Débora Zupeli Bossois

Metodologia de categorização de textos a partir de documentos não rotulados utilizando um processo de resolução de anáforas

Vitória - ES, Brasil

30 de agosto de 2010

Débora Zupeli Bossois

Metodologia de categorização de textos a partir de documentos não rotulados utilizando um processo de resolução de anáforas

Dissertação apresentada para obtenção do Grau de Mestre em Informática pela Universidade Federal do Espírito Santo.

Orientador:
Sérgio Antônio Andrade de Freitas

Departamento de Informática
Centro Tecnológico
Universidade Federal do Espírito Santo

Vitória - ES, Brasil

30 de agosto de 2010

Dissertação de Projeto Final de Mestrado sob o título "Metodologia de categorização de textos a partir de documentos não rotulados utilizando um processo de resolução de anáforas", defendida por Débora Zupeli Bossois e aprovada em 30 de agosto de 2010, em Vitória, Estado do Espírito Santo, pela banca examinadora constituída pelos professores:

Prof. Dr. Sérgio A. A. de Freitas
Orientador

Prof. Dr. Berilhes Borges Garcia
Universidade Federal do Espírito Santo

Dr. Emiliano Gomes Padilha
Universidade Federal do Rio Grande do Sul

Resumo

Com a constante expansão do conteúdo textual em formato eletrônico, surge a necessidade de organizar toda essa informação de forma operável. Desenvolveu-se, então, o processo de categorização de textos, visando facilitar a manipulação e recuperação da informação a partir da sua separação em categorias temáticas. Há diversas abordagens para a obtenção de um categorizador automático de textos e, dentre elas, o paradigma supervisionado é o mais tradicional. Apesar de a metodologia supervisionada apresentar uma precisão comparável àquela obtida por especialistas humanos, a obrigatoriedade de um corpus pré-classificado pode ser um fator limitador em certas aplicações.

Nessas situações, pode ser aplicada uma solução semi ou não supervisionada, que não exige um conjunto de treino completo e bem formado para a construção de um categorizador; pelo contrário, são somente fornecidos documentos não rotulados para o método. Tanto o paradigma de aprendizado de máquina supervisionado, quanto os paradigmas semi e não supervisionados, usualmente constroem uma representação dos textos baseado somente na ocorrência dos termos, não levando em conta fatores semânticos. Entretanto, muitas características intrínsecas da linguagem natural podem tornar o processo ambíguo, e um desses fatores é a utilização de termos diversos para a referência de uma entidade já apresentada no texto. A esse fenômeno linguístico, dá-se o nome de anáfora.

Esta dissertação propõe um método para a concepção de um categorizador não supervisionado, utilizando como base a Estrutura Nominal do Discurso (END), desenvolvida por Freitas com o propósito de resolução de anáforas, em [Freitas 2005]. Para isso, a técnica de *bootstrapping* para categorização é implementada, objetivando a obtenção da rotulação inicial para os documentos, a qual é utilizada para gerar um modelo de categorização através do paradigma supervisionado. Além de ter sido fundamentada a partir da END, a metodologia deste trabalho se beneficia do processo de resolução de anáforas de forma direta, utilizando os antecedentes identificados para as anáforas, durante a fase final da categorização.

O presente trabalho apresenta detalhes sobre a metodologia proposta, explanando os algoritmos desenvolvidos, bem como as experimentações realizadas para a avaliação do método. Os resultados mostram que a utilização do processo de resolução de anáforas é benéfica para um sistema de categorização não supervisionada.

Abstract

With the constant expansion of text content in electronic format comes the need to organize all this information in an operable way. Thus the text categorization process has been developed, aiming to make easier the manipulation and recovering of the information by separating it in thematic categories. There are many approaches to obtain an automatic text classification. Among then, the supervised learning is the most traditional. Though the supervised methodology is as much precise as the one obtained by human specialists, the obligatoriness of a pre-classified corpus might be a limiting factor in some applications.

In those situations, a semi- or unsupervised solution can be applied, wich does not demands a complete and well formed set of training to the building of a classifier; on the contrary, only unlabeled documents for the method are supplied. Both the supervised and the semi- and unsupervised learning usually built a text representation based only in the occurrence of the terms, not taking in consideration semantic factors. However, many intrinsic characteristics of the natural language can make the process ambiguous, and one of these factors is the use of diverse terms to refer to one entity already presented in the text. This linguistic phenomena is called anaphora.

This thesis proposes a method to concept an unsupervised classifier, using as a base the Nominal Structure of Speech (*Estrututra Nominal do Discurso* – END, in Portuguese), developed by Freitas with the objective of solving anaphora, in [Freitas 2005]. To accomplish the objective, the bootstrapping technique for classification is implemented, aiming to obtain the inicial labeled training data, wich is used to generate a classifying model through the supervised learning. Besides being grounded on the END, this paper methodology is benefited by the direct anaphora resolution process, using the antecedents identified for the anaphors, during the final classification phase.

This work presents details about the proposed methodol, as well as the trials and tests made to evaluate the method. The results show that the use of the anaphora resolution process is beneficial for an unsupervised learning system.

Dedicatória

Mais uma etapa vencida. Já foram muitas. Mas poucas, perto das que ainda virão. A cada batalha, o apoio incondicional; a cada conquista, o amor infindável. Mais uma vez, dedico a eles – meus pais.

Agradecimentos

Agradeço primeiramente à minha família, pelo apoio, força, incentivo, carinho, amor (...) incondicionais. Obrigada por me fazerem acreditar, todo dia, que sou uma pessoa especial.

Aos professores que, de alguma forma, me ajudaram na minha formação. Ao Sérgio, pela orientação, pelos conselho, pelos puxões de orelha e pelas doses de realidade quando eu mais precisava.

Aos meus "escraviários", principalmente ao Marquito (que me acompanha há mais tempo), por terem me ajudado, às vezes até altas horas da noite, implementando código, realizando testes... Obrigada pelo apoio, meninos!

Às minhas ex-companheiras de república, mas para sempre "irmãs" Paula e Dani – obrigada por compartilharem comigo, durante tanto tempo, a vida e a amizade de vocês.

À Pknucha, minha pequena grande amiga; obrigada pelas correções e, principalmente, por ter me dado o melhor presente que poderia receber nesta etapa final da dissertação!

À Lu, por ser amiga, irmã, confidente, "marida", às vezes mãe, às vezes filha. Não "só" por ter me ajudado em cada fase desta dissertação; muito mais do que isso: por ter me arrancado gargalhadas depois de noites mal dormidas, por ter feito comidinhas legais, preocupada com minha saúde, por ter entendido minha "cara de brava" quando tudo que queria era contar uma novidade. Obrigada!

Aos meus lindos amigos Cabelo e Salomão; obrigada por serem essa fofura toda (no bom sentido, que fique claro). À Lice, minha irmã gêmea adotada. Aos outros tantos amigos que praticamente não me viram enquanto eu estava na minha "bolha dissertística", mas que, tenho certeza, entenderam a minha ausência e estavam torcendo por mim.

Deus... a Ele, eu agradeço em orações.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Este capítulo apresenta as motivações, objetivos e metodologia deste trabalho, além de uma visão geral do que se encontra nesta dissertação.

1.1 Introdução

Com os avanços da tecnologia, há disponível uma capacidade cada vez maior de armazenamento e processamento de dados em larga escala. Atualmente, a grande maioria dos dispositivos de aquisição de dados são digitais e esse número só tende a crescer, principalmente com o aumento exponencial da Internet. A partir dessa revolução digital, logo surgiu a necessidade de organizar e gerenciar toda essa informação. O processo de Categorização de Textos – CT – apresenta este intuito: separar a informação em categorias de conhecimentos, que facilitem a sua manipulação e recuperação. Os documentos de interesse, sejam eles *online* ou não, são agrupados de forma que documentos que tratem do mesmo assunto permaneçam juntos.

Desde o surgimento da metodologia de Aprendizado de Máquina nos anos 90, algoritmos têm sido utilizados para classificação de textos através do paradigma supervisionado – ou simplesmente categorização supervisionada. O algoritmo de categorização supervisionada se baseia na construção automática de um classificador de textos através de um processo indutivo que aprende sobre as categorias de interesse, considerando um conjunto de instâncias previamente rotuladas – chamado conjunto de treino [Sebastiani 2002].

Há também a técnica de aprendizado semi ou não supervisionado visando a tarefa de categorização [Ghahramani 2004]. Nesse paradigma, não há a obrigatoriedade de um conjunto de treino completo e bem formado para a obtenção da classificação desejada e a construção de um categorizador. Muitas vertentes da abordagem semi ou não supervisionada são discutidas na literatura, tais como: utilização de uma técnica de *bootstrapping* para a definição da rotulação inicial [Mccallum e Nigam 1999, Liu et al. 2004, Adami, Avesani e Sona 2005, Gliozzo, Strapparava e Dagan 2009, Ko e Seo 2009]; agregação de documentos não rotulados a um conjunto pequeno de exemplos rotulados [Nigam et al. 1998, Ghani 2002]; e especificação de metodologias que consideram dados marcados de somente uma categoria de interesse [Jeon e Landgrebe 1999, Liu et al. 2002].

Tanto o paradigma de aprendizado de máquina supervisionado quanto os paradigmas semi e não supervisionados usualmente constroem uma representação dos textos baseado somente na ocorrência dos termos. De maneira geral, se uma entidade ocorre muito no texto de um dado documento D, sua relevância em D é grande; logo esse termo será importante na escolha da categoria à qual esse documento pertence.

Para exemplificação, considere que o texto do documento D é definido por:

"*Pedro gosta muito da sua bicicleta. Ele sonhou que estava pedalando pela*

> *cidade até ser acordado pela sua mãe o chamando. A pobre criança só queria ter dormido um pouco mais...*"

Uma pessoa, ao ler essas frases, facilmente percebe que *"Pedro"* é o assunto principal do texto. Entretanto, observa-se que o termo exato *"Pedro"* ocorre somente uma vez, o que, considerando os modelos de categorização tradicionais, resultaria em uma baixa relevância para esse termo em D. O problema é que esses modelos não identificam referências a um termo previamente mencionado. No texto em questão, os termos: *"sua"*, *"Ele"*, *"o"* e *"pobre criança"* referenciam exatamente à mesma entidade: *"Pedro"*. A esse fenômeno linguístico dá-se o nome de anáfora. O processo de resolução de uma anáfora consiste em estabelecer o relacionamento entre a entidade que introduz referência a um termo já apresentado e a entidade que é referenciada.

Este trabalho propõe um método de categorização não supervisionada, a partir documentos não rotulados e categorias pré-definidas, utilizando como base a estruturação desenvolvida para a resolução de anáforas apresentada em [Freitas 2005].

1.2 Motivação

Há muitas aplicações em que é necessário organizar determinados conteúdos em categorias de interesse. Em algumas situações, existe uma base de dados prévia disponível com informações sobre a rotulação de cada item; por exemplo, quando o usuário de um dado sistema define a classificação para cada conteúdo, à medida que ele é criado. Supondo que a demanda de conteúdo cresça, pode ser que se torne inviável a escolha da categoria de cada novo documento de forma manual, daí a necessidade de um método que realize essa categorização de forma automática. Nesse caso, é possível aplicar uma técnica de aprendizado supervisionado, já que existem, de antemão, dados pertencentes a esse domínio já rotulados.

Algoritmos que são baseados na abordagem supervisionada requerem uma grande quantidade de documentos rotulados para a construção do categorizador, visando uma aprendizagem mais apurada. Contudo, dados rotulados, além de não serem facilmente disponíveis para utilização, são também de difícil obtenção, uma vez que essa tarefa deve ser feita manualmente por um especialista de domínio, tornando o processo altamente custoso. Em alguns casos, a não obrigatoriedade dessa pré-rotulação pode ser mais interessante, mesmo que os resultados não sejam tão rigorosamente precisos.

Voltando ao exemplo, suponha que esse mesmo sistema deva fazer uma migração de

dados, mas que não seja possível recuperar a informação sobre o relacionamento entre os documentos e as categorias; ou, em uma outra situação, caso as regras do negócio se alterem e novas categorias sejam adicionadas e/ou as antigas sejam alteradas. Nesses casos, para realizar o trabalho de categorização utilizando uma abordagem supervisionada, seria necessário que o responsável pelo sistema definisse manualmente os rótulos dos documentos. Como essa tarefa exige um esforço grande do utilizador, surge a necessidade de um processo automático de categorização independente da disponibilidade de documentos marcados. Esse fato introduz uma grande motivação deste trabalho que, visando solucionar essa questão, propõe um método de categorização não supervisionada, no sentido de que são somente utilizados documentos não rotulados.

Além disso, este trabalho foi motivado por um outro fator, que leva em consideração justificativas linguísticas. Normalmente, as metodologias de categorização não utilizam técnicas de interpretação de texto, ou as consideram somente em determinados aspectos específicos. Entretanto, muitas características intrínsecas da linguagem natural podem tornar o processo ambíguo, e um desses fatores é a utilização de termos diversos para a referência de uma entidade já apresentada no texto – nomeadamente, anáforas. Os processos tradicionais de interpretação, que não consideram essa informação, não são capazes de capturar a semântica do texto de forma precisa. Isso ocorre pois um texto, por exemplo, pode tratar de um determinado assunto em todo o seu conteúdo, mas só apresentar o termo ao qual se refere pouquíssimas vezes; em todas as referências são utilizadas entidades que fornecem menor informação significativa. Se todas as referências são resolvidas, a semântica do texto se torna muito mais factível para interpretação. Visando a obtenção dessas vantagens, o método aqui proposto utiliza uma estruturação desenvolvida com o propósito de resolução de anáforas para a tarefa de categorização.

1.3 Objetivos

Esta dissertação apresenta como principais objetivos:

- Apresentar e implementar uma metodologia de categorização de textos não supervisionada baseada na estruturação desenvolvida para resolução de anáforas para a língua portuguesa. Para tal:
  - Construir a Estrutura Nominal do Discurso (END) para cada documento considerando as melhorias propostas no primeiro item;

  - Implementar um método de *bootstrapping* utilizando essa estrutura, visando a geração de uma rotulação inicial para os documentos;
  - Utilizar essa rotulação para a obtenção de um modelo de categorização.

- Analisar o processo de criação da END desenvolvido por Pereira em [Pereira 2009] e apresentar melhorias, buscando uma representação mais próxima da teoria fundamentada pela proposta de Freitas em [Freitas 2005].

- Avaliar o desempenho qualitativo do sistema através de dois mecanismos: (1) análise da Estrutura Nominal obtida para uma instância específica, explicando o passo-a-passo para sua criação e apresentando suas principais características, e (2) execução do método de categorização proposto a partir de um corpus de documentos marcados, permitindo, assim, a avaliação do mesmo.

1.4 Metodologia

Para a realização deste trabalho foi feita uma pesquisa aprofundada sobre técnicas de categorização de textos, passando por abordagens que utilizam o paradigma supervisionado, até o estabelecimento do foco de aplicação do método para o paradigma não supervisionado. Durante esse processo, foi realizado um levantamento de uma série de trabalhos relacionados à área. O trabalho apresentado por Ko e Seo em [Ko e Seo 2009] serviu como base para a definição do processo de *bootstrapping* para categorização.

A partir de um problema distinto da área de Processamento de Linguagem Natural (PLN), observou-se a possibilidade de junção dessas duas áreas de interesse visando uma tarefa de categorização com uma abordagem mais semântica. Foi realizado um estudo minuncioso acerca da proposta de Freitas sobre interpretação automatizada de textos para o processamento de anáforas [Freitas 2005], assim como acerca dos trabalhos subsequentes de Seibel Júnior [Júnior 2007] e de Pereira [Pereira 2009], que se baseiam nessa teoria, tendo como foco Recuperação de Informação. Foram identificadas algumas deficiências nesses trabalhos e propostas modificações visando a obtenção de um método mais preciso.

