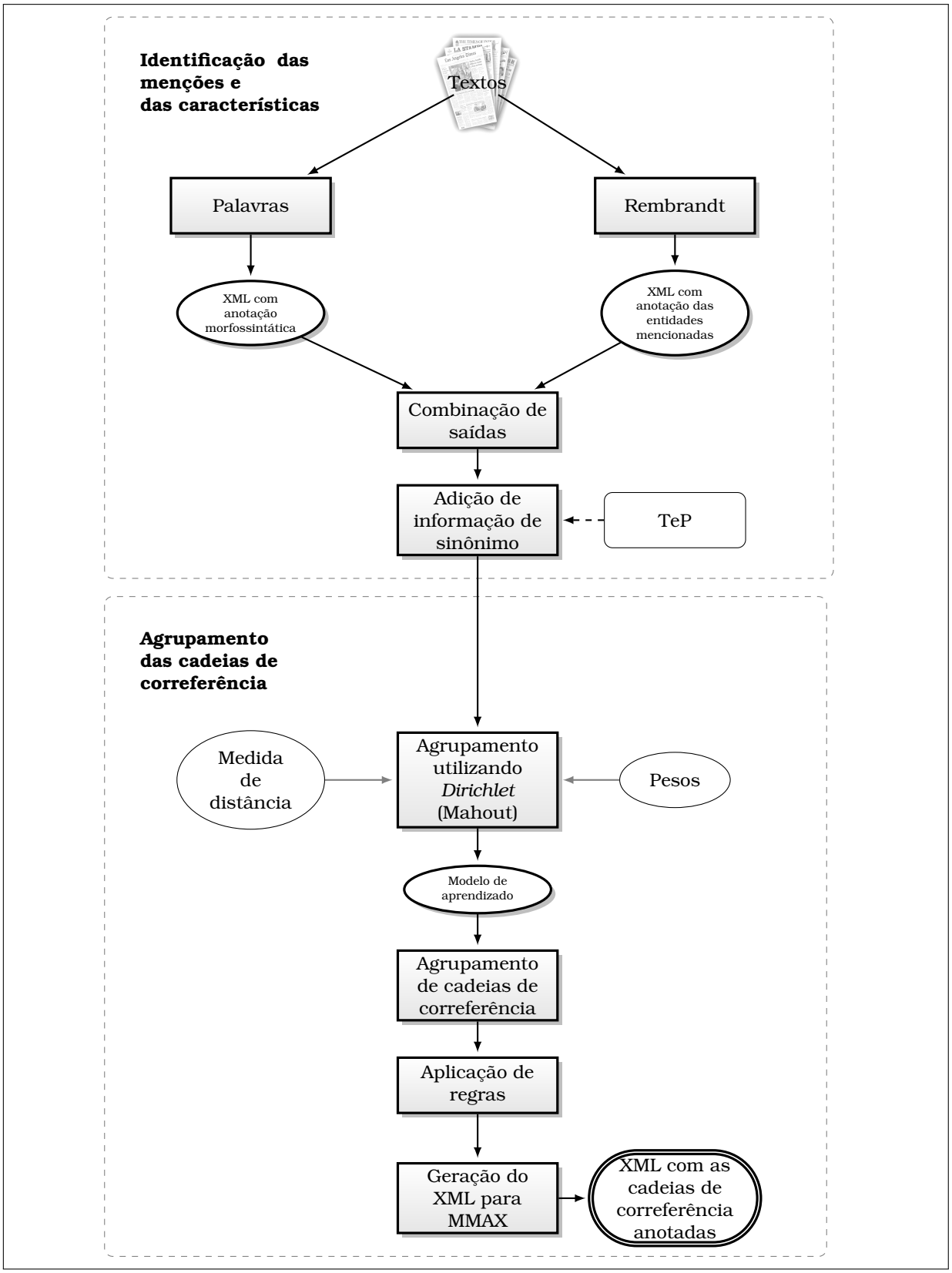

Figura 5.4: *Arquitetura do MemexLink detalhada apresentando as ferramentas de PLN utilizadas*

Como pode ser observado na Figura 5.4, o texto é submetido ao analisador sintático Palavras. Nessa fase são extraídos os SNs e feita a análise

morfossintática das palavras. Na segunda etapa de processamento, o texto é enviado para o Rembrandt para identificar as entidades mencionadas e realizar a marcação dos traços semânticos. O desempenho do Rembrandt, avaliado no Harem (Mota e Santos, 2008) quanto à medida de precisão, cobertura e medida-f, é, respectivamente, 0,63, 0,50 e 0,60. Observa-se que esses valores podem influenciar decisivamente no desempenho global do algoritmo, já que a maior parte das características são extraídas das saídas desses dois aplicativos. Outro problema é que não existe uma correspondência univoca entre as unidades anotadas pelo Palavras (SNs) e as anotadas pelo Rembrandt (entidades mencionadas). No MemexLink, como já foi tratado na seção 5.3, é feita uma tentativa de realizar a junção das características utilizando-se um método simples de retirar os artigos do SN na tentativa de encontrar a entidade mencionada anotada pelo Rembrandt. Para finalizar essa etapa, é feita a obtenção das características do MemexLink. Utiliza-se o TeP para obter os sinônimos dos nomes comuns. O TeP tem 8528 *synsets* de substantivos, a classe de palavras utilizada no MemexLink. Essa quantidade de *synsets* garante uma boa cobertura dos sinônimos.

Finalizada a etapa de extração dos atributos linguísticos, o sistema cria os vetores de características e os submete ao Mahout [4] (Apache, 2011) utilizando o agrupador baseado no processo de Dirichlet. Esse cria os modelos de aprendizado que são utilizados para construir as cadeias de correferência. Por fim, é feita a aplicação das regras e gerado o XML para ser visualizado no MMAX.

5.7 Considerações Finais

Neste capítulo, foi apresentada a arquitetura do MemexLink. Essa arquitetura é baseada nos trabalhos de sistemas não supervisionados para resolução de correferência. A diferença para essas arquiteturas é a utilização de regras com o objetivo de melhorar o sistema e suprir algumas lacunas deixadas pelo algoritmo de AM. Quanto às ferramentas utilizadas no MemexLink, salienta-se que utilizaram-se as consideradas melhores em suas tarefas, quando se trata do processamento automático da língua portuguesa. Apesar do protótipo desenvolvido ser para o português esse também pode ser aplicado para

[4]Mahout - é uma ferramenta que contém a implementação de um conjunto de algoritmos de AM. Essa ferramenta é disponível em licença livre, o que permitiu no caso desse trabalho a implementação de uma medida de distância própria.

outras línguas, bastando apenas a adequação do conjunto de ferramentas e os ajustes dos parâmetros do algoritmo.