Com o conhecimento adquirido sobre essas propostas que utilizam o processo de resolução de anáforas para sua fundamentação, e tendo como apoio outros trabalhos que consideraram a junção dessas duas grandes áreas [Mitkov et al. 2007, Yeh e Chen 2003], foram estabelecidas as vantagens dessa aplicação e definida a metodologia de categorização que utiliza, não só o resultado da tarefa de resolução de anáforas, mas também toda a

estruturação desenvolvida para esse propósito.

1.5 Estrutura da dissertação

Esta dissertação está estruturada da seguinte maneira: no capítulo 2 são abordados conceitos sobre a área de categorização de textos, bem como uma diversidade de trabalhos desenvolvidos para esse propósito, realizando uma pesquisa ampla em torno dos seus paradigmas.

O capítulo 3 introduz conceitos da área de PLN utilizados neste trabalho, em especial os que se referem a anáforas e à estruturação do discurso, que são as bases para o desenvolvimento da metodologia aqui apresentada. Em seguida, no capítulo 4, são apresentados os algoritmos desenvolvidos para a construção da estrutura e para a obtenção do modelo de categorização.

No capítulo 5 são apresentados os experimentos realizados para a avaliação do método de categorização proposto. Por fim, o capítulo 6 conclui este trabalho, apresentando as observações finais e direcionamentos para pesquisas futuras.

2 Categorização de textos

Neste capítulo são introduzidos conceitos e definições acerca do processo de categorização de textos, além de uma apresentação geral de uma série de trabalhos relacionados à área.

2.1 Introdução

Nas últimas décadas, observou-se um crescimento exponencial da informação textual em formato eletrônico, em níveis impensados, apoiado, obviamente, por um poderoso *hardware* em constante evolução. A Internet permitiu a disseminação mundial da informação e do conhecimento, além da colaboração e interação entre indivíduos independente de suas localizações geográficas. Com isso, o que se vê atualmente é a crescente substituição da utilização do conteúdo de forma analógica para sua utilização em formato digital.

Ao longo dessa revolução, surgiu a necessidade de organizar e gerenciar toda essa informação de forma operável, em termos de armazenamento e processamento. Desenvolveu-se, então, o processo de **categorização de textos** (CT – também conhecida como classificação automática de documentos), visando facilitar a manipulação e recuperação da informação a partir da sua separação em categorias temáticas. CT está sendo aplicada em muitos contextos, desde a indexação de documentos com base em um vocabulário controlado, a filtragem de documentos, a geração automática de metadados, *word sense disambiguation*, população de catálogos hierárquica de recursos da Web e, em geral, qualquer aplicação que exija organização de documento ou expedição selectiva e adaptativa de documento.

Até o final dos anos 80 a abordagem mais popular na comunidade de CT, principalmente nas aplicações de mundo real, foi a Engenharia do Conhecimento, que consiste em definir manualmente um conjunto de regras de codificação de conhecimentos específicos sobre como classificar os documentos sob as categorias determinadas.

A partir dos anos 90 esta abordagem perdeu popularidade em favor do paradigma de Aprendizado de Máquina, segundo ao qual um processo indutivo geral cria automaticamente um classificador de textos através da aprendizagem das características das categorias de interesse, a partir de um conjunto de documentos pré-classificados. As vantagens desta abordagem estão em uma precisão comparável àquela obtida por especialistas humanos, e em uma economia considerável em termos de força de trabalho de especialistas, uma vez que a intervenção de engenheiros de conhecimento é necessária, não somente para a classificação de um dado corpus, mas também para a criação de um construtor automático de classificadores, dado um conjunto de documentos manualmente classificados. Na terminologia do Aprendizado de Máquina, o problema de classificação é uma atividade de **aprendizado supervisionado**, uma vez que o processo de aprendizagem é

gerido pelo conhecimento das categorias e das instâncias de treino que pertencem a elas.

O **aprendizado não supervisionado**, por sua vez, é uma classe de problemas em que se procura determinar como os dados são organizados, não sendo necessária sua alimentação por um conjunto de dados rotulados[1]. O objetivo é a construção de representações da entrada, visando sua organização. Em certo sentido, a aprendizagem não supervisionada pode ser pensada como uma forma de encontrar padrões em dados que a princípio seriam considerados puros ruídos desestruturados. Voltado especificamente para categorização, a abordagem não supervisionada se refere à tarefa de classificar documentos, sem a necessidade de um conjunto prévio de instâncias rotuladas para sua execução.

O objetivo deste capítulo é explicar o processo de categorização de textos, descrevendo os paradigmas supervisionado (detalhado na seção 2.2) e semi e não supervisionado (seção 2.3). São mostradas as características intrínsecas das abordagens, ressaltando os pontos fortes e pontos fracos de cada uma delas. Além disso, na seção 2.4, é introduzida a técnica de *bootstrapping* para categorização e apresentada uma série de trabalhos desenvolvidos para esse propósito. Por fim, na seção 2.5, são mostrados os trabalhos que utilizam o processo de resolução de anáforas para categorização, e é estabelecido um breve histórico sobre a utilização da Estrutura Nominal do Discurso em diversas abordagens, partindo da sua proposta inicial.

2.2 Categorização supervisionada

De forma geral, categorização de textos [Sebastiani 2002] é a tarefa de associar um valor verdadeiro ou falso para cada par $\langle d_i, c_j \rangle \in D \times C$, onde D é o domínio dos documentos e $C = \{c_1, c_2, \ldots, c_{|C|}\}$ é o conjunto das categorias predefinidas. Um valor verdadeiro (V) é associado a $\langle d_i, c_j \rangle$, indicando a decisão de relacionar d_i a c_j, enquanto que um valor falso (F) indica a decisão de não relacioná-los. Mais formalmente, a tarefa é aproximar uma função alvo desconhecida $\breve{\Phi} : D \times C \rightarrow \{\mathrm{V}, \mathrm{F}\}$, que descreve como os documentos devem ser classificados, em relação a uma função $\Phi : D \times C \rightarrow \{\mathrm{V}, \mathrm{F}\}$ chamada de classificador, de tal forma que $\breve{\Phi}$ e Φ coincidam o máximo possível.

O paradigma de Aprendizado de Máquina, ou simplesmente, categorizador supervisionado, depende da disponibilidade de um corpus inicial $\Omega = \{d_1, \ldots, d_{|\Omega|}\} \subset D$

[1] A definição exata da expressão "não supervisionado" na literatura está ambígua. O termo "supervisão" pode remeter a qualquer atuação do operador no sistema, inclusive para a definição de palavras-chave para as categorias, por exemplo. Neste trabalho, contudo, foi assumido que um categorizador não supervisionado é aquele que não depende de dados rotulados para sua geração, independentemente de outras possíveis intervenções humanas.

de documentos pré-classificados sob $C = \{c_1, \ldots, c_{|C|}\}$. Ou seja, os valores da função $\breve{\Phi} : D \times C \rightarrow \{\mathrm{V}, \mathrm{F}\}$ são conhecidos para cada par $\langle d_i, c_j \rangle \in \Omega \times C$. Um documento d_i é um exemplo positivo de c_j se $\breve{\Phi}(d_i, c_j) = \mathrm{V}$, e um exemplo negativo de c_j se $\breve{\Phi}(d_i, c_j) = \mathrm{F}$.

Em contextos de pesquisa, uma vez que um classificador Φ foi construído é desejável avaliar a sua eficácia. Nesse caso, antes da construção do classificador, o corpus inicial de documentos é dividido em dois conjuntos, não necessariamente de mesmo tamanho:

- um conjunto de treino $Tr = \{d_1, \ldots, d_{|Tr|}\}$. O classificador Φ para as categorias $C = \{c_1, \ldots, c_{|C|}\}$ é indutivamente construído pela observação das características desses documentos;
- um conjunto de teste $Te = \{d_{|Tr|+1}, \ldots, d_{|\Omega|}\}$, usado para testar a eficácia do classificador. O classificador é alimentado por cada $d_i \in Te$, e a decisão $\Phi(d_i, c_j)$ é comparada com a decisão do especialista $\breve{\Phi}(d_i, c_j)$. Uma medida da eficáfia da classificação é baseada em quantas vezes os valores $\Phi(d_i, c_j)$ correspondem aos valores de $\breve{\Phi}(d_i, c_j)$.

Esta abordagem é chamada de treino-e-teste. Uma alternativa é a abordagem chamada validação cruzada (*k-fold crossvalidation approach*) [Mitkov 2005], na qual k diferentes classificadores $\Phi_1, \ldots, \Phi_k$ são construídos pelo particionamento do corpus inicial em k conjuntos disjuntos $Te_1, \ldots, Te_k$ e pela aplicação iterativa da abordagem de treino-e-teste nos pares $\langle Tr_i = \Omega - Te_i, Te_i \rangle$. A eficácia final é obtida pelo cálculo da eficiência individual de $\Phi_1, \ldots, \Phi_k$, e em seguida a média dos resultados individuais de alguma forma.

Há uma série de algoritmos de categorização que utilizam o paradigma supervisionado. Dentre eles, pode-se destacar o *naive Bayes*, que é definido como um classificador probabilístico baseado na aplicação do teorema de Bayes [McCallum e Nigam 1998], assumindo a premissa de que as variáveis envolvidas são independentes. Apesar de sua concepção aparentemente simplificada, esse tipo de categorizador trabalha muito bem em situações complexas do mundo real – como sua utilização para ambiente web [Zhang et al. 2009], aplicação do método visando dados imprecisos [Ren et al. 2009], categorização de sequências de RNA em uma dada taxonomia de bactérias [Wang et al. 2007], abrangendo, neste último caso, inclusive a área médica. Em [Mitchell 1997] são apresentados maiores detalhes sobre o algoritmo de *naive Bayes* para categorização de textos.

Um outro exemplo de algoritmo para categorização supervisionada é o *k-Nearest Neighbor* (k-NN) [Yang, Slattery e Ghani 2002, Goldstein 1972], que consiste na associ-

ação de um dado documento à categoria dos exemplos mais próximos, considerando que o corpus está disposto em um espaço vetorial. Em casos nos quais ocorre uma alteração frequente na base de dados, a utilização desse categorizador é uma boa escolha, já que não é necessário um processamento inicial para a geração de um classificador a partir do conjunto de treino. Em [Khan, Ding e Perrizo 2002, Han, Karypis e Kumar 2001], o método de k-NN é adaptado, visando a obtenção de melhores resultados.

O classificador *Rocchio* é definido como um método de realimentação de relevância (*relevance feedback*), encontrado em sistemas de recuperação de informação [Rochio 1971]. Os seguintes trabalhos utilizam a implementação desse algoritmo: [Ragas e Koster 1998, Rogati e Yang 2002, Lewis et al. 1996].

Os métodos de classificação supervisionada citados, normalmente, utilizam o espaço vetorial como representação dos documentos do corpus considerados. Além dessa representação, citam-se duas outras abordagens: *Support Vector Machine* (SVM) [Joachims 2002, Vapnik 1995] e redes neurais [Rogova 1994].

Algoritmos baseados no método supervisionado geralmente apresentam bons resultados. Quando há disponível uma base de dados rotulados bem definida e bem caracterizada, a aplicação da maioria das metodologias supervisionadas permite a obtenção de uma boa classificação das instâncias consideradas. Este, entretanto, é um limitador para esse tipo de abordagem: caso não haja um corpus pré-rotulado ou caso ele seja impreciso, o algoritmo tenderá a uma classificação equívoca.

2.3 Categorização semi e não supervisionada

Na prática, nem sempre há disponíveis grandes quantidades de dados marcados para a aplicação de um esquema supervisionado a um problema real, da mesma forma que, em muitos casos aplicados, a tarefa de coletar manualmente os dados rotulados é muito custosa. Por outro lado, coleções de textos não identificados são, em geral, facilmente disponíveis. Muitos usuários de um sistema prático preferem algoritmos que apresentam resultados minimamente satisfatórios àqueles que exigem uma grande quantidade de rotulagem manual. Sendo assim, um esquema semi ou não supervisionado pode ser aplicado nesses casos.

No paradigma não supervisionado [Ghahramani 2004], a máquina simplesmente recebe os dados genéricos de entrada $X = \{x_1, x_2, \ldots, x_{|X|}\}$, sem qualquer tipo de *feedback* do operador ou especialistas de domínio, e apresenta como objetivo a construção de repre-

sentações dessa entrada. Voltado especificamente para categorização, a abordagem não supervisionada – ou categorização não supervisionada – é a tarefa de associar um valor verdadeiro ou falso para cada par $\langle d_i, c_j \rangle \in D \times C$, sendo que os valores da função $\breve{\Phi} : D \times C \rightarrow \{\mathrm{V}, \mathrm{F}\}$, para cada um desses pares, são **desconhecidos**.

Clustering [Jain, Murty e Flynn 1999, Kaufman e Rousseeuw 1990], ou agrupamento, é um exemplo clássico de aplicação do paradigma não supervisionado, sendo baseado, de forma geral, na divisão do conjunto de entrada X em conjuntos menores, de tal modo que os pontos em um mesmo subconjunto sejam mais semelhantes entre si do que os pontos nos demais subconjuntos. Existem trabalhos que sugerem a utilização de técnicas de *clustering* visando a tarefa de classificação de documentos não supervisionada. Em [El-Yaniv e Souroujon 2001, Slonim e Tishby 2000, Slonim, Friedman e Tishby 2002], é dada uma coleção de documentos não rotulados e o objetivo é determinar *clusters* (agrupamentos) que estão altamente correlacionados com as verdadeiras categorias dos documentos. Adami et al., em [Adami, Avesani e Sona 2003], propõe um processo semiautomático cujo objetivo é minimizar o esforço requerido dos administradores do sistema ao criar, modificar e manter as taxonomias com documentos rotulados.

Quando há disponível uma certa quantidade de dados previamente classificados, ou quando a execução dessa tarefa é factível, podem ser utilizados algoritmos que são capazes de aprender através de um número pequeno de exemplos rotulados. Ghani, em [Ghani 2002], utiliza essa asserção no desenvolvimento de um *framework* que incorpora dados não rotulados na configuração da técnica de *Error-Correcting Output Coding* (ECOC), proposta em [Ghani 2000], através da decomposição de problemas multi-classes em múltiplos problemas binários e, então, utilizando um algoritmo de *Co-Training* para o aprendizado dos problemas binários de classificação. Além de apresentar bons desempenhos em termos de exatidão, essa abordagem também proporciona um balanço suave entre as grandezas precisão e *recall*, o que é útil em aplicações que requerem resultados de alta precisão.

Seguindo a mesma linha de Ghani, Nigam et al. em [Nigam et al. 1998] mostram que a acurácia de classificadores de textos treinados com um número pequeno de documentos rotulados pode ser melhorada através do aumento desse conjunto de treino com uma série de documentos não rotulados. O método se baseia na combinação da técnica de *Expected Maximization* (EM) com um classificador supervisionado *naive Bayes*. O algoritmo treina um classificador usando os documentos rotulados à disposição, e probabilisticamente rotula os documentos não marcados; em seguida, treina um novo classificador usando os

rótulos de todos os documentos. Resultados experimentais mostram que o uso de dados não rotulados reduz o erro do categorizador em 33%.