No próximo capítulo é apresentada a avaliação do MemexLink quanto às métricas apresentadas no capítulo 3.

CAPÍTULO

6

Avaliação do MemexLink

Neste capítulo é apresentado o método de avaliação desenvolvido para o sistema proposto nesta dissertação. O MemexLink foi avaliado intrinsecamente quanto ao seu desempenho na resolução de correferência em mono e em múltiplos documentos. O resultado da avaliação do sistema é comparado com um corpus anotado manualmente.

Este capítulo está dividido como segue: a seção 6.1 descreve o método de anotação do corpus; na seção 6.2 são apresentados os métodos *baselines*; na seção 6.3 são mostrados os resultados da avaliação do MemexLink e na última seção deste capítulo é feita uma discussão acerca dos resultados obtidos.

6.1 Corpus de Avaliação

O método de avaliação do protótipo utilizado é baseado na comparação das cadeias de correferência obtidas pelo sistema com as cadeias anotadas manualmente. Até o desenvolvimento deste trabalho não havia corpus para o português anotado com informações de correferência em múltiplos documentos. Então, para que fosse possível realizar a avaliação do MemexLink, foi feita a anotação de um corpus. Foi utilizado um subconjunto com 3 grupos de textos do CST-News (Maziero et al., 2010). Esse corpus é composto por grupos

de textos que tratam de um mesmo fato jornalístico. Dessa forma, ele se torna adequado para a utilização como corpus de validação do método proposto.

Como ferramenta de anotação, foi utilizado o MMAX (Mueller e Strube, 2001). Apesar do MMAX não disponibilizar uma forma para anotação de textos em múltiplos documentos foi desenvolvido no escopo desse trabalho um módulo que provê essa característica a ferramenta. A Figura 6.1 mostra o módulo para anotação em múltiplos documentos.

Figura 6.1: *MMAX alterado para tratar com anotação de múltiplos documentos*

A anotação do corpus foi baseada no método adotado por Collovini et al. (2007) e Hasler et al. (2006). Seguem os passos realizados para obter o corpus anotado.

1. Definição de um guia de anotação para múltiplos documentos.
2. Anotação de teste de 3 grupos de textos com 2 pessoas.
3. Refinamento e simplificação do guia de anotação.
4. Anotação com 3 grupos de textos com 7 pessoas, duas para cada grupo e um juiz. Esse corpus foi utilizado para definição do conjunto de pesos.

5. Anotação com 6 grupos de textos com 3 pessoas utilizado nos testes do sistema.

A quantidade de pessoas envolvidas está de acordo, por exemplo, com a quantidade envolvida na anotação do corpus SummitCollovini et al. (2007) quanto às cadeias de correferência. Quanto as expertise dos anotadores são divididos entre 6 cientistas da computação e 1 linguista. Já quantidade de textos anotados é reduzida se comparada com a Hasler et al. (2006), pois nesse trabalho são anotados em média 10 textos por grupo, em um total de 5 grupos. A quantidade de textos foi apenas 9, porque a anotação dos textos exige um tempo considerável por parte do anotadores, o que no âmbito deste trabalho esse tempo era reduzido. O guia de anotação foi definido baseado em Collovini et al. (2007). O guia é bastante simplificado, tendo como foco apenas anotação das relações de correferências. A anotação inicia-se pela correção dos SNs, já anotados previamente pelo analisador morfossintático Palavras (Bick, 2000). Após corrigida a anotação dos SNs, é feita a anotação das ligações de correferência.

A anotação de correferência é dividida em duas etapas. Na primeira etapa é realizada a anotação das correferências em mono-documento e na segunda etapa é feita a anotação das correferência entre os documentos.

A primeira anotação foi realizada com o intuito de fazer os ajustes no guia de anotação. Para a avaliação da concordância entre os 2 anotadores dessa etapa, foi utilizado o método apresentado em Passonneau (1997), no qual a medida estatística *kappa* (Carletta, 1996) é calculada utilizando-se as ligações de correferência. Os resultados dessa anotação são apresentados na Tabela 6.1. As colunas da tabela referente aos anotadores mostram o que cada anotado, nesse caso apenas 2, anotou quanto a quantidade de SNs. Na coluna "Textos" os elementos entre parênteses representam a identificação do grupo de textos no corpus CST-News (Maziero et al., 2010).

Textos	Quantidade de textos	Quantidade de SNs		*Kappa*
		Anotador 1	Anotador 2	
Grupo 1 (C1)	3	152	147	0,29
Grupo 2 (C8)	3	171	183	0,11
Grupo 3 (C15)	3	167	154	0,04
Total	**9**	**490**	**484**	**0,11**

Tabela 6.1: *Kappa para os textos anotados com as cadeias de correferência do CST-New (Primeira Anotação)*

Na Tabela 6.1, os resultados apresentados revelaram que o guia de anotação e forma de anotação deveriam ser revistos. Segundo a análise da estatística *kappa* realizada por Landis e Koch (1977) apresentada na Tabela 6.2, o valor total de concordância é considerada pobre.

Intervalos	Interpretação
<0	Sem concordância
0-0.19	Concordância pobre
0.20-0.39	Concordância fraca
0.40-0.59	Concordância moderada
0.60-0.79	Concordância substancial
0.80-1.00	Concordância quase perfeita

Tabela 6.2: *Interpretação dos valores da estatística $kappa$. Extraído de Landis e Koch (1977)*

Analisados os resultados do corpus, foram verificados os problemas para essa baixa concordância. Um desses problemas era a baixa concordância na anotação dos SNs, principalmente com relação aos modificadores do núcleo. Então, como solução, foi inserido no guia de anotação um conjunto de exemplos esclarecendo como realizar a anotação quanto aos SNs e seus modificadores.