Em [Jeon e Landgrebe 1999] e [Liu et al. 2002], os autores definem métodos de categorização parcialmente supervisionada, considerando a necessidade de somente uma classe, ou um número pequeno de classes, de documentos rotulados positivamente. O problema tratado em [Jeon e Landgrebe 1999] parte de 1 (um) *cluster* conhecido; a definição e as estatísticas das outras classes são automaticamente desenvolvidas através de um processo de *clustering* não supervisionado ponderado, mantendo a identidade da classe de interesse. Depois que todas as classes são desenvolvidas, um classificador supervisionado convencional é utilizado na categorização. Os resultados experimentais, tanto com dados reais quanto simulados, verificaram a eficácia do método. Liu et al. em [Liu et al. 2002] tratam a categorização como um problema de otimização restrito e mostram que, sob condições adequadas, as soluções para o problema de otimização fornecem bons resultados para o problema de classificação parcialmente supervisionado. Em seu trabalho, os autores apresentam a técnica desenvolvida para tal e demonstram a sua eficácia através de experimentação extensiva.

Todos os métodos citados nesta seção que são baseados em categorização semissupervisionada requerem pelo menos um corpus pequeno de documentos rotulados, para a partir daí utilizarem a metodologia específica de cada um. Além disso, nenhum deles – incluindo os métodos apresentados que são focados no paradigma não supervisionado, que desconsideram essa necessidade – utiliza as informações semânticas do texto, que podem melhorar seu entendimento e, com isso, a tarefa de categorização. O foco deste trabalho, entretanto, são aplicações para as quais não há disponível nenhuma rotulação prévia e, mais do que isso, a proposta considera o ganho semântico introduzido por um processo de interpretação de textos.

2.4 *Bootstrapping* para categorização

Bootstrapping é uma técnica desenvolvida para melhorar iterativamente o aprendizado de um dado sistema utilizando dados não rotulados. É inicializado com uma pequena quantidade de informação que pode assumir muitas formas. Cada iteração apresenta duas etapas: (1) as etiquetas dos dados não marcados são estimadas a partir do modelo de aprendizado e (2) os dados não rotulados e os rótulos estimados são incorporados como dados de treinamento dentro do sistema.

Abordagens de *bootstrapping* têm sido utilizadas para várias vertentes, como: extração de informação [Riloff e Jones 1999], *word sense disambiguation* [Yarowsky 1995], classificação de hipertexto [Blum e Mitchell 1998] e categorização de textos de forma geral. Essa última abordagem se caracteriza pela geração de um modelo de categorização utilizando um classificador supervisionado, sem a dependência de um corpus completo de documentos marcados. O processo geral de funcionamento dessa abordagem é que, inicialmente, deve ser obtida uma rotulação para os documentos, que servirá como entrada para o categorizador supervisionado. Algoritmos que apresentam essas características seguem uma metodologia chamada, neste trabalho, de técnica de *bootstrapping* para categorização.

Em [Ko e Seo 2004, Ko e Seo 2009], Ko e Seo definem um método de categorização de documentos não rotulados utilizando um processo de *bootstrapping* e uma técnica de projeção de características. O *framework* proposto pelos autores é descrito pelos seguintes passos:

- Pré-processamento: os documentos são reestruturados através de *contextos* [Manning e Schtze 1999] e os termos de interesse (*content words*) são extraídos dos mesmos.

- Construção dos *clusters* de contextos para treino: são definidas palavras-chave para cada categoria, através de um processo que utiliza informação de coocorrência nos documentos, entre os nomes das categorias e os demais termos contidos no corpus. Em seguida, são definidos como contextos-centroides aqueles que apresentam em seu conteúdo o nome da categoria ou alguma palavra-chave. Por fim, os contextos restantes são associados às categorias através de métricas de similaridade.

- Aprendizagem do classificador: é utilizado o categorizador TCFP para a geração do modelo final. TCFP foi desenvolvido em um trabalho anterior dos mesmos autores [Ko e Seo 2002], a partir da técnica de projeção de características, e tem a propriedade de ser mais robusto para dados ruidosos do que outros algoritmos de aprendizagem.

Ko e Seo relataram resultados comparáveis aos dos classificadores supervisionados, com a grande vantagem, obviamente, de não necessitarem de uma base de dados rotulados.

O conceito da definição de palavras-chave para as categorias também é explorado por McCallum e Nigam em [Mccallum e Nigam 1999]. Inicialmente, são estipulados rótulos aos documentos com base na correspondência entre os termos presentes no texto e as

palavras-chave. Essa rotulação prévia se torna o ponto de partida do processo de *bootstrapping*, que gera um classificador *naive Bayes* utilizando um algoritmo de *Expectation-Maximization* – visando estimar os rótulos para os documentos restantes – e uma técnica estatística de *shrinkage* – com o objetivo de melhorar a estimativa obtida. Essa metodologia foi posteriormente aplicada na criação de um sistema que automatiza a construção de portais de Internet em [McCallum et al. 2000].

Seguindo a mesma linha, em [Gliozzo, Strapparava e Dagan 2009], é proposto um algoritmo de categorização que parte de informações de interesse para a caracterização das categorias. São introduzidas duas técnicas para melhorar a rotulação obtida no processo de *bootstrapping*: utilização de espaços semânticos latentes para a estimativa de similaridade entre documentos e termos, e aplicação do algoritmo de misturas de Gaussianas (*Gaussian mixture algorithm*), capaz de diferenciar as informações sobre as categorias que são relevantes daquelas não relevantes, a partir dos exemplos não rotulados. A performance qualitativa obtida pelo método mostrou ser equiparável com uma abordagem supervisionada.

O trabalho de Adami et al. [Adami, Avesani e Sona 2003, Adami, Avesani e Sona 2005] propõe um modelo chamado TaxSOM, que agrupa um conjunto de documentos em uma determinada hierarquia de classes, explorando diretamente o conhecimento sobre a organização topológica e descrição léxica das categorias. Os experimentos realizados, segundo os autores, apresentaram bons resultados.

Esta seção apresentou várias abordagens existentes na literatura para o processo de *bootstrapping* para categorização. Considerando aplicações para as quais não há disponível uma base de dados rotulados, essa técnica é um excelente artifício para a resolução da tarefa de categorização de textos, como pôde ser demonstrado a partir dos resultados alcançados pelos trabalhos citados. Contudo, existe um ganho ainda maior que pode ser considerado: a utilização do Processamento de Linguagem Natural para a interpretação dos textos. A próxima seção apresenta referências bibliográficas acerca do processo de resolução de anáforas voltado para categorização, que é a base principal deste trabalho.

2.5 Utilização de anáforas em categorização

Considerando a tarefa de resolução de anáforas voltada especificamente para a área de categorização, o trabalho de Mitkov et al. [Mitkov et al. 2007] mostra como um sistema de resolução de anáforas pronominais para o inglês pode melhorar a performance de

três problemas que envolvem linguagem natural: resumo de textos, extração de termos e categorização de documentos. Para o caso particular da categorização, em todos os documentos, os pronomes são substituídos pelos sintagmas nominais reconhecidos como seus antecedentes pelo sistema de resolução de anáforas considerado. A partir desses novos documentos, foram testados quatro tipos de métodos de categorização, obtendo resultados melhores em relação ao modelo inicial.

Seguindo esse mesmo princípio, Yeh e Chen em [Yeh e Chen 2003] empregam um método de resolução de elipses (chamado por eles de *zero anaphora resolution*) com o objetivo de recuperar as anáforas omitidas no texto, considerando o idioma chinês. O documento resultante desse processo alimenta o sistema de categorização, de forma que as referências anteriormente omitidas no conteúdo através das elipses passem a contribuir no cálculo da geração do categorizador. Os resultados dos experimentos mostram que o método de resolução de elipses aumenta a exatidão do categoridador de textos de 79% para 84%.

Os trabalhos de Mitkov et al. [Mitkov et al. 2007] e Yeh e Chen [Yeh e Chen 2003] se assemelham pois utilizam o resultado do processo de resolução de anáforas (anáfora pronominal e elipse, respectivamente) em um sistema de categorização supervisionada, visando melhorar seus resultados. Isso é possível devido ao fato de que uma anáfora não apresenta um conteúdo semântico apropriado, uma vez que ela referencia, através de um outro termo ou sua omissão, a uma entidade de interesse já apresentada no texto. Portanto, um processo de categorização que considere a substituição das anáforas pelas suas entidades antecedentes, que são as mais relevantes para a semântica geral do texto, atinge esses benefícios.

Este trabalho segue essa linha de propósito, entretanto com três diferenças básicas. A primeira delas é a mais evidente: o objetivo é o idioma português, para o qual não foi encontrado nenhuma proposta semelhante na literatura. A segunda remete à abordagem do sistema de resolução de anáforas considerado, que visa tratar os seguintes tipos: anáforas pronominais e anáforas nominais definidas. Além disso, os trabalhos anteriores se diferem em relação a esta proposta no que diz respeito à utilização do categorizador supervisionado. Mais do que simplesmente considerar a resolução anafórica na tarefa de categorização, o objetivo é tratar um problema ainda mais complexo: categorização não supervisionada.

Para isso, este trabalho toma como base a abordagem de Freitas, em [Freitas 2005], para resolução de anáforas pronominais, anáforas nominais definidas e elipses, utilizando

regras pragmáticas para a identificação dos antecedentes das anáforas. Além disso, Freitas proporciona uma metodologia para a obtenção de uma representação estruturada do documento, chamada Estrutura Nominal do Discurso (END). O capítulo 3 apresenta detalhes sobre a definição e utilização de anáforas, bem como sobre as características intrínsecas da END e o seu processo de criação.

Seibel Júnior, em [Júnior 2007], utiliza a Estrutura Nominal do Discurso proposta por Freitas [Freitas e Lopes 1993, Freitas e Lopes 1994, Freitas e Lopes 1995] para a tarefa de recuperação de informação, apresentando uma metodologia para a realização de buscas considerando essa estrutura. Seibel propõe uma modificação na estrutura de maneira que a mesma armazene somente os termos indexados e seus valores de relevância para o documento.

Em [Pereira, Morellato e Freitas 2009], os autores apresentam um modelo de recuperação estrutural de informação também utilizando a END. O trabalho fez uso de sintagmas nominais a fim de permitir uma melhor representação de texto. Esse trabalho buscou mostrar os benefícios que a área de recuperação de informação alcança ao utilizar a Estrutura Nominal do Discurso, além de apresentar uma comparação do sistema desenvolvido baseado em anáfora, com o tradicional modelo vetorial.

A partir também da proposta de Freitas [Freitas 2005], e das modificações propostas por Seibel Júnior [Júnior 2007], em [Pereira, Júnior e Freitas 2009] é apresentada uma nova metodologia para a RI baseada na resolução de anáforas. A construção da estrutura para buscas é realizada transpondo todas as entidades identificadas durante o processo de resolução anafórica, o que possibilita uma melhora na forma de representação do texto dos documentos e na qualidade dos resultados obtidos pelas pesquisas. Pereira, em [Pereira 2009], detalha a proposta descrita em [Pereira, Júnior e Freitas 2009], apresentando os algoritmos envolvidos na sua definição e experimentações sobre a nova metodologia de buscas baseada na resolução de anáforas.

Com essa varredura na literatura, primeiramente sobre os trabalhos relacionados à categorização e, por fim, sobre os trabalhos que utilizam a Estrutura Nominal do Discurso para resolução de anáforas, a proposta deste trabalho pode ser apresentada. No capítulo seguinte, ela é descrita, bem como todos os conceitos e informações vinculados ao seu entendimento.

3 Método de categorização proposto

Este capítulo apresenta os conceitos da área de Processamento de Linguagem Natural utilizados no desenvolvimento do trabalho e detalhes sobre o método proposto de categorização de textos não supervisionada, a partir de um processo de resolução de anáforas.

3.1 Introdução

Neste capítulo são apresentados conceitos da área de linguagem natural necessários para o entendimento e a construção do modelo de categorização proposto neste trabalho. A base da proposta está na utilização das vantagens do processo de resolução de anáforas na área de categorização de textos não supervisionada. Define-se anáfora como o fenômeno linguístico de realizar uma referência a uma entidade já apresentada no texto. Por sua vez, o processo de resolução de anáforas consiste em estabelecer o relacionamento entre a anáfora e a entidade que está sendo referenciada, resultando em uma relevância apropriada para essa entidade em relação ao documento, mesmo que sejam utilizados termos diversos para sua referência. Em CT, essa característica é importante para o seu bom desempenho.

Ao realizar a leitura de um texto, uma pessoa consegue identificar naturalmente esse relacionamento entre anáforas e seus antecedentes. Para isso, entretanto, podem ser utilizadas diversas informações, baseadas na estruturação léxica do texto, sintática e/ou semântica, de forma que o processo pode se tornar complexo [Hobbs 1986]. Na literatura, há várias abordagens para a automatização desse processo [Lappin e Leass 1994, Iida, Inui e Matsumoto 2005, Palomar et al. 2001, Chaves e Rino 2008]. O modelo de categorização proposto parte dos métodos e estruturas desenvolvidos para o problema de resolução de anáforas em [Freitas 2005]. Freitas apresenta a Estrutura Nominal do Discurso, sobre a qual foi aplicada a técnica de *bootstrapping* – geração da rotulação inicial dos documentos, servindo como entrada para um categorizador supervisionado.

Há uma série de algoritmos para categorização supervisionada que resultam em uma boa classificação (como foi mostrado na seção 2.2). Todavia, considerando domínios complexos, esses algoritmos geralmente requerem conjuntos de treino extremamente grandes para alcançarem resultados precisos. A criação desses conjuntos de dados rotulados, se já não disponíveis, é altamente custosa, uma vez que devem ser feitas por um especialista humano [Mccallum e Nigam 1999]. Este fato leva à necessidade da obtenção de um método não supervisionado, ou pelo menos que exija o mínimo possível da atuação humana (método semissupervisionado), tendo como entrada documentos não rotulados e as categorias nas quais eles devem ser classificados. A técnica de *bootstrapping* proposta permite a obtenção desse categorizador.

Na seção 3.2 são apresentadas definições e descrições relativas ao conceito de anáforas. Em seguida, na seção 3.3, são mostradas as principais características da Estrutura Nominal do Discurso para enfim, nas duas seções seguintes – seção 3.4 e seção 3.5 –,

ser apresentado o método de categorização de documentos não rotulados, baseado nessa estrutura e no processo desenvolvido para resolução de anáforas, especificando o procedimento de *boostrapping* proposto. Na seção 3.6, o método de categorização é concluído, com a aplicação do resultado da rotulação em um categorizador supervisionado.

3.2 Anáforas

Anáfora é definida como o fenômeno linguístico de referenciar a um item previamente mencionado no texto através de uma expressão linguística mais simples [Mitkov 2005]. Esse fenômeno é altamente frequente em produções discursivas em linguagem natural. No processo de escrita de um texto, o escritor constrói um discurso com base em uma estruturação coerente das ideias que ele deseja transmitir ao leitor. Por sua vez, o leitor realiza uma interpretação incremental do discurso formulado pelo escritor. À medida que o discurso progride, o escritor pode vir a referenciar uma entidade que já tenha sido citada anteriormente, a partir da utilização de um termo diferente. Nesse caso, a expressão de referência utilizada é uma anáfora.

Formalmente, a expressão linguística que introduz uma referência a uma entidade já apresentada no discurso é denominada **expressão anafórica** ou simplesmente **anáfora**. A informação previamente introduzida é denominada **antecedente** e o processo pelo qual é identificado o antecedente de uma expressão anafórica é denominado **resolução anafórica** ou **resolução de anáforas**. Tanto a anáfora quanto o antecedente são representados como referentes do discurso.

Considere o seguinte texto:

"*José organizou uma festa.*
Ele não esqueceu de nenhum detalhe." (3.1)

O pronome *"Ele"* do texto (3.1) é uma entidade anafórica, que referencia o antecedente *"José"*. Como a anáfora neste caso trata-se de um pronome, ela é classificada como anáfora pronominal.