Um problema que foi reportado pelo anotadores foi a dificuldade de se realizar as ligações entre os documentos, pois exigia a leitura dos textos várias vezes. Na tentativa de minimizar esse problema, a estratégia adotada foi reduzir a quantidade de textos para anotação, por anotador. Cada anotador ficou apenas com um grupo de textos para anotar. Dessa forma, a quantidade de anotadores subiu para 6, reduzindo, então, o trabalho realizado por cada um deles, e assim, podendo obter-se resultados melhores. Na Tabela 6.4 são apresentados os resultados obtidos por essa anotação. Nessa tabela as colunas referentes aos anotadores devem ser interpretadas da seguinte maneira: o anotador 1 e o anotador 2 são diferentes para cada grupo de textos, por exemplo, o anotador 1 do grupo 1 é diferente do anotador 1 do grupo 2. Diferente do que ocorre na primeira anotação, onde os resultados são mostrados na Tabela 6.1, nessa anotação cada grupo de texto tem um conjunto de anotadores distintos.

A nova anotação resultou em valores como os apresentados na Tabela 6.2 com concordância quase perfeita. No entanto, para realizar a avaliação era necessário um corpus com anotação sem discordâncias. Para obter esse corpus, um juiz analisou os casos em que não houve concordância e decidiu por uma ou outra anotação. No fim do processo, foi obtido um corpus com as informações de correferência em múltiplos documentos e mono documento.

Textos	Quantidade de textos	Quantidade de SNs Anotador 1	Quantidade de SNs Anotador 2	*Kappa*
Grupo 1 (C1)	3	140	140	0,95
Grupo 2 (C8)	3	191	183	0,68
Grupo 3 (C15)	3	144	145	0,88
Total	**9**	**475**	**468**	**0,82**

Tabela 6.3: *Kappa para os textos anotados com as cadeias de correferência do CST-New*

Para realizar a avaliação do protótipo desenvolvido nessa dissertação foi realizar a construção um corpus com 18 textos. Os resultados da avaliação são apresentados na tabela abaixo.

Textos	Quantidade de textos	Quantidade de SNs Anotador 1	Quantidade de SNs Anotador 2	*Kappa*
Grupo 1	3	74	65	0,85
Grupo 2	3	72	61	0,83
Grupo 3	3	123	112	0,68
Grupo 4	3	56	51	0,67
Grupo 5	3	90	86	0,61
Grupo 6	3	109	104	0,63
Total	**18**	**524**	**479**	**0,62**

Tabela 6.4: *Kappa para os textos anotados com as cadeias de correferência do CST-New*

Esses resultados mostram que a anotação de correferência em múltiplos documentos para esse conjunto de texto tem uma boa concordância. Também é possível verificar que a tarefa é passível de automatização, pois essa tarefa pode ser replicada por diferentes humanos e produz resultados semelhantes.

6.2 Sistemas baseline

Os *baselines* apresentados nesta seção são baseados nos descritos em Cardie et al. (1999). Foram definidos dois sistemas *baselines* para realizar a comparação dos resultados como o MemexLink. São eles:

a) *Baseline* 1: todos os sintagmas nominais são considerados da mesma cadeia de correferência;

b) *Baseline* 2: são considerados correferentes os sintagmas que têm o mesmo núcleo.

Segundo Cardie et al. (1999), o *baseline* 1 é definido para verificar qual a cobertura máxima que pode ser obtida para o corpus. Essa cobertura máxima é limitada pelo desempenho na extração dos SNs por parte do analisador sintático. Nas tabelas 6.5 e 6.6 são apresentados os resultados obtidos na extração dos SNs para o corpus anotado descrito anteriormente.

Textos	Quantidade de SN no CR[a]	Quantidade de SN extraídos	SNs corretos	SNs parciais	SNs não identificados
Grupo 1 (C1)	140	124	101	16	23
Grupo 2 (C8)	188	160	128	8	52
Grupo 3 (C15)	144	149	94	5	45
Total	**472**	**433**	**323**	**29**	**120**

Tabela 6.5: *Detalhes da identificação dos SNs no corpus anotado pelo Palavras (Bick, 2000)*

[a]Corpus de Referência

Textos	Cobertura	Precisão	Medida-f
Grupo 1 (C1)	67,01%	64,76%	65,87%
Grupo 2 (C8)	77,85%	87,9%	82,57%
Grupo 3 (C15)	70,21%	82,5%	75.86%
Total	**71,5%**	**77,94%**	**74,58%**

Tabela 6.6: *Resultados da identificação dos SNs no corpus anotado pelo Palavras (Bick, 2000) quanto as medidas de precisão e de cobertura*

É observado na Tabela 6.5 que a quantidade de SN não identificados é cerca de 25,4% do total de SNs. Esse valor define um teto no qual o algoritmo de resolução de correferência pode obter quanto a cobertura.

Na Tabela 6.7, são apresentados os resultados para a avaliação das cadeias em múltiplos documentos obtidos pelo *baseline* 1. Os resultados obtidos são avaliados quanto às medidas de cobertura, precisão e medida-f, utilizando-se a medida empregada no MUC (Vilain et al., 1995) e a B-CUBEB (Bagga e Baldwin, 1998b). Observa-se que apesar de definir todos os SNs pertencendo a apenas uma cadeia de correferência, o valor de cobertura, que deveria ser 100%, é de 71,02% e 68,09%, para o MUC e B-CUBEB, respectivamente. O valor de precisão para o MUC foi de 40,46%. Esse valor foi obtido porque a medida MUC privilegia a formação de cadeias de correferência, ao contrário da medida B-CUBEB que privilegia a identificação de cadeias com um elemento (*singletons*), obtendo o valor de precisão de 2,38%.