Agora considere o próximo texto:

"*José gostou muito do bolo.*
O doce não poderia estar melhor." (3.2)

No processo de interpretação, humano ou computacional, a utilização de um artigo definido qualquer na precedência de um substantivo é um indicativo de que a entidade já foi anteriormente introduzida no discurso, apresentando um caráter anafórico. Esta entidade é classificada sintaticamente como um Sintagma Nominal Definido (SND) e, devido à sua característica anafórica, é chamada como Anáfora Nominal Definida (AND). Em (3.2), o termo *"O doce"* é uma AND, que possui como antecedente o termo *"bolo"*.

Há casos em que o escritor simplesmente omite uma anáfora em um texto, pelo fato de considerar a referência ao elemento em questão explícita o suficiente para o leitor, como acontece no texto abaixo:

"*Os enfeites estavam lindos.* (3.3)
θ Combinaram perfeitamente com a festa."

O fenômeno linguístico exemplificado no texto (3.3) chama-se elipse: o leitor compreende facilmente que o assunto da segunda frase continua sendo *"Os enfeites"* da primeira frase. O símbolo θ apresenta o ponto em que a anáfora seria apresentada.

3.2.1 Resolução de anáforas

O processo de resolução de anáforas consiste em estabelecer o relacionamento entre a anáfora e a entidade sendo referenciada – seu antecedente. No caso de anáforas pronominais, esse relacionamento é resolvido apenas pela identificação do antecedente; diferente do que ocorre com anáforas nominais definidas, em que, além da determinação do antecedente, é necessária a identificação da relação existente entre o antecedente e a expressão anafórica. Para exemplificar, considere o texto:

"*O bolo estava uma delícia.* (3.4)
O recheio era de dar água na boca."

No texto (3.4), a anáfora nominal definida *"O recheio"* deve ser resolvida não somente pela identificação do antecedente – *"O bolo"* –, mas também pela identificação da relação existente entre o antecedente e a expressão anafórica – no contexto do exemplo, fica claro que recheio é **parte** do bolo.

Assim, a interpretação das anáforas nominais definidas ou de qualquer outro fenômeno anafórico pode ser generalizada como um processo que atribui valores aos itens da seguinte

equação:

$$\mathcal{R}(\mathcal{A}, \mathcal{T}), \tag{3.5}$$

onde: $\mathcal{A}$ denota uma entidade introduzida pela expressão anafórica via pronome, SND ou elipse, $\mathcal{T}$ denota o seu antecedente e $\mathcal{R}$ denota a relação existente entre $\mathcal{A}$ e $\mathcal{T}$. O processo de resolução da equação, que é propriamente o processo de resolução de anáforas, consiste em descobrir $\mathcal{T}$ e $\mathcal{R}$ dado $\mathcal{A}$.

Anáforas pronominais apresentam uma forte dependência para com o seu antecedente e surgem somente com a função de substituir o antecedente no decorrer do discurso. Nesse caso, basta identificar o antecedente $\mathcal{T}$ para a anáfora $\mathcal{A}$, sendo a única relação possível a de correferência.

Já as anáforas nominais definidas podem ter significado de forma independente de seus antecedentes, podendo inclusive fornecer mais informação sobre o antecedente no decorrer do discurso. Assim, para o processo de interpretação das ANDs, além de identificar o antecedente $\mathcal{T}$, é necessário identificar a relação $\mathcal{R}$ existente entre $\mathcal{A}$ e $\mathcal{T}$, sem a qual não é possível obter uma interpretação plausível.

Freitas propõe em [Freitas 2005] uma metodologia computacional que interpreta as anáforas nominais definidas cuja relação $\mathcal{R}$ é uma dentre: **correferência**, "**membro de**", "**parte de**", "**subcategorizado por**" e **acomodação**. A seção 3.2.2 detalha a obtenção dessas relações através das regras pragmáticas. Em seguida, a seção 3.2.3 define o foco de um discurso, que é utilizado na criação da Estrutura Nominal do Discurso – tratada na seção 3.3.

3.2.2 Regras pragmáticas

Baseado no conhecimento que as pessoas têm sobre a língua que falam, é possível estabelecer um conjunto pragmático de regras a serem utilizadas na determinação da relação entre a expressão anafórica e seus antecedentes. As informações sobre gênero, número, coletivos e animacidade podem ser utilizadas na determinação das seguintes relações:

- Correferência: indicando que tanto $\mathcal{A}$ quanto $\mathcal{T}$ denotam a mesma entidade, ou seja, $\mathcal{A} = \mathcal{T}$.
- "Membro de": indicando que a entidade denotada por $\mathcal{A}$ é um membro do conjunto de entidades denotada por $\mathcal{T}$.

- "Parte de": indicando que a entidade denotada por $\mathcal{A}$ é parte (estrutural) da entidade denotada por $\mathcal{T}$.
- "Subcategorizado por": indicando que a entidade denotada por $\mathcal{A}$ é, de alguma forma, uma parte conceitual da entidade denotada por $\mathcal{T}$.
- Acomodação: pseudorrelação utilizada para categorizar relações que não puderam ser enquadradas nas outras relações.

3.2.2.1 Relação de correferência

Esta é a relação tradicional usada na resolução de anáforas pronominais, elipses e em algumas ANDs. Considere o exemplo:

"*Pedro não gostou da festa.*" (3.6a)

"*Ele nem provou o bolo.*" (3.6b)

"*θ Nem provou o bolo.*" (3.6c)

"*O chato nem provou o bolo.*" (3.6d)

A frase (3.6b) correferencia o antecedente *"Pedro"* da frase (3.6a) através do pronome *"Ele"*; na frase (3.6c) existe uma elipse, na qual ocorre a relação de correferência ao mesmo antecedente; a frase (3.6d) apresenta a anáfora nominal definida *"O chato"* também em uma situação de correferência com o antecedente *"Pedro"*. Quando um transmissor usa um SND em vez de usar um pronome, como acontece na frase (3.6d), ele está tentando enriquecer o conhecimento do receptor com mais informações sobre uma mesma entidade (*"Pedro"*).

No exemplo (3.7), tanto *"multidão"* quanto *"pessoas"* são entidades coletivas. Neste contexto, os dois termos são sinônimos entre si.

"*Uma multidão atacou a mesa para pegar os docinhos.* (3.7)
As pessoas brigavam pelo brigadeiro."

Freitas define em [Freitas 2005] uma proposta para identificação da relação de correferência para os casos expressos nos exemplos (3.6) e (3.7). Entretanto, no que diz respeito à regra proposta para o tratamento de entidades coletivas (exemplo (3.7)), houve um equívoco na lógica apresentada. Sendo assim, este trabalho propõe uma nova regra para este caso, de forma que o conjunto de regras para a identificação da relação de

correferência seja dado por:

a) *Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de um pronome ou de uma elipse, então $\mathcal{R}$ é uma relação de correferência.*

b) *Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de um SND e $\mathcal{A}$ e $\mathcal{T}$ concordam em número e gênero, então $\mathcal{R}$ pode ser uma relação de correferência.*

c) *Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de um SND e $\mathcal{A}$ e $\mathcal{T}$ são coletivos qualificados*[1]*, então $\mathcal{R}$ pode ser uma relação de correferência.*

3.2.2.2 Relação "membro de"

A relação "membro de" pode ser estabelecida entre indivíduos e conjuntos de indivíduos, como em:

"*Marina ganhou muitos presentes.*
O presente mais divertido foi a vuvuzela." (3.8)

O exemplo a seguir mostra a utilização da relação "membro de" quando se considera uma entidade coletiva:

"*Uma multidão atacou a mesa para pegar os docinhos.*
A primeira pessoa ficou com os melhores." (3.9)

A proposta para a identificação da relação "membro de" de Freitas ficou imprecisa em certos aspectos e, devido a isso, este trabalho sugere uma nova definição para a mesma, da seguinte forma:

> *Considere $\mathcal{A}$ e $\mathcal{T}$ como sendo, respectivamente, um SND e um dos seus possíveis antecedentes. Considere $T_{\mathcal{T}}$ como sendo o tipo de $\mathcal{T}$, o qual é determinado da seguinte forma: se $\mathcal{T}$ está no plural então $T_{\mathcal{T}}$ é um conjunto único formado pela cabeça linguística de $\mathcal{T}$ no singular; se $\mathcal{T}$ é uma entidade coletiva então $T_{\mathcal{T}}$ é o conjunto de sinônimos de $\mathcal{T}$. De forma análoga, considere $T_{\mathcal{A}}$ como sendo o tipo de $\mathcal{A}$, dado por: se $\mathcal{A}$ está no singular então $T_{\mathcal{A}}$ é formado pela cabeça linguística de $\mathcal{A}$. Se $T_{\mathcal{A}} \cap T_{\mathcal{T}} \neq \{\} \wedge (\textbf{\textit{plural}}(\mathcal{T}) \Rightarrow \textbf{\textit{genero}}(\mathcal{A}) = \textbf{\textit{genero}}(\mathcal{T}))$ então pode-se assumir uma relação de "membro de".*

[1] A identificação de uma entidade coletiva é obtida através de um dicionário de coletivos da língua portuguesa.

Considerando os tipos $T_{\mathcal{T}}$ e $T_{\mathcal{A}}$ descritos na definição acima e destrinchando a lógica utilizada, é possível reescrevê-la da seguinte forma:

> *Se* $(\textbf{\textit{plural}}(\mathcal{T}) \vee \textbf{\textit{coletivo}}(\mathcal{T})) \wedge (\textbf{\textit{singular}}(\mathcal{A})) \wedge (T_{\mathcal{A}} \cap T_{\mathcal{T}} \neq \{\}) \wedge (\neg\textbf{\textit{plural}}(\mathcal{T}) \vee \textbf{\textit{genero}}(\mathcal{A}) = \textbf{\textit{genero}}(\mathcal{T}))$ *então pode-se assumir uma relação de "membro de".*

3.2.2.3 Relação "parte de"

Esta relação é definida quando uma entidade é parte estrutural de outra. No exemplo (3.10), o SND *"O pavio"* é considerado como parte da *"vela"*.

> *"O aniversariante mal conseguiu assoprar a vela.*
> *O pavio era muito curto."* (3.10)

Em um outro contexto, como nas frases do exemplo (3.11), pode-se assumir que o *"brigadeiro"* é parte da *"cesta de doces"*.

> *"Cada convidado recebeu uma cesta de doces.*
> *O brigadeiro estava gostosão!"* (3.11)

A regra para a determinação da relação "parte de" foi proposta por Freitas da seguinte maneira:

a) *Se o antecedente $\mathcal{T}$ está no singular, $\mathcal{A}$ é a entidade introduzida por um SND, $\mathcal{A}$ não é uma entidade coletiva (determinada pelo plural ou por estar presente num dicionário de coletivos), então pode-se assumir a relação de "parte de".*

b) *A relação "parte de" somente será válida se não existir nada em contrário no contexto de interpretação.*

A anormalidade da relação é quando existe informação que impossibilite que um objeto seja parte de outro, como por exemplo: caso o tamanho de $\mathcal{A}$ seja maior que o tamanho de $\mathcal{T}$.

3.2.2.4 Relação "subcategorizado por"

A relação "subcategorizado por" ocorre quando uma entidade é parte conceitual da outra. É semelhante à relação "parte de", exceto pela necessidade de o antecedente

ser uma entidade autônoma e intencional – denominada **animacidade** [Sidner 1979, Dahl e Fraurud 1996, Orasan e Evans 2007]. Como exemplo, as frases (3.12) trazem *"O motorista"* como uma entidade subcategorizada por *"limusine"*.

> *"Uma limusine estacionou para buscar o aniversariante.*
> *O motorista estava muito bem vestido."* (3.12)

Segundo a proposta, uma relação de "subcategorizado por" é definida da seguinte forma:

> *Se o antecedente $\mathcal{T}$ está no singular, $\mathcal{A}$ é um SND animado e $\mathcal{T}$ não é um indivíduo coletivo, pode-se assumir a relação "subcategorizado por" entre $\mathcal{A}$ e $\mathcal{T}$.*

3.2.2.5 Pseudorrelação de acomodação

A acomodação surge quando todas as outras possibilidades de interpretação de um SND terminaram e nenhuma das relações anteriores pôde ser estabelecida. Assim, a regra é dada por:

> *Se $\mathcal{A}$ é um SND e não é possível estabelecer nenhuma das relações anteriores (correferência, "membro de", "parte de", "subcategorizado por") entre $\mathcal{A}$ e o antecedente $\mathcal{T}$, então assume-se a relação de acomodação.*

3.2.3 Focos do discurso e listas de entidades relevantes

Foco do discurso é o termo utilizado para designar a entidade mais saliente do discurso [Hajičová, Skoumalová e Sgall 1995, Freitas e Lopes 1994, Sidner 1981, Grosz 1977]. Tipicamente, o foco é a entidade sobre a qual o transmissor centra sua atenção em determinado ponto do discurso, sendo que a utilização continuada de uma determinada entidade através do uso de anáforas é um forte indício de que esta entidade está em foco [Grosz, Joshi e Weinstein 1995, Sidner 1981]. Para o tratamento das anáforas nominais definidas, consideram-se dois tipos de componentes salientes no discurso: o **foco implícito**, que consiste do assunto sobre o qual versa determinado conjunto de frases, e o **foco explícito**, que determina a entidade sobre a qual é centrada a atenção em cada frase.

Existe uma íntima ligação entre o uso de anáforas e os focos, mais especificamente:

- O foco explícito é resultante da utilização de anáforas pronominais, elipses e ANDs diretas (relação de correferência);

- O foco implícito é resultante da utilização de ANDs indiretas (demais relações).

Para exemplificar, considere o seguinte texto:

"*Marcelo ganhou uma cesta de salgadinhos.*" (3.13a)

"*A empada estava desmanchando.*" (3.13b)

"*Os croquetes estavam frios.*" (3.13c)

"*Ela estava muito gostosa.*" (3.13d)

Na frase (3.13a) é introduzida a entidade *"cesta de salgadinhos"*. Na frase (3.13b) essa entidade é referenciada através da AND *"A empada"*, de forma que o assunto do discurso continue a ser implicitamente a *"cesta de salgadinhos"*, mas *"A empada"* passa a ser a entidade mais saliente da frase. Em seguida, consideram-se duas possíveis continuações: as frases (3.13c) e (3.13d). Na frase (3.13c) é usada novamente um SND, indicando que existe uma referência ao assunto do discurso e não à frase anterior, logo *"Os croquetes"* está ligado ao foco implícito do discurso até o momento, *"cesta de salgadinhos"*, e não ao foco explícito *"A empada"* da frase anterior. Por outro lado, se a continuação fosse a frase (3.13d), a anáfora pronominal *"Ela"* faria referência direta ao foco explícito, que no momento é *"A empada"* da frase (3.13b).

Dado um discurso D constituído das frases $f_1, \cdots, f_i, \cdots, f_n$, considere o conjunto de referentes do discurso $Refs_i = [u_1, u_2, \cdots, u_n]$ introduzidos pela interpretação semântica da frase corrente f_i. Define-se **lista de entidades explícitas relevantes** (LR_i^{exp}) como a lista ordenada dos referentes $Refs_i$. Essa lista servirá, a priori, para a determinação dos antecedentes para as anáforas na interpretação da frase seguinte. Além disso, a entidade melhor classificada em LR_i^{exp} será o foco explícito da frase f_i, ou $foco_i^{exp}$.

Considera-se, também, a **lista de entidades implícitas relevantes** (LR_i^{imp}), composta somente pelas entidades referenciadas (antecedentes) por anáforas nominais definidas da frase f_i. A entidade melhor classificada em LR_i^{imp} será o foco implícito da frase f_i, ou $foco_i^{imp}$. Note que, para a interpretação da primeira frase do discurso f_1, não existe nada previamente interpretado, portanto a LR_1^{imp} é vazia. Logo, o foco implícito da primeira frase será nulo: $foco_1^{imp} = nulo$.