Na Tabela 6.8, são apresentados os resultados obtidos para as cadeias em múltiplos documentos pelo *baseline* 2. Observa-se que os valores de precisão obtidos por essa abordagem são: MUC 70,17% e B-CUBEB 79,17%.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Grupo 1 (C1)	83,54%	53,65%	65,34%	79,18%	3,21%	6,18%
Grupo 2 (C8)	69,76%	37,73%	48,97%	66,98%	2,59%	4,99%
Grupo 3 (C15)	60%	32,43%	42,1%	58,75%	1,46%	2,85%
Total	**71,02%**	**40,46%**	**51,55%**	**68,09%**	**2,38%**	**4,6%**

Tabela 6.7: *Resultados da avaliação em múltiplos documentos baseline 1 quanto as medidas de MUC e B-CUBEB*

Esses valores sugerem que essa característica é importante para a resolução de correferência, o que também pode ser verificado no resultado geral quanto a medida-f 26,49% e 52,52%, para MUC e B-CUBEB, respectivamente.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Grupo 1 (C1)	6,25%	33,33%	10,52%	33,75%	63,98%	44,19%
Grupo 2 (C8)	39,24%	91,17%	54,86%	46,51%	92,2%	61,83%
Grupo 3 (C15)	4,65%	50,00%	8,51%	38,17%	83,22%	52,34%
Total	**16,32%**	**70,17%**	**26,49%**	**39,29%**	**79,17%**	**52,52%**

Tabela 6.8: *Resultados da avaliação do baseline-2 quanto as medidas de MUC e B-CUBEB*

Na Tabela 6.9, são apresentados os resultados dos dois algoritmos para mono documento. Um ponto que deve ser observado, é o resultado muito inferior para o *baseline* 2 quanto a medida MUC, 1,56% de medida-f. Isso se deve às características do corpus utilizado, que é composto de textos jornalísticos pequenos, que na maioria das vezes apenas introduzem um conjunto de entidades nas frases iniciais e nas próximas frases é feito o uso da anáfora indireta para retomar as entidades mencionadas. Já os valores altos quanto à medida B-CUBEB se deve ao fato de que a maioria das menções forma cadeias unitárias.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Baseline 1	71,33%	3,17%	6,08%	72,13%	20,75%	32,23%
Baseline 2	0,81%	16,66%	1,56%	54,2%	80,71%	64,85%

Tabela 6.9: *Resultados da avaliação do baselines em mono-documento quanto as medidas de MUC e B-CUBEB*

6.3 Resultados obtidos pelo MemexLink no corpus de testes

Foram realizados quatro experimentos para avaliação do MemexLink utilizando o corpus de teste, são eles: a) sistema sem regras e sem informações semânticas do Rembrandt (Cardoso, 2008), b) sistema com regras, mas sem utilizar informação semântica, c) sistema sem as regras e utilizando as informações semânticas do Rembrandt e d) o sistema completo, com regras e as informações semânticas.

Os experimentos foram realizados dessa forma por dois motivos: (1) verificar qual a importância das características semânticas na resolução de correferência, pois a extração dessas características tem um elevado custo quanto ao tempo; e (2) verificar qual o efeito das regras para o sistema.

Os experimentos foram avaliados utilizando as medidas do MUC e a B-CUBEB, tanto considerando apenas cadeias de correferência em mono documento como as cadeias inter e intra documentos. Os resultados obtidos pelo MemexLink são apresentados a seguir.

a) Experimento 1 – MemexLink sem regras e sem informação semântica do Rembrandt.

   São apresentados nas Tabelas 6.10 e 6.11 os resultados obtidos pelo sistema para essa configuração.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Grupo 1 (C1)	72,15%	82,6%	77,02%	68,63%	82,55%	73,13%
Grupo 2 (C8)	50%	54,43%	52,12%	52,05%	64,61%	57,66%
Grupo 3 (C15)	46,25%	45,67%	45,96%	51,1%	50,88%	50,99%
Total	**55,91%**	**59,82%**	**57,8%**	**55,79%**	**65,02%**	**60,05%**

Tabela 6.10: *Resultados da avaliação do MemexLink sem regras e sem informação do Rembrandt quanto às cadeias em múltiplos documentos*

MUC			B-CUBEB		
Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
42,62%	54,73%	47,92%	62,25%	74,36%	67,77%

Tabela 6.11: *Resultados da avaliação do MemexLink sem regras e sem informação do Rembrandt quanto às cadeias em mono documento*

Os resultados apresentados nas tabelas 6.10 e 6.11 superam percentualmente todos os obtidos pelos *baselines* 1 e 2 quanto a medida-f para as duas medidas, MUC e B-CUBEB. Esses resultados mostram que com a combinação de atributos, mesmo que sejam atributos superficiais, obtém-se um desempenho superior se comparado com as estratégias simples como as adotadas nos *baselines*. Esses resultados também mostram que uma estratégia não supervisionada é viável e obtém resultados melhores que as estratégias simples.

b) Experimento 2 – Memex com regras e sem informação semântica do Rembrandt

Nas Tabelas 6.12 e 6.13 são apresentados os resultado obtidos pelas as avaliação do MemexLink utilizando a configuração com regras e sem informação do Rembrandt.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Grupo 1 (C1)	73,41%	80,55%	76,82%	69,68%	74,79%	72,15%
Grupo 2 (C8)	50%	60,56%	54,77%	52,01%	68,96%	59,33%
Grupo 3 (C15)	43,75%	43,75%	43,75%	49,71%	49,78%	49,74%
Total	**55,51%**	**60,98%**	**58,11%**	**56,57%**	**64,03%**	**60,07%**

Tabela 6.12: *Resultados da avaliação do MemexLink utilizando regras e sem informação semântica do Rembrandt quanto às cadeias em múltiplos documentos*

MUC			B-CUBEB		
Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
47,54%	47,15%	47,34%	63,95%	69,94%	66,81%

Tabela 6.13: *Resultados da avaliação do MemexLink utilizando regras e sem informação semântica quanto às cadeias em mono documento*

Nas Tabelas 6.12 e 6.13 os resultados apresentados mostram que em comparação com o experimento 1, os resultados praticamente não tiveram alteração. No entanto, observa-se que existe uma tendência de aumentar a cobertura quando é feita a adição de regras, principalmente para a medida B-CUBEB. Isso se deve ao fato da natureza das regras, que em sua maioria tentam desfazer cadeias de correferência com problemas, que é o caso das regras para SNs indefinidos e SNs com modificador numeral. Essas regras tendem a criar mais grupos com apenas uma menção a entidade. A medida B-CUBEB é capaz de pontuar a detecção dos grupos com uma só menção (Bagga e Baldwin, 1998b).

c) Experimento 3 – MemexLink sem regras e utilizando a informação semântica extraída do Rembrandt

Nas Tabelas 6.14 e 6.15 são apresentados os resultado obtidos pelas avaliação do MemexLink utilizando a configuração sem regras e com informação do Rembrandt.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Grupo 1 (C1)	72,15%	77,02%	74,5%	68,63%	76,01%	70,44%
Grupo 2 (C8)	50%	54,43%	52,12%	52,05%	64,61%	57,66%
Grupo 3 (C15)	46,25%	45,67%	45,96%	51,1%	50,88%	50,99%
Total	**55,91%**	**58,54%**	**57,2%**	**55,79%**	**63,15%**	**59,24%**