Para a ordenação da lista de entidades implícitas relevantes LR^{imp}, Freitas em

[Freitas 2005] propõe a seguinte regra[2]:

$$\text{sujeito} > \text{objeto direto} > \text{objeto indireto} \tag{3.14}$$

Já para a lista de entidades explícitas relevantes LR^{exp}, outros fatores devem ser observados, já que um referente do discurso pode ser anafórico ou não. Assim, Freitas propõe a seguinte regra para a ordenação de LR^{exp}:

$$\begin{array}{ccc} \text{entidades anafóricas} & > & \text{entidades não anafóricas} \\ \text{elipse} > \text{pronomes} > \text{SND} & & \text{sujeito} > \text{objeto direto} > \text{objeto indireto} \\ \text{sujeito} > \text{objeto direto} > \text{objeto indireto} & & \end{array} \tag{3.15}$$

3.3 Estrutura Nominal do Discurso

A Estrutura Nominal do Discurso (END) [Freitas 2005] é uma estrutura que surge a partir da interpretação de um documento, no qual elementos linguísticos que sugerem a utilização de anáforas, tais como pronomes, elipses e sintagmas nominais definidos, são identificados juntamente com os antecedentes candidatos ao estabelecimento de uma relação anafórica. Para a interpretação automática, o documento é analisado frase a frase em um processo denominado **interpretação fora de contexto**, no qual para cada frase é criada a representação semântica com base em uma DRS (*Discourse Representation Structure*) [Kamp e Reyle 1993, Freitas 1992, Freitas e Lopes 1993]. A representação obtida da frase é denominada **segmento** do texto. Caso a frase apresente predicados que sugerem uma anáfora, o segmento relativo à frase deve ser interpretado em relação à parcela do texto já interpretada em um processo denominado **interpretação em contexto** da frase.

Na interpretação em contexto o segmento criado é interpretado com base nos outros segmentos já interpretados na estrutura. Com isso, é possível identificar qual entidade é o seu antecedente, caso realmente exista uma anáfora na frase.

A Figura 3.1 apresenta as informações que são armazenadas em um segmento S_i interpretado. $foco_i^{exp}$ e $foco_i^{imp}$ representam, respectivamente, os focos explícito e implícito do segmento S_i, LR_i^{exp} e LR_i^{imp}, as listas de entidades relevantes explícita e implícita, respectivamente, e $Conds_i$, os predicados identificados no segmento S_i. O elemento $tipo_i$ armazena o tipo do segmento: segmentos básicos , que indica a relação que permite a

[2] A > B significa que A está melhor classificado na lista do que B.

interpretação do segmento junto à parcela já interpretada do texto.

S_i	$tipo_i$: básico
$foco_i^{exp}$ $foco_i^{imp}$	LR_i^{exp} LR_i^{imp}
$Conds_i$	

Figura 3.1: Representação de um segmento da END.

A teoria por trás da END baseia-se no acompanhamento dos focos do discurso (explícito e implícito) e no agrupamento de todo o material semântico existente na interpretação de uma frase, em uma estrutura em árvore onde somente os nós mais à direita estão abertos para interpretação [Polanyi, Berg e Ahn 2003, Polanyi 1988]. Os nós folhas, chamados **segmentos básicos**, encapsulam toda a informação semântica das frases, e os demais, nós internos da árvore denominados **segmentos compostos** ou simplesmente **segmentos**, são compostos de material semântico herdado de seus nós filhos. A Figura 3.2 mostra essa estruturação.

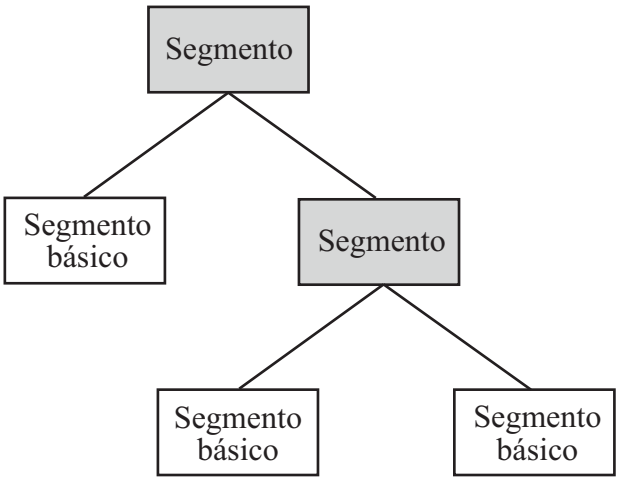


Figura 3.2: Segmentos da Estrutura Nominal do Discurso.

O procedimento para inserção de um novo segmento na estrutura primeiramente deve determinar qual é o **ponto de interpretação**, que será o segmento que possibilite a resolução do maior número de anáforas do novo segmento. Para isso, somente a raiz e os subsequentes filhos à direita são considerados na busca – chamados de **segmentos visíveis** –, iniciando do segmento de maior profundidade até atingir a raiz, como pode ser observado na Figura 3.3. Os elementos anafóricos do "Novo segmento" na figura serão resolvidos ao ser possível estabelecer relacionamentos (a partir das relações introduzidas pelas regras pragmáticas – seção 3.2.2) entre suas partículas anafóricas e os elementos

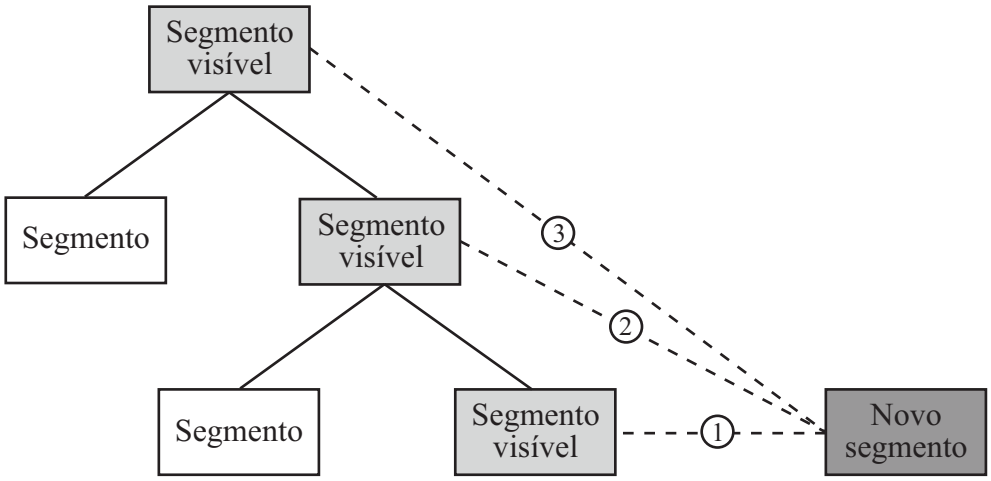


Figura 3.3: Ordem da interpretação de um segmento na END.

de um dos três segmentos visíveis apresentados na END em questão. O segmento visível que fornecer o maior número de resoluções será considerado como o segmento antecedente – onde está localizado o ponto de interpretação. Não sendo possível identificar relacionamentos entre as anáforas do "Novo segmento" e entidades dos segmentos visíveis da estrutura, a relação de acomodação será utilizada representando que o termo do segmento comporta-se como um indefinido.

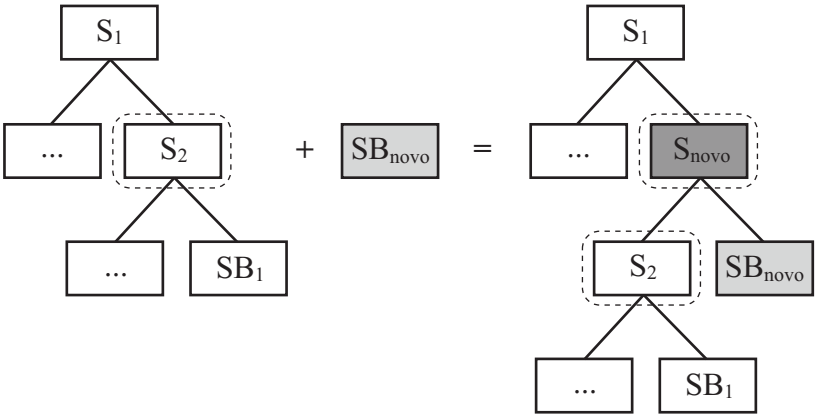


Figura 3.4: Composição de um novo segmento na END.

Uma vez localizado o ponto de interpretação, a Figura 3.4 mostra o processo de composição de um novo segmento à estrutura. Um novo segmento $\mathbf{S_{novo}}$ é criado e inserido no ponto de ancoragem, o qual herda atributos do segmento básico $\mathbf{SB_{novo}}$ resultante da interpretação fora de contexto (filho à direita) e do segmento antecedente $\mathbf{S_2}$ (filho à esquerda). Os atributos herdados constituem a única forma de inserção de material semântico em um segmento composto. Devido a esta herança, um segmento composto tem duas funções bem definidas: resumir as informações dos seus subsegmentos imediatos e servir de ponto de interpretação para as próximas frases. A determinação da herança de-

pende exclusivamente da comparação entre os focos de dois segmentos, sejam eles básicos ou não. O resultado desta comparação dá origem a quatro tipos de segmentos:

- elaboração – um segmento do tipo elaboração indica que o assunto de seus segmentos-filhos é o mesmo e há uma elaboração sobre uma mesma entidade do discurso (tópico) nestes segmentos;
- mudança de assunto – este segmento indica que o discurso passa a dissertar sobre um novo assunto que não tem ligação nenhuma com o tópico anterior;
- mudança de tópico – este segmento indica que o discurso disserta sobre um tópico (entidade) diferente de um mesmo assunto;
- manutenção de tópico – neste segmento houve uma mudança de assunto, porém algumas entidades do assunto anterior continuarão a ser referenciadas no novo assunto.

Esses tipos de segmentos são obtidos aplicando as regras da Tabela 3.1.

	$foco_1^{exp} = foco_2^{exp}$	$foco_1^{exp} \neq foco_2^{exp}$
$foco_1^{imp} = foco_2^{imp}$	**elaboração**	**mudança de tópico**
$foco_1^{imp} \neq foco_2^{imp}$	**manutenção de tópico**	**mudança de assunto**

Tabela 3.1: Relações entre $foco^{imp}$, $foco^{exp}$ e o tipo de segmento gerado.

Tendo definido o tipo do segmento composto recém-criado, seus atributos devem ser ajustados, considerando o material herdado do filho mais à esquerda S_1 (segmento antecedente) e do filho mais à direita S_2 (segmento anafórico). Para um melhor entendimento desses valores, a seguir é apresentada uma análise da herança de cada atributo, considerando os quatro tipos de segmentos compostos. Os índices 1 e 2 utilizados nos atributos são relativos aos segmentos S_1 e S_2, respectivamente.

- **Elaboração**:
  - $foco^{imp}$: focos implícitos iguais nos subsegmentos indicam que o transmissor está falando sobre um mesmo assunto e, portanto, a entidade implícita mais saliente é o foco implícito (não nulo) comum.
    $\therefore\ foco^{imp} \leftarrow foco_1^{imp}$, se $foco_1^{imp} \neq$ nulo, ou $foco_2^{imp}$, caso contrário
  - LR^{imp}: com a continuação do assunto expressa pela continuação dos focos implícitos dos subsegmentos, a lista de entidades implícitas terá o $foco^{imp}$ à

cabeça.
$\therefore\ LR^{imp} \leftarrow [foco^{imp}]$

- $foco^{exp}$: focos explícitos iguais nos subsegmentos indicam que o transmissor está falando sobre uma mesma entidade, sendo o foco explícito, portanto, esta entidade em comum, representando um resumo do tópico de seus subsegmentos.
  $\therefore\ foco^{exp} \leftarrow foco_1^{exp}$
- LR^{exp}: a continuação do tópico indica que o $foco^{exp}$ é o resumo, não havendo necessidade da herança direta ou da combinação dos outros elementos das LRs. O resultado é uma lista de entidades explícitas com somente um elemento.
  $\therefore\ LR^{exp} \leftarrow [foco^{exp}]$

- **Mudança de assunto**:
  - $foco^{imp}$: focos implícitos diferentes nos subsegmentos indicam que o transmissor mudou de assunto, caso em que o assunto antigo deve ser arquivado. O resultado é que o segmento antigo não deve ser esquecido, mas sim apenas a sua subárvore.
    $\therefore\ foco^{imp} \leftarrow foco_1^{imp}$
  - LR^{imp}: visando dar uma maior amplitude à retomada de assunto posteriormente, a lista de relevantes implícita deve conter todos os elementos do segmento antigo.
    $\therefore\ LR^{imp} \leftarrow LR_1^{imp}$
  - $foco^{exp}$: focos explícitos diferentes nos subsegmentos indicam que o transmissor mudou de tópico, além de já ter mudado de assunto. O segmento composto deve então herdar seus atributos do segmento mais antigo, permitindo uma futura retomada do tópico.
    $\therefore\ foco^{exp} \leftarrow foco_1^{exp}$
  - LR^{exp}: utilizando o mesmo critério aplicado ao foco explícito, implica que a lista de relevantes explícita também deve ser herdada do segmento mais antigo.
    $\therefore\ LR^{exp} \leftarrow LR_1^{exp}$

- **Mudança de tópico**:
  - $foco^{imp}$: focos implícitos iguais indicam continuação do assunto, logo o segmento composto herda o foco em comum (diferente de nulo).
    $\therefore\ foco^{imp} \leftarrow foco_1^{imp}$

- LR^{imp}: como o assunto centra-se sobre uma determinada entidade e esta já é o foco implícito, logo não há necessidade da heranças dos elementos da LR implícita, à exceção do foco implícito.
  $\therefore\ LR^{imp} \leftarrow [foco_1^{imp}]$
- $foco^{exp}$: diferentes focos explícitos significam que houve uma mudança de entidades no foco local. Para que estas entidades possam ser reutilizadas nas próximas interpretações, deve-se herdar as entidades explícitas.
  $\therefore\ foco^{exp} \leftarrow foco_1^{exp}$
- LR^{exp}: idem ao foco explícito.
  $\therefore\ LR^{exp} \leftarrow LR_1^{exp}$

- **Manutenção de tópico**:
  - $foco^{imp}$: será herdado o foco implícito do segmento mais antigo.
    $\therefore\ foco^{imp} \leftarrow foco_1^{imp}$
  - LR^{imp}: visando a dar uma maior amplitude à retomada de assunto posteriormente, a lista de relevantes implícita deve conter todos os elementos do segmento antigo.
    $\therefore\ LR^{imp} \leftarrow LR_1^{imp}$
  - $foco^{exp}$: o segmento composto não necessitará das informações das entidades explícitas, apenas herdará o assunto anterior possibilitando futuras referências via anáforas.
    $\therefore\ foco^{exp} \leftarrow \perp$
  - LR^{exp}: idem ao foco explícito.
    $\therefore\ LR^{exp} \leftarrow \perp$

A Tabela 3.2 apresenta um resumo dos valores dos atributos do novo segmento composto gerado, considerando cada tipo de segmento.