Tabela 6.14: *Resultados da avaliação do MemexLink sem regras e com informação semântica do Rembrandt quanto às cadeias em múltiplos documentos*

MUC			B-CUBEB		
Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
42,62%	50,48%	46,22%	62,25%	72,59%	67,03%

Tabela 6.15: *Resultados da avaliação do MemexLink com regras e sem informação semântica do Rembrandt quanto às cadeias em mono documento*

Observa-se que adição de informação semântica do Rembrandt praticamente não alterou os resultados do sistema, se comparado com os experimentos 1 e 2. Esse resultado deve-se à dificuldade de se realizar a junção das informações do Rembrandt, que anota entidade nomeada, com as informações do Palavras, que anota SNs. Outro fato que influenciou esse resultado foi a baixa cobertura do Rembrandt para entidades do tipo Organização, cerca de 32% (Cardoso, 2008). O tipo Organização é bastante presente no tipo de texto jornalistico e a sua detecção adequada poderia aumentar o desempenho do MemexLink.

d) MemexLink utilizando regras e informação semântica.

Nas Tabelas 6.16 e 6.17 são apresentados os resultados obtidos pela avaliação do MemexLink utilizando a configuração com regras e informação do Rembrandt.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Grupo 1 (C1)	68,42%	65%	66,6%	76,33%	81,85%	78,99%
Grupo 2 (C8)	30,43%	38,88%	34,14%	56,29%	74,99%	64,31%
Grupo 3 (C15)	42,1%	42,1%	42,1%	59,83%	58,53%	59,17%
Total	**55,51%**	**58,11%**	**56,78%**	**56,35%**	**62,27%**	**59,17%**

Tabela 6.16: *Resultados da avaliação do MemexLink com regras e informação semântica do Rembrandt quanto às cadeias em múltiplos documentos*

MUC			B-CUBEB		
Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
45,9%	49,12%	47,45%	63,31%	71,29%	67,07%

Tabela 6.17: *Resultados da avaliação do MemexLink com regras e informação semântica do Rembrandt quanto às cadeias em mono documento*

Observa-se que nas Tabelas 6.16 e 6.17 que a utilização do sistema com regras e as informações semânticas do Rembrandt não obteve uma melhoria esperada quanto as medidas MUC e B-CUBEB. Os principais motivos para isso são os erros nos analisadores sintático e semântico. Quanto ao analisador sintático, um erro que influência na aplicação de regras é a não identificação de um aposto ou, o que ainda é mais prejudicial, a identificação errada do aposto. Já quanto ao analisador semântico, sua baixa cobertura e a dificuldade de relacionar as entidade nomeadas com os SNs impossibilita a identificação correta de várias ligações de correferências.

6.4 Resultados obtidos pelo MemexLink para o corpus de testes

Foram realizados experimentos para avaliação do MemexLink como um corpus de testes com 18 textos os resultados são apresentados na tabela 6.18.

Textos	MUC			B-CUBEB		
	Cobertura	Precisão	Medida-f	Cobertura	Precisão	Medida-f
Corpus de teste	**77,77%**	**40,32%**	**53,10%**	**53,94%**	**65,15%**	**59,01%**

Tabela 6.18: *Resultados da avaliação do MemexLink com regras e informação semântica do Rembrandt quanto às cadeias em múltiplos documentos*

6.5 Discussão dos resultados obtidos

Os resultados apresentados pelo conjunto de experimentos descritos neste capítulo mostram que um algoritmo não supervisionado para resolução de correferência em múltiplos documentos obtém resultados melhores que métodos simples como os proposto nos *baselines*. No entanto, observa-se que o método proposto é fortemente dependente do conjunto de ferramentas que são utilizadas, como os analisadores sintáticos e semânticos. Os resultado obtidos dependem diretamente do desempenho dessas ferramentas.

O MemexLink foi superior quanto à medida-f utilizando a medida MUC em 30,08% e quanto à medida B-CUBEB em 7,53%, quando avaliado em múltiplos documentos em comparação ao *baseline* 2 (elementos são correferentes se concordam em núcleo), que um *baseline* que obtém bons resultados na tarefa de resolução de correferência.

Na tarefa de resolução de correferência em mono documento os resultados foram 46,36% e 2,92% superiores, para MUC e B-CUBEB quanto à medida-f, respectivamente.

Os resultados quanto à medida MUC mostram que o sistema conseguiu identificar um número de ligações de correferência muito maior que a do *baseline*. Já as melhorias quanto à medida B-CUBEB mostram que o algoritmo também consegue identificar as menções que não são correferentes.

O resultado do *baseline* 1 (todos os elementos pertencentes a uma mesma cadeia de correferência) mostra o limite que um algoritmo de resolução automática pode obter quanto à medida de cobertura utilizando para avaliação corpus anotado com o analisador sintático (Palavras). O MemexLink foi apenas 15,42% para MUC e 11,3% para B-CUBEB menor que o resultado máximo para múltiplos documentos, mostrando que o método identifica a maioria das ligações de correferências em múltiplos documentos.

Para mono-documento os resultados são 23,79% e 8,18%, para MUC e B-CUBEB respectivamente. Esses resultados mostram que ainda existe um grande espaço para melhoria do sistema para que ele possa se tornar mais efetivo na identificação das cadeias de correferência em mono documento.

Apesar dos resultados do MemexLink serem superiores aos obtidos pelo *baseline*, vale ressaltar que a avaliação do sistema deve ser realizada com um corpus maior, pois a quantidade de textos utilizados na avaliação, se comparada com outros trabalhos, é reduzida. Em outros trabalhos para a área de resolução de correferência o corpus utilizado é muitas vezes maior, como o corpus utilizado no trabalho de Baron e Freedman (2008) com 400 documentos.

Outro ponto importante que deve ser abordado é que com a adição de conhecimento semântico e simbólico, o sistema não apresentou uma melhora significativa; pelo contrário, ocorreu uma pequena piora nos resultados obtidos pelos experimentos. Existem duas causas possíveis para a ocorrência desses resultados: (1) problemas na anotação feita pelos analisadores sintáticos e semânticos, e (2) os atributos e/ou as regras não estão bem definidos

para o problema de resolução de correferência. Uma forma de verificar qual a causa real do problema seria utilizar um corpus com as anotações sintáticas e semânticas corrigidas manualmente. No entanto, um corpus com esse tipo de anotação é difícil de obter ou criar, o que dificulta esse tipo de avaliação.

Porém, o sistema MemexLink contém algumas características que fazem dele uma opção como um sistema de resolução de correferência em múltiplos documentos: (1) tenta estabelecer ligações de correferência entre todos os SNs contidos nos textos, ou seja, não se restringindo a tratar SNs de apenas alguns tipos de entidades mencionadas; (2) não haver necessidade de informar o número de cadeias de correferência, o algoritmo não supervisionado induz automaticamente o número; e (3) possibilitar a de adição de regras semânticas, que apesar de, na avaliação do sistema não ter obtido bons resultados com as regras utilizadas, isso não implica que a definição de outras regras não poderia melhorar os resultados do sistema.

CAPÍTULO

7

Considerações Finais

Nesta dissertação apresentou-se um método de resolução de correferência para múltiplos documentos não supervisionado. As características do método investigado são: a quantidade reduzida de parâmetros do algoritmo, a possibilidade de inclusão de conhecimento simbólico e a utilização do SN com núcleo nominal como unidade para as cadeias de correferência, ou seja, o métodos proposto tenta resolver correferência entre todos os SNs encontrados nos textos sem aplicar filtro quanto ao tipo de menção.

Dentre os parâmetros, o que é recorrente nos métodos não supervisionados para resolução de correferência é a necessidade de determinação de uma limiar ou do próprio número de cadeias de correferência. O método proposto supera essa dificuldade utilizando o método Dirichlet de agrupamento, que é baseado em distribuições estatísticas.

Quanto ao conhecimento simbólico, o método pode utilizar um conjunto de regras com a finalidade de tratar casos que o algoritmo de aprendizado não trata. Um diferencial desse trabalho é a utilização do SN com núcleo nominal como unidade das cadeias de correferência para múltiplos documentos, pois nos trabalhos anteriores é feito um filtro das entidades a serem tratadas; por exemplo, tratar apenas menções a pessoas e/ou organizações. Neste trabalho, no entanto, não é feito filtro de quais SNs são tratados.

Para validação do método proposto foi construído um protótipo que foi denominado de MemexLink. O MemexLink foi desenvolvido para resolver

correferência em múltiplos documentos para o português. O sistema utilizou o conjunto de ferramentas disponíveis para o português, que se comparadas com ferramentas para outras línguas, como o inglês, ainda podem melhorar seus resultados. Até o momento da escrita desta dissertação, esse é o primeiro sistema com esse propósito desenvolvido para o português.

Os resultados da avaliação do MemexLink mostraram-se promissores, sendo superior em 30,08% utilizando a medida MUC e em 7,53% quanto a medida B-CUBEB, quando avaliado em múltiplos documentos, em comparação ao *baseline* cujos elementos são correferentes se concordam em núcleo. Obtendo valores brutos 58,11% de MUC e 60,07% de B-CUBEB de medida-f. Entretanto, os resultados obtidos da utilização das regras (conhecimento simbólico) e das informações semânticas das entidade nomeadas não representaram um acréscimo no desempenho do sistema. Esse problema deve-se a adição de erros pelas ferramentas de análise sintática e semântica. Porém, uma avaliação mais detalhada deve ser feita para identificar quais os problemas gerados por essas ferramentas que ocasionam erros no MemexLink.

Apesar de não ser possível uma comparação direta dos resultados obtidos com outros da literatura, devido a diferença de corpus, de língua e do tipo de entidade tratada, um trabalho parecido com o Baron e Freedman (2008), obteve valor de 71,5% de medida-f utilizando a medida B-CUBEB, apenas para entidades do tipo pessoas e organizações.

Nas próximas seções são apresentadas as principais contribuições, algumas limitações apresentadas por este trabalho e diversos trabalhos futuros que podem ser desenvolvidos.

7.1 Contribuições

Apresenta-se nessa seção as principais contribuições desta dissertação. São elas:

a) Método de resolução de correferência não supervisionado, capaz de identificar a quantidade de cadeias de correferência automaticamente, sem a necessidade da definição de um limiar ou do próprio número de cadeias, e que pode combinar conhecimento estatístico com simbólico;

b) Investigação, pela primeira vez para um português, de métodos para a resolução de correferência em múltiplos documentos;

c) Investigação e escolha dos atributos utilizados para resolução de correferência não supervisionada para o português;

d) Sistema de resolução de correferência não supervisionado para o português para mono e múltiplos documentos. Esse sistema além de anotar as cadeias de correferência, produz como saída um conjunto de XMLs que podem ser utilizados pela ferramenta de anotação MMAX (Mueller e Strube, 2001). Dessa forma, o sistema pode ser utilizado também com uma ferramenta de auxilio a anotação de corpus e/ou de estudo do fenômeno de correferência;

e) Criação de um módulo adicional para o MMAX, para que ele seja capaz de tratar com múltiplos documentos;

f) Investigação e validação de métodos para anotação de corpus para correferência em múltiplos documentos para o português;

g) Criação de um guia para a anotação de corpus em múltiplos documentos para o português;

h) Corpus com anotação das correferência tanto em mono como em múltiplos documentos; sendo esse último inédito para o português.

7.2 Limitações

As limitações identificadas pelo método apresentado nesta dissertação são listadas a seguir.

a) Na avaliação foi utilizado um corpus com um total com 9 textos, o que, apesar de ter sido útil para verificar o desempenho do sistema, não é um corpus significativo. No entanto, optou-se por esse tamanho de corpus, devido ao tempo necessário para sua anotação.

b) Não foi realizado pelo sistema o agrupamento dos textos para definir se esses tratam do mesmo assunto. A entrada do sistema é composta de textos agrupados.

c) A definição dos pesos utilizados pelo algoritmo de agrupamento pode ser melhor sistematizada. Utilizando, por exemplo, algoritmos de otimização para encontrar a melhor combinação dos valores dos pesos.

d) Os modelos de aprendizado gerados são difíceis de extrair conhecimento sobre o fenômeno da correferência, pois cada grupo apenas representa um conjunto de menções. Diferente de um modelo supervisionado que, por exemplo, utiliza uma árvore de decisão, por esse modelo é possível identificar quais são os atributos mais revelantes para o fenômeno da correferência.