A Estrutura Nominal do Discurso é criada com a finalidade específica de resolver anáforas. No entanto, ela apresenta características que podem ser úteis para um sistema de categorização de textos. Primeiramente, para sua criação são excluídas as classes gramaticais que são consideradas sem informação semântica para o objetivo de resolução de anáforas e, analogamente, para categorização, como artigos, preposições, conjunções, verbos, entre outras. Além disso, a própria característica principal da END, ou seja, a identificação das múltiplas referências a uma mesma entidade, introduz ganho de informação para a categorização, uma vez que torna possível saber o número de vezes que

	$\boldsymbol{foco^{imp}}$	$\boldsymbol{LR^{imp}}$	$\boldsymbol{foco^{exp}}$	$\boldsymbol{LR^{exp}}$
Elaboração	$foco_2^{imp}$	$[foco^{imp}]$	$foco_1^{exp}$	$[foco^{exp}]$
Mudança de assunto	$foco_1^{imp}$	LR_1^{imp}	$foco_1^{exp}$	LR_1^{exp}
Mudança de tópico	$foco_1^{imp}$	$[foco_1^{imp}]$	$foco_1^{exp}$	LR_1^{exp}
Manutenção de tópico	$foco_1^{imp}$	LR_1^{imp}	$\perp$	$\perp$

Tabela 3.2: Valores de $foco^{imp}$, LR^{imp}, $foco^{exp}$ e LR^{exp} para cada tipo de segmento.

uma entidade é realmente referenciada, de forma que a relevância no documento onde ela ocorre é melhor estimada.

Nas próximas seções deste capítulo será abordada a proposta para categorização, tendo como base a Estrutura Nominal do Discurso.

3.4 Proposta de categorização

Este trabalho apresenta uma proposta de categorização de textos, sem a necessidade de um conjunto de documentos previamente rotulados ou da atuação de um especialista humano para a realização da tarefa de rotulação – altamente custosa, como já foi discutido. Partindo somente de documentos não rotulados, categorias pré-definidas e palavras relacionadas às mesmas, o categorizador proposto apresenta dois passos gerais: (1) a utilização da técnica de *bootstrapping*, associando, automaticamente, rótulos aos documentos, e (2) incorporação do conjunto de documentos rotulados a um categorizador supervisionado de respaldo na literatura.

Para o passo 1 supracitado, considera-se como representação de cada documento a Estrutura Nominal do Discurso. A partir dela, é definido um conjunto de termos para caracterizar cada categoria, de forma a permitir a aplicação de medidas de similaridade para a associação dos documentos às categorias.

Já em relação ao passo 2 da proposta de categorização, não é o foco deste trabalho avaliar o melhor categorizador supervisionado a ser utilizado, muito menos desenvolver um novo método. O objetivo é avaliar a utilização da rotulação retornada pelo processo de *bootstrapping* em diferentes categorizadores amplamente conhecidos na literatura. Além disso, são feitas manipulações nos conteúdos dos documentos, utilizando informações obtidas pelo processo de resolusão de anáforas durante a criação da END. Os documentos originais são substituídos por diferentes variações de seu conteúdo, para posteriormente

serem considerados nas fases de treino e teste do categorizador.

As seções 3.5 e 3.6 apresentam os detalhes dos passos 1 e 2, respectivamente.

3.5 Método de *bootstrapping*

O termo *bootstrapping* está relacionado com o seguinte conceito geral: "promover ou desenvolver pela iniciativa e esforço com pouca ou nenhuma assistência" [Adami, Avesani e Sona 2003]. Voltado para este trabalho específico de categorização, seu conceito pode ser definido como: dado um conjunto de categorias/rótulos e um conjunto de documentos não rotulados, *bootstrapping* é a tarefa de estipular rótulos para esses documentos, de forma que essa rotulação sirva como entrada em um categorizador supervisionado.

O processo de *bootstrapping* proposto neste trabalho, de forma geral, introduz uma métrica de similaridade entre categorias e documentos. Baseado no valor obtido por essa métrica, cada documento é associado à categoria devida. Para que esse relacionamento seja eficaz e a métrica de similaridade apresente valores satisfatórios, é necessário que ela seja alimentada com o máximo de informação possível sobre cada categoria. Nessa etapa, quanto mais bem definidas as categorias estejam, melhor será o resultado da rotulação.

Para definir uma categoria, um texto em linguagem natural não seria ideal, já que muitos termos sem valor semântico seriam levados em consideração desnecessariamente. Este trabalho utiliza para a definição das categorias as aqui chamadas **palavras características** – termos relacionados às categorias, independentes entre si, que possuem sentido próprio. Para exemplificar a vantagem da utilização de um conjunto de palavras características para cada categoria, considere a seguinte situação: um documento que apresenta um conteúdo sobre a última geração de telefones celulares está visivelmente relacionado a uma categoria de título "Tecnologia"; porém, caso o termo "tecnologia" não seja citado no documento, considerando somente o título da categoria, esse relacionamento torna-se impraticável. Por outro lado, caso o termo "telefonia" esteja presente na lista de palavras características da categoria "Tecnologia", e também ocorra no documento em questão (o que é bem provável), o relacionamento entre o documento e a categoria pode ser facilmente detectado. Os próximos passos, portanto, visam a obtenção de palavras características bem definidas para cada categoria.

Inicialmente, no processo de *bootstrapping*, só estão disponíveis: os documentos a serem rotulados e os nomes das categorias. Há bases de dados que fornecem informações

suplementares sobre cada categoria, com maiores detalhes e/ou palavras relacionadas. Nesse caso, esses termos poderiam ser levados em consideração no processo, facilitando a geração das palavras características das categorias. Contudo, o objetivo deste trabalho é tratar o problema de uma forma mais extensiva, sem fazer quaisquer limitações sobre o conjunto de dados a ser considerado.

Sem a disponibilização de informações extras sobre as categorias, em alguns casos pode ser impraticável definir relacionamentos somente a partir dos títulos das mesmas. Os nomes podem ser vagos ou abrangentes demais, dificultando a identificação do assunto do qual trata a categoria em questão. Considerando como exemplo a coleção CHAVES que será utilizada para testes no capítulo 5, podem-se destacar as seguintes categorias de sentido impreciso: "Cotidiano", "Mais!" e "Opinião". Um texto que disserte sobre o cotidiano de um jogador de futebol, por exemplo, mesmo apresentando por várias vezes em seu conteúdo o termo "cotidiano", deve estar associado à categoria "Esporte" (também presente na coleção CHAVES) e não à categoria "Cotidiano" (que possivelmente diz respeito a notícias gerais de interesse). Exemplos similares podem ser facilmente imaginados para as outras duas categorias citadas.

Sendo assim, o sistema desenvolvido neste trabalho permite que sejam previamente fornecidas **palavras relacionadas** às categorias. Mesmo assumindo a obrigatoriedade da atuação humana de um conhecedor da coleção para definir um certo número de palavras relacionadas para cada categoria, este esforço ainda continua mínimo em relação ao custo de rotular manualmente toda a coleção. O operador do sistema pode definir a quantidade de palavras para cada categoria que julgar necessário.

Como se trata de dados obtidos de forma manual, essas palavras relacionadas normalmente estarão em um número pequeno e podem ser insuficientes para a caracterização definitiva da categoria. O próximo passo proposto neste trabalho é ampliar esse conjunto. Para isso, foi considerado um *thesaurus* ou dicionário de sinônimos [Gomes 1990] da seguinte forma: para cada categoria, são adicionados os sinônimos das palavras relacionadas fornecidas pelo especialista humano, formando o **conjunto expandido de palavras relacionadas**.

Essa primeira expansão do conjunto de palavras relacionadas, utilizando um dicionário de sinônimos, baseou-se somente no sentido semântico dos termos relacionados. A próxima etapa proposta é inserir informação que apresente relação direta com o conjunto de documentos considerado, visando uma representação mais específica. A partir de padrões de coocorrência, são extraídos termos dos documentos que apresentam maior similaridade

com a categoria considerada, formando assim o **conjunto de palavras-chave**. A seção 3.5.1 detalha como esse conjunto é obtido.

Por fim, o processo de geração das palavras características das categorias pode ser resumido da seguinte forma:

1. Definir palavras relacionadas para cada categoria, a partir da atuação de um especialista humano.
2. Adicionar os sinônimos dessas palavras relacionadas, utilizando um dicionário de sinônimos – formando o conjunto expandido de palavras relacionadas.
3. Utilizar informações de coocorrência entre os termos do conjunto expandido e os termos contidos nos documentos, para gerar o conjunto de palavras-chave de cada categoria.
4. A união desses três conjuntos forma o **conjunto de palavras características** de cada categoria.

A seção 3.5.1 mostra como o passo 3 é obtido. Com isso, o conjunto de palavras-chave para cada categoria está definido e, na seção 3.5.2, é mostrado por fim o processo de relacionamento dos documentos com as categorias.

3.5.1 Definição das palavras-chave de cada categoria

Ko e Seo, em [Ko e Seo 2009], definiram um método para a determinação automática de um conjunto de palavras-chave semanticamente relacionadas aos nomes de cada categoria. Isso é feito utilizando informação de coocorrência entre os nomes das categorias e os termos contidos dos documentos: se o grau de similaridade semântica entre o nome de uma categoria e um dado termo for suficientemente alto, este termo é adicionado à lista de palavras-chave da categoria. O método de Ko e Seo foi adaptado ao problema em questão, considerando a representação dos documentos através das ENDs.

O valor de similaridade semântica (sim) entre um termo t e uma categoria c_j é calculado pela seguinte métrica:

$$sim(t, c_j) = \frac{\sum_{p \in P_j} \left(\frac{\sum_{i=1}^{n} sim\_doc(t, p, d_i)}{n} \right)}{|P_j|}, \tag{3.16}$$

onde P_j representa o conjunto de palavras características da categoria c_j – neste momento representado somente pelo conjunto expandido de palavras relacionadas –, n representa o número total de documentos e $sim\_doc(t, p, d_i)$ retorna um valor de similaridade semântica considerando a coocorrência dos termos t e p no documento d_i. Esse valor é dado pela seguinte fórmula:

$$sim\_doc(t, p, d_i) = w_{ti} \cdot w_{pi} \cdot (1 - |w_{ti} - w_{pi}|), \tag{3.17}$$

sendo w_{xi} o peso do termo x no documento d_i, definido como se segue:

$$w_{xi} \quad = \quad 0.6 \cdot f_{xi} + 0.4 \cdot g_{xi}, \tag{3.18}$$

onde f_{xi} representa a taxa de ocorrência do termo x em focos (implícito ou explícito) da END relativa ao documento d_i, definido da seguinte forma: $f_{xi} = tf_f(x, d_i)/m$, onde tf_f é a frequência do termo em questão (*term frequency*) em focos da END, e m é o número de segmentos da END. g_{xi} é a taxa de ocorrência do termo x em listas de entidades relevantes (implícitas ou explícitas) do documento d_i: $g_{xi} = tf_g(x, d_i)/m$, sendo tf_g a frequência do termo em listas de relevantes.

A equação (3.18) fornece uma métrica para a importância do termo x no documento, assumindo um fator de 0.6 para a ocorrência em focos e 0.4, em listas de relevantes (valores obtidos experimentalmente), dando maior relevância para a ocorrência em focos da END.

Voltando à análise da função de similaridade que considera a coocorrência de dois termos em um dado documento – equação (3.17) –, observam-se dois fatores em sua formulação:

1. a multiplicação entre os valores w_{ti} e w_{pi}: quanto maior o resultado, maior a similaridade entre os termos t e p;

2. o módulo da diferença entre w_{ti} e w_{pi} subtraído de um: quanto menor for o valor em módulo da diferença, mais similares os termos são; para expressar isso, como w_{ti} e w_{pi} são valores entre 0 e 1, a diferença foi subtraída de 1.

Em relação ao primeiro fator, vale ressaltar que, caso w_{ti} ou w_{pi} (pelo menos um) sejam nulos, o valor de similaridade entre esses dois termos será zero. Caso w_{ti} seja zero, por exemplo, significa que o termo t não ocorre no documento d_i, o que implica em uma coocorrência nula com o termo p nesse mesmo documento, independente do valor w_{pi}. Essa situação é tratada na formulação com a multiplicação entre esses pesos, fornecendo:

$sim\_doc(t, p, d_i) = 0$.

Para justificar o segundo fator, considere o seguinte exemplo. Em um dado documento, há uma ocorrência grande do termo *"jogador"* e uma pequena aparição do termo *"entrevista"*. Como ambos os termos estão presentes no documento em questão, haverá, obviamente, um valor de coocorrência entre eles. Por outro lado, nesse mesmo documento, suponha que os termos *"técnico"* e *"clube"* apresentam ocorrências relativamente grandes e de valores próximos. Por terem valores próximos, estes termos provavelmente são mais similares entre si do que os dois primeiros termos. A formulação proposta trata essa situação através do fator que utiliza o módulo da diferença entre os pesos.

Visando afirmar essa situação, considere o mesmo exemplo em números. Suponha que os pesos dos dois primeiros termos (t_1:*"jogador"* e p_1:*"entrevista"*) sejam dados por: $w_{t_1 i} = 0.5$ e $w_{p_1 i} = 0.125$. O valor da função sim_doc nesse caso é dado por: $sim\_doc_1 = 0.0625 \cdot (1 - 0.375) = 0.039$. Considere agora que os pesos dos dois segundos termos (t_2:*"técnico"* e p_2:*"clube"*) sejam iguais: $w_{t_2 i} = w_{p_2 i} = 0.25$. O valor obtido pela multiplicação é exatamente o mesmo do primeiro caso, mas o módulo da diferença é nulo, fornecendo o seguinte cálculo para sim_doc: $sim\_doc_2 = 0.0625 \cdot (1 - 0) = 0.063$. O valor obtido foi maior do que na primeira situação, o que significa que os termos *"técnico"* e *"clube"* são mais similares entre si.

A equação (3.16) fornece o critério para a seleção de boas palavras-chave para as categorias. Todavia, há um outro ponto que deve ser observado, como foi ressaltado por Ko e Seo em [Ko e Seo 2009]. Caso um termo apresente um alto valor de similaridade com duas ou mais categorias, este termo não deve ser considerado como palavra-chave, já que ele não possui o poder discriminativo entre essas categorias. Visando sanar esta questão, cada termo é inicialmente selecionado como um candidato a palavra-chave da categoria que apresente o maior valor de similaridade. Em seguida, o peso (*score*) de cada termo é recalculado de acordo com a seguinte fórmula:

$$score(t, c_{max}) = sim(t, c_{max}) + \big(sim(t, c_{max}) - sim(t, c_{secondmax})\big), \tag{3.19}$$

onde c_{max} é a categoria com maior valor de similaridade com o termo t, e $c_{secondmax}$ é a categoria com o segundo maior valor de similaridade com o termo t.

Os termos candidatos a palavras-chave de cada categoria são ordenados de forma decrescente de acordo com o valor de *score*, dado pela fórmula (3.19). Por fim, são escolhidos como palavras-chave de cada categoria os k primeiros termos nessa ordenação.

Em outras palavras, para cada categoria são selecionados os termos que apresentam maior valor de *score* como palavras-chave.

A fórmula (3.19), portanto, significa o seguinte: um termo apresenta um valor de *score* maior se ele possui uma alta similaridade com a categoria e uma grande diferença de similaridade com as demais categorias.

Assim, cada categoria está representada por uma lista de palavras-chave, além dos termos já determinados que estão presentes no conjunto expandido de palavras relacionadas – formando o conjunto de palavras características da categoria. Tendo isso disponível, a seguir é mostrado como é feito o relacionamento entre documentos e categorias.

3.5.2 Cálculo da similaridade entre documentos e categorias

Em [Pereira 2009], Pereira propõe um método de recuperação de informação, modificado a partir da proposta de [Júnior 2007], considerando a Estrutura Nominal do Discurso. Para o cálculo de relevância entre uma consulta q (*query*) e um documento d, Pereira et al. [Pereira, Júnior e Freitas 2009] propôs uma modificação em relação ao método original de Seibel Júnior [Júnior 2007], obtendo a seguinte fórmula:

$$VR(q,d) = \sum_{i=0}^{n} \sum_{v=0}^{m} \frac{0.6 \cdot f(i,v,q) + 0.4 \cdot g(i,v,q)}{v+1}, \tag{3.20}$$

onde f é uma função que verifica se o termo q ocorre como foco implícito ou explícito do segmento $[i,v]$. Caso o termo apareça em uma dessas funções no segmento ela retorna o valor um, caso contrário ela retorna zero. v representa a profundidade do segmento em teste no momento. A função g retorna um número entre zero e a função $\frac{o^{exp}}{n^{exp}} + \frac{o^{imp}}{n^{imp}}$. Zero é retornado caso não tenha sido encontrada uma ocorrência do termo q no segmento $[i,v]$. Caso o termo q seja encontrado no segmento $[i,v]$ da estrutura, é retornado um valor com base na função $\frac{o^{exp}}{n^{exp}} + \frac{o^{imp}}{n^{imp}}$, onde o^{exp} e o^{imp} representam, respectivamente, o número de ocorrências do termo q na lista de entidades relevantes explícitas e o número de ocorrências de q na lista de entidades relevantes implícitas no segmento $[i,v]$. As variáveis n^{exp} e n^{imp} representam, respectivamente, o número de entidades explícitas do segmento e o número de entidades implícitas do segmento.

A fórmula (3.20), como já mencionado, foi desenvolvida para propósitos de recuperação de informação, visando determinar o valor de relevância de uma consulta – formada por um conjunto de termos – em um dado documento, tendo apresentado bons resultados

nessa tarefa. Limitando a consulta a um termo único, essa métrica pode ser tranquilamente utilizada para o cálculo de relevância de uma entidade genérica em um documento.

O objetivo desta seção é determinar uma forma de realizar a associação dos documentos às categorias, através de uma medida de similaridade entre eles. As informações disponíveis sobre as categorias são dadas pelo conjunto de palavras características, cuja obtenção foi tratada na seção anterior. Logo, o cálculo de similaridade entre um documento d_i e uma categoria c_j pode ser visto como a similaridade entre d_i e o conjunto de palavras características relativo a c_j. Especificando ainda mais, esse cálculo pode ser obtido pela média das similaridades entre d_i e cada entidade presente no conjunto de palavras características de c_j. Sendo assim, como medida de similaridade entre um documento e um termo, este trabalho considera a fórmula de valor de relevância, dada pela equação (3.20).

A similaridade, portanto, entre um documento d_i e uma categoria c_j, utilizando a fórmula (3.20), é definida da seguinte maneira:

$$simVR(d_i, c_j) = \frac{\sum_{p \in P_j} VR(p, d_i)}{|P_j|}, \tag{3.21}$$

onde P_j é o conjunto de palavras características da categoria c_j.

Assim, cada documento d_i possui um valor de $simVR$ relativo a cada categoria c_j. A categoria que fornecer o maior valor de $simVR$ com d_i será a categoria escolhida.

Com isso, o processo de *bootstrapping* está concluído, tendo gerado uma rotulação para cada documento, baseado nas categorias de interesse. O modelo de categorização pode ser então obtido, através da aplicação dessa rotulação em um categorizador supervisionado. A próxima seção apresenta como isso é feito.

3.6 Aplicação da rotulação em categorização supervisionada

Na seção anterior foi proposto um método de *bootstrapping* para a obtenção de documentos rotulados em categorias estabelecidas, visando a utilização desse corpus préclassificado em um categorizador supervisionado. O método foi baseado em uma estrutura desenvolvida para o problema de resolução de anáforas. Não cabe a este trabalho desenvolver uma abordagem supervisionada apoiada nesses conceitos, nem buscar melho-

rar abordagens já existentes. Serão utilizadas técnicas de categorização já fundamentadas e bem conceituadas na literatura.

Apesar de todo o trabalho de rotulação inicial dos documentos ter sido construído a partir da Estrutura Nominal do Discurso, a aplicação imediata dessa rotulação em um categorizador supervisionado – utilizando os documentos originais sem nenhum processamento – não se beneficia da resolução de anáforas em si. A proposta de Freitas [Freitas 2005] serve como base para o método de *bootstrapping*, mas não interfere no processo de obtenção do categorizador supervisionado. No entanto, ainda há informações e características intrínsecas da estrutura desenvolvida com o objetivo de resolver anáforas que podem ser utilizadas nesse processo final. Esta seção trata de como isso pode ser alcançado.

Para a construção do categorizador supervisionado, o corpus inicial de documentos pré-classificados é dividido em dois conjuntos, não necessariamente de mesmo tamanho: conjunto de treino T_r e conjunto de teste T_e. A partir do conjunto de treino, o categorizador é indutivamente construído observando as características dos documentos e, a partir do conjunto de teste, ele tem sua eficácia testada. A medida de eficácia da classificação é baseada no número de acertos obtidos.

Os textos dos documentos não são diretamente interpretados pelo classificador – eles devem ser inicialmente indexados, ou seja, mapeados para uma representação compacta do seu conteúdo. Normalmente, um documento d_j é representado por um vetor de pesos $\vec{d_j} = \langle w_{1j}, \ldots, w_{mj} \rangle$, onde m é o número de termos que ocorrem na coleção de documentos e $0 \leq w_{kj} \leq 1$ representa a importância (genericamente falando) do termo t_k para o documento d_j [Sebastiani 2002][3].

Considere o seguinte trecho da obra literária de José de Alencar:

> "*O* ***guerreiro*** *cristão atravessou a* ***cabana*** *e sumiu-****se*** *na* ***treva****. (...)*
> *Quando* ***ele*** *transmontou o* ***vale*** *e ia penetrar na* ***mata****,*
> *surgiu um* ***vulto*** *de* ***Iracema****.*
> *A* ***virgem*** *seguira o* ***estrangeiro*** *como a* ***brisa*** *sutil*
> *que resvala sem murmurejar por entre a* ***ramagem****.*"
>
> (Iracema, José de Alencar) (3.22)

[3]Para o objetivo deste exemplo será considerada a representação vetorial dos documentos devido à sua simplicidade, visando um melhor entendimento da explicação proposta. Entretanto, há uma série de categorizadores que indexam os documentos de forma distinta, como pôde ser visto no capítulo 2.

Uma pessoa que realize a leitura do trecho acima entende facilmente que há duas entidades em destaque: *"guerreiro"* e *"Iracema"*, pois ambas são referenciadas ao longo do texto, a patir de pronomes e sintagmas nominais definidos (sublinhados em (3.22)). Apesar de esses termos em destaque só ocorrerem uma única vez em todo o texto, um leitor humano é capaz de identificar esse relacionamento intuitivamente.

Suponha que o trecho (3.22) seja o texto relativo a um documento d_1. As entidades em negrito são as que apresentam maior valor semântico (substantivos e pronomes[4]) e, portanto, somente elas são consideradas na indexação. Sendo assim, levando em conta somente a frequência no texto – que é a base para muitas medidas de cálculo de peso –, a representação do documento d_1 seria data através do seguinte vetor: $\vec{d_1} = \langle 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 \rangle$, onde cada posição denota o peso relativo aos termos em negrito em (3.22), na ordem de aparecimento. Com esse vetor, fica difícil identificar a entidade em destaque no texto, uma vez que todas apresentam o mesmo valor de importância. Mesmo empregando medidas mais sofisticadas para o cálculo dos pesos, não é possível estabelecer, por exemplo, que o termo *"guerreiro"* é o mais relevante do texto sem a aplicação de técnicas que utilizam o conceito de anáforas.

Seguindo essa ideia, considere agora que o texto (3.22), antes de ser indexado, teve suas anáforas resolvidas[5]. O pronome reflexivo *"se"*, o pronome pessoal *"ele"* e o SND *"o estrangeiro"* estão referenciando a mesma entidade *"guerreiro"*. Da mesma forma, o SND *"A virgem"* referencia o substantivo próprio *"Iracema"*. Com isso, o texto (3.22) poderia ser reescrito da seguinte forma:

"*O* ***guerreiro*** *cristão atravessou a* ***cabana*** *e*
o ***guerreiro*** *sumiu na* ***treva***.
Quando o ***guerreiro*** *transmontou o* ***vale*** *e ia penetrar na* ***mata***,
surgiu um ***vulto*** *de* ***Iracema***.
Iracema *seguira o* ***guerreiro*** *como a* ***brisa*** *sutil*
que resvala sem murmurejar por entre a ***ramagem***." (3.23)

Considerando, da mesma forma, os termos em negrito no texto (3.23) em sua ordem de aparecimento, o vetor de pesos do documento d_1 passaria a ser representado por: $\vec{d_1'} = \langle 4, 1, 1, 1, 1, 1, 2, 1, 1 \rangle$, onde os valores das posições 1 e 7 se referem aos termos

[4]Em [Freitas 2005], Freitas leva em consideração a utilização de elementos verbais, além dos elementos nominais. Entretanto, os nomes são mais significativos para a tarefa de categorização, por isso, visando simplificar, no exempo só serão considerados os elementos nominais.

[5]As anáforas introduzidas por elipses não estão sendo consideradas neste exemplo.

"guerreiro" e *"Iracema"*, respectivamente. Claramente, em relação ao vetor $\vec{d_1}$, este novo vetor $\vec{d_1'}$ representa melhor o conteúdo do texto em questão, fornecendo um peso maior para os termos mais relevantes no contexto. Um categorizador que se baseie no texto (3.23), ao invés de utilizar o texto (3.22), independente do método escolhido, terá uma representação melhor do documento, tendendo, portanto, a uma classificação mais precisa.

Em [Mitkov et al. 2007], é mostrado como um sistema desenvolvido para resolução de anáforas pronominais pode ser utilizado para melhorar a performance de três aplicações que envolvem linguagem natural, incluindo categorização de textos. Todos os documentos do corpus considerado tiveram suas anáforas pronominais substituídas pelos sintagmas nominais reconhecidos como seus antecedentes pelo sistema. A partir desses novos documentos, foram testadas quatro abordagens de categorização supervisionada, obtendo resultados superiores em relação à utilização dos documentos originais.

No ponto atual deste trabalho, com a rotulação dos documentos já obtida pelo método de *bootstrapping* (seção 3.5), propõe-se a utilização do mesmo processo apresentado por Mitkov et al., à exceção de que, além de anáforas pronominais, também são consideradas anáforas nominais definidas[6]. Assim, os documentos rotulados são passados para o categorizador supervisionado observando as seguintes características principais: (1) só são indexados termos considerados com alto valor semântico – substantivos, que introduzem semântica ao texto, e pronomes, que fazem referência a termos já apresentados –, e (2) as anáforas existentes são resolvidas de acordo com a proposta de Freitas [Freitas 2005].

De maneira geral, o processo proposto neste trabalho pode ser resumido na Figura 3.5.

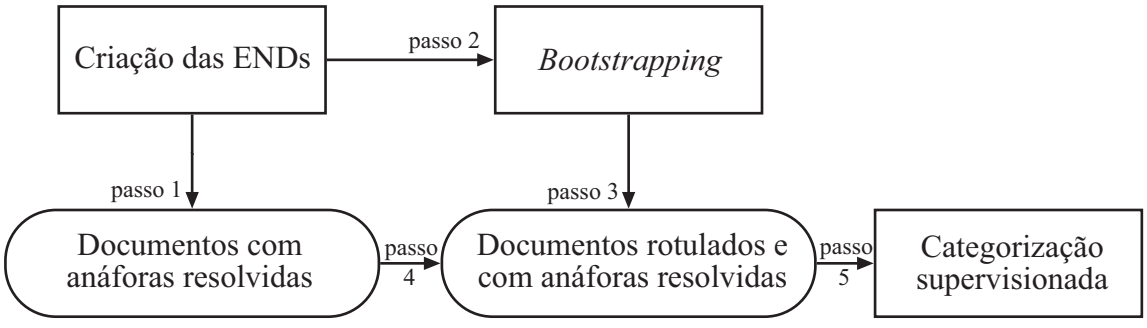


Figura 3.5: Processo geral da proposta deste trabalho.

Durante a criação das estruturas nominais do discurso, são obtidos os documentos, de forma que eles apresentem as duas características citadas acima (passo 1 na Figura 3.5).

[6]A proposta inicial de Freitas [Freitas 2005] abrange a resolução de anáforas pronominais, anáforas nominais definidas e anáforas introduzidas por elipse. Todavia, neste trabalho as elipses foram desconsideradas, pois sua identificação e seu tratamento são muito complexos e fogem do escopo desta dissertação.

Uma vez com as ENDs construídas, o processo de *bootstrapping* é executado (passo 2), gerando uma rotulação para os documentos. Esta rotulação é aplicada aos documentos que possuem suas anáforas resolvidas (passos 3 e 4) que, por fim, é passado como entrada para o categorizador supervisionado (passo 5).

No próximo capítulo serão mostrados os algoritmos implementados para a obtenção do objetivo proposto.

4 Algoritmo e implementação

Neste capítulo são detalhados os algoritmos desenvolvidos para a implementação do protótipo de categorização.

4.1 Introdução

Neste capítulo são apresentados os algoritmos utilizados no método de categorização proposto neste trabalho. A metodologia é baseada na teoria de Freitas [Freitas 2005] para resolução de anáforas, que utiliza regras pragmáticas para a identificação dos antecedentes de anáforas pronominais, anáforas nominais definidas e elipses, através da criação da Estrutura Nominal do Discurso. Em [Júnior 2007], o autor propõe uma adaptação da END para o problema de recuperação de informação. Dando continuidade a esse trabalho, em [Pereira, Júnior e Freitas 2009] é apresentada uma nova metododologia para RI baseada na resolução de anáforas, a qual é detalhada em [Pereira 2009]. Para tornar possível essa abordagem em uma linguagem de programação que não utilize o paradigma lógico, foi necessário definir uma série de considerações relativas à END proposta inicialmente, já que são observados os seguintes fatores: (1) um sistema de recuperação de informação deve ser computacionalmente eficiente, por isso houve a necessidade de limitar o processamento a certos tipos de anáforas e de armazenar somente informações mais relevantes na estrutura; (2) como o foco é recuperar informação, não existe a obrigatoriedade de ter as anáforas resolvidas em seus melhores resultados. Com isso, foi possível desenvolver um sistema de RI baseado na Estrutura Nominal do Discurso de forma simplificada, tornando computacionalmente factível sua criação e manipulação.

Este trabalho toma como base a END simplificada desenvolvida para o sistema de recuperação de informação apresentado em [Pereira 2009]. Como foi visto no capítulo 3, durante a etapa final categorização supervisionada, o processo de resolução de anáforas é utilizado de forma direta. Com isso, o método fica mais dependente de bons resultados no que diz respeito à tarefa de resolução de anáforas. Há, portanto, uma preocupação maior em tornar a Estrutura Nominal do Discurso mais fidedigna à proposta original sem, no entanto, que a performance computacional seja muito prejudicada.

Nas próximas seções são mostradas as considerações assumidas durante a criação da END e na aplicação do método de categorização. Na seção 4.2, são apresentados todos os algoritmos envolvidos no processo de criação da Estrutura Nominal, bem como as melhorias assumidas em relação à implementação inicial do trabalho de Pereira. Em seguida, na seção 4.3, é mostrado o algoritmo geral para a obtenção do categorizador proposto. A seção 4.4 conclui o capítulo, apresentando as considerações finais sobre o que foi discutido.

4.2 Criação da END

A proposta de Freitas para resolução de anáforas assume que o relacionamento de uma anáfora deve ser determinado a partir de termos presentes em frases já apresentadas anteriormente no texto. Ou seja, anáforas intrafrases estão fora do escopo de sua proposta e, portanto, a mesma assunção é feita neste trabalho. Devido a isso, a unidade de representação do texto é baseada em **frases**.