7.3 Trabalhos Futuros

Algumas possíveis formas de extensão deste trabalho são apresentadas a seguir:

a) Realizar a avaliação do sistema desenvolvido em um corpus maior, para ser possível verificar se o método se comporta de maneira parecida em um corpus mais significativo, como o corpus utilizado no trabalho de Baron e Freedman (2008) com 400 documentos;

b) Realizar a avaliação detalhada do impacto das ferramentas utilizadas no sistema, a fim de verificar se é necessário o ajuste do método proposto e identificar os erros das ferramentas que mais prejudicam o sistema;

c) Realizar uma avaliação extrínseca, por exemplo, com uma aplicação de perguntas e respostas;

d) Implementar um método para agrupar os texto quanto ao mesmo assunto, tornando, assim, o sistema totalmente automático.

e) Ampliar o conjunto de regras simbólicas, com o intuito de verificar se o desempenho do sistema pode ser melhorado.

APÊNDICE

A

Tipos semânticos do Harem

Na Tabela A.1 são apresentados os tipos semântico utilizado no Harem (Mota e Santos, 2008):

Tabela A.1: *Tipos semântico do Harem (Mota e Santos, 2008)*

Categorias	**Tipos**	**Subtipos**
ABSTRACCAO (5)	DISCIPLINA	
	ESTADO	
	IDEIA	
	NOME	
	OUTRO	
ACONTECIMENTO (4)	EFEMERIDE	
	EVENTO	
	ORGANIZADO	
	OUTRO	

Continua...

Categorias	**Tipos**	**Subtipos**
COISA (5)	CLASSE	
	MEMBROCLASSE	
	OBJECTO	
	SUBSTANCIA	
	OUTRO	
LOCAL (4)	FISICO (7)	ILHA, AGUACURSO, PLANETA, REGIAO, RELEVO, AGUAMASSA, OUTRO
	HUMANO (6)	RUA, PAIS, DIVISAO, REGIAO, CONSTRUCAO, OUTRO
	VIRTUAL (4)	COMSOCIAL, SITIO, OBRA, OUTRO
	OUTRO	
OBRA (4)	ARTE	
	PLANO	
	REPRODUZIDA	
	OUTRO	
ORGANIZACAO (4)	ADMINISTRACAO	
	EMPRESA	
	INSTITUICAO	
	OUTRO	
PESSOA (8)	CARGO	
	GRUPOCARGO	
	GRUPOIND	
	GRUPOMEMBRO	
	INDIVIDUAL	
	MEMBRO	
	POVO	
	OUTRO	

Continua...

Categorias	**Tipos**	**Subtipos**
TEMPO (5)	DURACAO	
	FREQUENCIA	
	GENERICO	
	TEMPO_CALEND (4)	HORA, INTERVALO, DATA, OUTRO
	OUTRO	
VALOR (4)	CLASSIFICACAO	
	MOEDA	
	QUANTIDADE	
	OUTRO	
OUTRO (1)		

Referências Bibliográficas

Apache. Mahout, 2011.

Amit Bagga e Breck Baldwin. How Much Processing Is Required for Cross-Document Coreference? In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98)*, number 919, pages 106–111. Association for Computational Linguistics, 1998a.

Amit Bagga e Breck Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85, Morristown, NJ, USA, 1998b. Association for Computational Linguistics.

Alex Baron e Marjorie Freedman. Who is who and what is what: experiments in cross-document co-reference. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 274–283, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000. ISBN 8772889101.

Thiago Ianez Carbonel. *Estudo e validação de teorias do domínio linguístico com vistas à melhoria do tratamento de cadeias de co-referência em Sumarização Automática*. Dissertação (mestrado), Universidade Federal de São Carlos, 2007.

Claire Cardie e David Pierce. Error-driven pruning of Treebank grammars for base noun phrase identification. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics* -, page 218, Morristown, NJ, USA, 1998. Association for Computational Linguistics. doi: 10.3115/980845.980881.

Claire Cardie e Kiri Wagstaf. Noun Phrase Coreference as Clustering. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, number 1995, pages 82–89. Assocciation for Computational Linguistics, 1999.

Claire Cardie, Kiri Wagstaff, e Others. Noun phrase coreference as clustering. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, 1999. doi: 10.1.1.29.4600.

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

Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, June 1996. ISSN 0891-2017.

José Castaño, Jason Zhang, e James Pustejovsky. Anaphora Resolution in Biomedical Literature. In *In Proceedings of the 2002 International Symposium on Reference Resolution*, 2002.

S. Collovini, T.I. Carbonel, J.T. Fuchs, J.C. Coelho, L. Rino, e R. Vieira. Summit: Um corpus anotado com informações discursivas visando à sumarização automática. In *5º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007)*, Rio de Janeiro, RJ, 2007. Proceedings of the SBC.

C Fellbaum. *WordNet: An Electronical Lexical Database*. MIT Press, Cambridge, MA, USA, 1998.

John F Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, e A. Schlichting Toncheva. The Diverse and Exploding Digital Universe. Technical report, An Information Data Center (IDC), Framingham, MA 01701 USA, 2008.

Ralph Grishman. Whither written language evaluation? In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 120–125, Plainsboro, NJ, 1994. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075812.1075836.

Aria Haghighi e Dan Klein. Unsupervised coreference resolution in a nonparametric bayesian model. *ANNUAL MEETING-ASSOCIATION FOR*, 2007.

M. A. K. Halliday e Rugaia Hasan. *Cohesion in English*. Longman Pub Group, 1976. ISBN 978-0582550414.

Laura Hasler, Constantin Orasan, e Karin Naumann. NPs for events: Experiments in coreference annotation. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006)*, pages 1167–1172. Citeseer, 2006.

J.R. Hobbs. *Pronoun resolution*, page 61. Association for Computing Machinery, 28 edition, 1977.

Nancy Ide e Dan Cristea. A hierarchical account of referential accessibility. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*, pages 416–424, Morristown, NJ, USA, October 2000. Association for Computational Linguistics. doi: 10.3115/1075218.1075271.

Ingedore Grunfeld Villaça Koch. *A coesão textual*. Contexto, São Paulo, 10 edition, 1998. ISBN 85-85134-46-1.

J Richard Landis e Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X. doi: 10.1007/BF00163035.

Vladimir I Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1(1):8–17, 1965.

W.C. Mann e S.A Thompson. Rhetorical Structure Theory: A Theory of Text Organization. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1987.

E. G. Maziero, Thiago Alexandre Salgueiro Pardo, A Di Felipo, e B. C Dias-da Silva. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *Workshop em Tecnologia da Informação e da Linguagem Humana - TIL 2008*, Vilha Velha - ES, 2008. Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana TIL 2008.

E.G. Maziero, MLC Jorge, e T.A.S. Pardo. Identifying Multidocument Relations. In *the Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science. June*, page 10, Funchal/Madeira, Portugal, 2010.

George a. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, e Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database *. *International Journal of Lexicography*, 3(4):235–244, 1990. ISSN 0950-3846. doi: 10.1093/ijl/3.4.235.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, e Ramez Zakhary. ACE 2004 Multilingual Training Corpus, 2004.

Ruslan Mitkov. *Anaphora Resolution*, volume 11. Longman, London, 1 edition, 2002. doi: 10.1017/S1351324905214006.

Cristina Mota e Diana Santos. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

MUC-6. Coreference task definition (v2.3, 8 Sep 95). In *In Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 335–344, 1995.

MUC-7. Coreference task definition (v3.0, 13 Jul 97). In *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1997.

Christoph Mueller e Michael Strube. MMAX: A Tool for the Annotation of Multimodal Corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50, 2001. doi: 10.1.1.18.3322.

Vincent Ng. Supervised Noun Phrase Coreference Research: The First Fifteen Years. *aclweb.org*, (July):1396–1411, 2010.

Vincent Ng e Claire Cardie. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111. ACL, 2002. doi: 10.1.1.20.896.

Rebecca J. Passonneau. Applying Reliability Metrics to Co-Reference Annotation. *CoRR*, (3):10, June 1997.

Xuan-hieu Phan, Le-minh Nguyen, e Susumu Horiguchi. Personal Name Resolution Crossover Documents by a Semantics-Based Approach. *IEICE TRANSACTIONS on Information and Systems*, (2):825–836, 2006. doi: 10.1093/ietisy/e89.

S.P. Ponzetto e Michael Strube. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL*, volume 6, pages 192–199, Morristown, NJ, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220835.1220860.

Hoifung Poon e Pedro Domingos. Joint unsupervised coreference resolution with Markov Logic. *Proceedings of the Conference on Empirical*, (October): 650, 2008. doi: 10.3115/1613715.1613796.

Lance Ramshaw, Elizabeth Boschee, Sergey Bratus, Scott Miller, Rebecca Stone, Ralph Weischedel, e Alex Zamanian. *Experiments in multi-modal automatic content extraction*. HLT '01. Association for Computational Linguistics, Morristown, NJ, USA, 2001. doi: 10.3115/1072133.1072176.

Lucia Helena Machado Rino e Eloize Rossi Marques Seno. A importância do tratamento co-referencial para a sumarização automática de textos. *Estudos Lingüísticos*, XXXV:1179–1188, 2006.

Horacio Saggion. Experiments on Semantic-based Clustering for Cross-document Coreference. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, volume I, pages 149–156, Sheffield, England, UK, 2007. ACL.

Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of In- formation by Computer*. Addison-Wesley, Reading, MA, 1989.

Denis Neves de Arruda Santos. *Resolução de anáfora pronominal em português utilizando o algoritmo de Hobbs*. Dissertação de mestrado, UNICAMP, 2008.

Wee Meng Soon, Daniel Chung, Daniel Chung Yong Lim, e Hwee Tou Ng. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, 2001. doi: 10.1.1.18.8040.

José Guilherme Souza, Patricia Nunes Gonçalves, e Renata Vieira. Learning Coreference Resolution for Portuguese Texts. In *Proceedings of*

the 8th international conference on Computational Processing of the Portuguese Language(Lecture Notes In Artificial Intelligence; Vol. 5190), pages 153–162, Berlin, Heidelberg, 2008. Springer-Verlag. doi: 10.1007/978-3-540-85980-2_16.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, e David Hysom. Coreference resolution with reconcile. In *Proceeding ACLShort 10 Proceedings of the ACL 2010 Conference Short Papers*, pages 156–161, July 2010.

M. Strube. NLP approaches to reference resolution. *Tutorial notes, ACL*, 2:124, 2002.

Pang-Ning Tan, Michael Steinbach, e Vipin Kumar. *Introduction to Data Mining*, volume 19 of *Pearson International Edition*. Addison Wesley, 2005. ISBN 0321321367. doi: 10.1016/0022-4405(81)90007-8.

YW Teh, MI Jordan, MJ Beal, e DM Blei. Hierarchical dirichlet processes. *Journal of the American Statistical*, pages 1–41, 2006.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 2 edition, 1995. ISBN 0-387-94559-8.

Renata Vieira e Massimo Poesio. An Empirically Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4):539–593, December 2000. ISSN 0891-2017. doi: 10.1162/089120100750105948.

Renata Vieira, P.N. Gonçalves, e J.G.C. de Souza. Processamento computacional de anáfora e correferência. *Revista de Estudos da Linguagem*, 16(1): 22, 2008.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, e Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding - MUC6 '95*, page 45, Morristown, NJ, USA, 1995. Association for Computational Linguistics. ISBN 1558604022. doi: 10.3115/1072399.1072405.

Christopher Walker, Stephanie Strassel, Julie Medero, e Kazuaki Maeda. ACE 2005 Multilingual Training Corpus, 2005.

Xiaojun Wan. Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval*, 11(1): 25–49, 2008. ISSN 1386-4564. doi: 10.1007/s10791-007-9037-5.

Xiaofeng Yang e Guodong Zhou. Improving noun phrase coreference resolution by matching strings, 2004.

Xiaofeng Yang, Guodong Zhou, Jian Su, e Chew Lim Tan. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*, pages 176–183, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075119.

Xiaofeng Yang, Jian Su, Guodong Zhou, e Chew Lim Tan. An NP-cluster based approach to coreference resolution. In *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, pages 226–es, Morristown, NJ, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220388.

Xiaofeng Yang, Jian Su, e Chew Lim Tan. Kernel-based pronoun resolution with structured syntactic knowledge. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 41 – 48, Morristown, NJ, USA, 2006a. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, e Chew Lim Tan. Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 41–48, Morristown, NJ, USA, July 2006b. Association for Computational Linguistics. doi: 10.3115/1220175.1220181.