Para a construção da END, é requerida uma série de informações sobre os elementos de cada frase do texto, dentre elas: (1) informação morfológica: gênero, número, pessoa e morfema, (2) informação sintática: sujeito, objeto e objeto direto, (3) informação temática: agente, ator, paciente, tema e localização, (4) informação semântica: tipagem de argumentos dos verbos e (5) informação pragmática: foco de atenção, animacidade das entidades, estrutura de língua e conhecimento de senso comum. Com exceção de (1) e (2), essas informações não estão facilmente disponíveis, seja qual for o corpus de documentos considerado. No caso de (1) e (2), é possível a utilização de um ***POS-tagger*** – *Part-Of-Speech tagger* –, responsável pela atribuição de etiquetas morfológicas e de informações gramaticais e sintáticas para cada palavra do texto. Existem sistemas já desenvolvidos e disponíveis através da Internet, inclusive para a língua portuguesa [Bick 2000, FORMA], que retornam esse tipo de informação.

Com isso disponível, o processo de criação da END é iniciado através da interpretação fora de contexto, que define uma representação estrutural com os elementos apresentados em cada frase do texto. Neste ponto, Freitas considera ainda informações sobre os verbos presentes na frase, entretanto Pereira não leva isso em conta no algoritmo pelo fato de que, para a realização de buscas, elementos como verbos não são tão significativos quanto os nomes apresentados por uma frase [Oliveira e Quental 2003]. No caso do problema de categorização atual, essa afirmativa também é válida.

Dessa maneira, o processo de interpretação fora de contexto beneficia a identificação de elementos nominais, sendo responsável por converter a lista de sintagmas, juntamente com as informações sobre eles, para uma representação computacional da frase, denominada segmento. Para isso, cada termo reconhecido na frase é adicionado à lista de entidades relevantes explícita do segmento recém-criado.

Esse processo é utilizado no Algoritmo 1, no qual é criada uma representação fora de contexto para cada frase do texto considerado, resultando em um novo segmento (linha 3). O passo seguinte é a interpretação em contexto deste segmento, responsável por anexá-lo

à Estrutura Nominal em construção (linha 4). Ao final desse processo iterativo, a END relativa ao documento de entrada estará devidamente gerada e pronta para ser utilizada nos métodos subsequentes (linha 6).

Algoritmo 1 Cria a Estrutura Nominal do Discurso para um dado documento.

Função CriaEND(Documento D)
Pré-condição: Documento D: conteúdo do documento
Pós-condição: END E: estrutura gerada
1: $E \leftarrow [\,]$
2: **Para todo** Frase f em D **faça**
3: Segmento $S \leftarrow$ Interpretação fora de contexto de f
4: InterpretacaoEmContexto(S, E)
5: **Fim Para**
6: **Retorne** E

O processo de interpretação em contexto é o ponto-chave da construção da END, pois além de anexar fisicamente o segmento relativo a uma nova frase à representação das demais frases já interpretadas do texto, é realizada a resolução anafórica dos elementos candidatos à anáfora do segmento. O Algoritmo 2 mostra esse processo[1]. De forma geral, inicialmente ele determina qual é o ponto de interpretação, buscando dentre os segmentos visíveis já anexados aquele que fornece o maior número de resoluções anafóricas para o novo segmento. A partir daí, são estabelecidos os relacionamentos entre os candidatos à anáfora e os respectivos elementos antecedentes dos dois segmentos envolvidos, considerando as relações definidas pelas regras pragmáticas. Além disso, para a anexação do novo segmento à estrutura, é criado um segmento composto localizado no ponto de interpretação, contendo material herdado de seus dois filhos que, de acordo com a nova organização da árvore, são os segmentos anafórico e antecedente.

Da linha 1 à 4 do Algoritmo 2 é feita a verificação do caso em que a Estrutura Nominal E está vazia. Esta situação indica que o segmento é o primeiro identificado do texto e, portanto, não há interpretação a ser realizada, de forma que o segmento S_A represente a estrutura E no momento (linha 2). Na linha 5, a lista dos segmentos visíveis de E é atribuída à variável SV, com os elementos ordenados do mais profundo para a raiz da árvore. Na linha 6 é inicializada uma variável auxiliar para a determinação do ponto de interpretação, onde estará localizado o segmento S_T, inicializado na linha 7. Na linha 8 começa a iteração sobre os elementos de SV. Nesse processo iterativo, nas linhas 8 a 15, o algoritmo determina qual dos segmentos visíveis $S \in SV$ é o que apresenta elementos que permitem o maior número de resoluções para os elementos de S_A.

[1]Nos algoritmos é assumido que uma END é formada por um conjunto de segmentos e um segmento, por sua vez, é formado por um conjunto de termos.

Algoritmo 2 Realiza a interpretação em contexto do segmento recém-criado em relação à END.

Procedimento InterpretacaoEmContexto(Segmento S_A, END E)
Pré-condição: Segmento S_A: novo segmento; END E: estrutura corrente
Pós-condição: Segmento S_A devidamente anexado à estrutura E

1: **Se** $E = [\ ]$ **então**
2: $E \leftarrow$ **novo** END(S_A)
3: **Retorne**
4: **Fim Se**
5: Lista<Segmento> $SV \leftarrow$ Segmentos visíveis de E, do mais profundo para a raiz
6: Inteiro $max \leftarrow 0$
7: Segmento $S_T \leftarrow [\ ]$
8: **Para todo** Segmento S em SV **faça**
9: {*Conta o número de resoluções anafóricas possíveis para os termos de* S_A *a partir dos termos de* $S \in SV$*, e determina qual foi o segmento que possibilitou o maior número de resoluções.*}
10: $n \leftarrow$ **COUNT**(EhCorreferencia(t_1, t_2) **OU** EhMembroDe(t_1, t_2): $t_1 \in S_A$; $t_2 \in S$)
11: **Se** $n > max$ **então**
12: $max \leftarrow n$
13: $S_T \leftarrow S$
14: **Fim Se**
15: **Fim Para**
16: {$S_T = [\ ]$ *se nenhum segmento visível possibilitou resolução das anáforas de* S_A.}
17: **Se** $S_T = [\ ]$ **então**
18: $S_T \leftarrow$ Segmento visível mais à direita
19: **Fim Se**
20: EstabeleceRelacao(S_A, S_T)
21: AnexaSegmento(S_A, S_T, E)

Para a identificação do relacionamento anafórico entre duas entidades, Freitas define cinco tipos de relações: correferência, "membro de", "parte de", "subcategorizado por" e acomodação. A utilização das relações de "parte de" e "subcategorizado por" requer a disponibilização de informações sobre coletivos e animacidade, as quais são de difícil obtenção para um conjunto de dados genéricos. Sendo assim, visando simplificar o método como um todo, tornando-o de implementação factível, somente as seguintes relações são consideradas: correferência, "membro de" e acomodação. Como já foi dito, acomodação é, na verdade, uma pseudorrelação utilizada quando as demais não se enquadram. Este trabalho, portanto, considera que duas entidades estão anaforicamente relacionadas entre si, ou seja, a entidade anafórica é resolvida pela entidade antecedente, quando elas apresentam uma das relações: correferência ou "membro de" (linha 10 do Algoritmo 2).

Ao concluir as iterações das linhas 8 a 15, caso nenhum segmento $S \in SV$ apresente elementos que resolvam as anáforas de S_A, então S_T será vazio (linha 16). Nesse caso, considera-se como ponto de interpretação o segmento visível mais à direita na árvore (linha 17). Seguindo com o algoritmo, na linha 19 é estabelecida a relação entre o segmento anafórico S_A e seu antecedente S_T (função definida no Algoritmo 3) e na linha 20, o segmento é anexado à estrutura (Algoritmo 6).

O próximo algoritmo (Algoritmo 3) trata de estabelecer o relacionamento entre dois segmentos, através da utilização das regras pragmáticas para a definição da relação – correferência, "membro de" ou acomodação – entre os termos presentes no segmento anafórico e no segmento antecedente. Pereira em seu trabalho não considerou o tipo da anáfora em questão ao estabelecer essa relação. Entretanto, de acordo com a proposta de Freitas, esta informação é utilizada visando agregar o conhecimento já adquirido sobre a utilização dos focos e das listas de relevantes em determinados tipos de anáforas.

Sabe-se que o foco explícito assinala a tendência do transmissor em continuar falando sobre um mesmo indivíduo, permitindo assim introduzir mais informações sobre o mesmo. Os pronomes contêm reduzido material informativo, o que torna mais presente a necessidade do foco. Sendo assim, para a resolução de anáforas pronominais, o foco explícito do segmento antecedente S_T é utilizado como primeira opção. Não sendo possível, são consideradas as entidades presentes na lista de entidades explícitas relevantes. Neste caso, pode haver uma troca de foco explícito de S_T, sinalizando a mudança de atenção das entidades atuais em favor de novas entidades.

Já no caso das anáforas nominais definidas é sabido que a expressão anafórica $\mathcal{A}$ e seu antecedente $\mathcal{T}$ não correferenciam a mesma entidade, sendo necessário introduzir

Algoritmo 3 Estabelece a relação entre um segmento anafórico e seu antecedente.

Procedimento EstabeleceRelacao(Segmento S_A, Segmento S_T)

Pré-condição: Segmento S_A: segmento anafórico; Segmento S_T: segmento antecedente

Pós-condição: Segmento S_A devidamente relacionado com o segmento S_T

1: **Para todo** Relacao r em [CORREFERENCIA, MEMBRO_DE] **faça**
2: **Para todo** Termo t_A em S_A **e** t_A é candidato a anáfora **e** t_A.Relacao $= \perp$ **faça**
3: Termo $t_T = \perp$
4: **Se** t_A é Pronome **então**
5: **Se** Relaciona(t_A, S_T.FocoExplicito, r) **então**
6: $t_T \leftarrow S_T$.FocoExplicito
7: **Senão Se** Relaciona(t_A, $t' : \forall$ Termo $t' \in S_T$.ListaExplicita, r) **então**
8: $t_T \leftarrow t'$
9: S_T.FocoExplicito $\leftarrow t_T$ {*Fazendo os devidos ajustes para a troca.*}
10: **Fim Se**
11: **Senão Se** t_A é SND **então**
12: **Se** Relaciona(t_A, S_T.FocoImplicito, r) **então**
13: $t_T \leftarrow S_T$.FocoImplicito
14: **Senão Se** Relaciona(t_A, $t' : \forall$ Termo $t' \in S_T$.ListaImplicita, r) **então**
15: $t_T \leftarrow t'$
16: S_T.FocoImplicito $\leftarrow t_T$
17: **Senão Se** Relaciona(t_A, S_T.FocoExplicito, r) **então**
18: $t_T \leftarrow S_T$.FocoExplicito
19: **Senão Se** Relaciona(t_A, $t' : \forall$ Termo $t' \in S_T$.ListaExplicita, r) **então**
20: $t_T \leftarrow t'$
21: S_T.FocoExplicito $\leftarrow t_T$
22: **Fim Se**
23: **Fim Se**
24: **Se** $t_T \neq \perp$ **então**
25: S_A.ListaImplicita.Adiciona(t_T)
26: Ajusta listas de entidades explícita e implícita, ordenando-as e atualizando os respectivos focos
27: **Fim Se**
28: **Fim Para**
29: **Fim Para**
30: **Para todo** Termo t_A em S_A **e** t_A é candidato a anáfora **e** t_A.Relacao $= \perp$ **faça**
31: t_A.Antecedente $\leftarrow$ **nulo**
32: t_A.Relacao $\leftarrow t_T$.Relacao $\leftarrow$ ACOMODACAO
33: **Fim Para**

uma relação $\mathcal{R}$ entre ambos: $\mathcal{R}(\mathcal{A}, \mathcal{T})$. Devido a isso, a utilização do foco implícito do segmento S_T é propícia para a resolução de ANDs, pois ele acompanha as entidades que foram correferenciadas através de SNDs ao longo do discurso. Se não for possível a utilização do foco implícito, é então considerada a lista de entidades implícitas relevantes, sinalizando a mudança de assunto do discurso. Caso ainda não tenha sido identificado nenhum antecedente, são também considerados o foco explícito e a lista de entidades explícitas relevantes do segmento antecedente.

Seguindo os passos acima para resolução de anáforas pronominais e ANDs, caso o antecedente seja determinado, a lista de entidades relevantes implícitas do segmento anafórico é atualizada. Sempre que há uma alteração em uma lista de relevantes, seja ela implícita ou explícita, sua ordenação deve ser refeita. Adequando a regra de ordenação definida na seção 3.2.3 para o problema considerado, obtém-se a seguinte regra geral:

$$\begin{array}{ccc} \text{entidades anafóricas} & > & \text{entidades não anafóricas} \\ \text{pronomes} > \text{SND} & & \text{sujeito} > \text{objeto} \\ \text{sujeito} > \text{objeto} & & \end{array} \tag{4.1}$$

Se, por outro lado, não for possível identificar o antecedente através das regras de correferência ou "membro de", é utilizada a relação de acomodação. Os procedimentos para a resolução de anáforas descritos anteriormente são aplicados a cada termo do segmento anafórico S_A que é candidato a anáfora. Como neste trabalho só são consideradas anáforas pronominais e nominais definidas, são definidas como candidatas a anáfora aquelas entidades que: (1) estão classificadas como pronomes pessoais ou (2) estão classificadas como substantivos, tendo como precedência um artigo definido.

O Algoritmo 3 consiste de um processo iterativo, que percorre todos os termos t_A do segmento anafórico que são candidados a anáfora (linha 2), considerando as relações correferência e "membro de" separadamente (linha 1), exatamente nessa ordem[2]. Nas linhas 4 a 23, são aplicados os procedimentos para identificação do antecedente t_T de uma anáfora, especificando as situações em que o termo anafórico é um pronome (linhas 5 a 10) e em que ele é um sintagma nominal definido (linhas 12 a 22). No primeiro caso, inicialmente é feita a tentativa de relacionar t_A com o foco explícito do segmento S_T considerando a relação r, na linha 5 do algoritmo (função definida no Algoritmo 4), atualizando a variável t_T para esse valor em caso de positivo (linha 6). Se não foi

[2]Os atributos "FocoImplicito" e "FocoExplicito" de um segmento utilizados nos algoritmos retornam um termo, e os atributos "ListaImplicita" e "ListaExplicita", representando a lista de entidades relevantes implícita e explícita, respectivamente, retornam um conjunto de termos.

possível, a tentativa é refeita para a lista de entidades explícitas relevantes do segmento S_T, experimentando cada termo t' pertencente a essa lista (linha 7). Se foi possível relacionar algum t' ao termo t_A, então a variável t_T é associada na linha 8 e é feita a troca do foco explícito do segmento antecedente na linha 9, realizando os seguintes ajustes: o antigo foco passa a integrar a lista de entidades relevantes explícitas e o novo foco é atualizado; como a lista de relevantes foi alterada, ela deve ser novamente ordenada, seguindo a regra definida em (4.1).

Se o termo anafórico é um SND, são feitas as tentativas de relacionamento de t_A com o foco implícito (linha 12), lista de entidades implícitas (linha 14), foco explícito (linha 17) e lista de entidades explícitas (linha 19), só partindo para o caso seguinte, na ordem, se o atual falhar. Em todos os casos, se o relacionamento foi possível, a variável t' é atualizada para o valor correspondente e os focos (nas situações em que são analisadas as listas de relevantes) são alterados, realizando os devidos ajustes para a troca.

Seguindo o algoritmo, na linha 24 é feita a verificação do valor da variável t_T: se ela apresentar um valor diferente daquele com o qual ela foi inicializada na linha 3, significa que foi encontrado um antecedente para o termo anafórico t_A. Neste caso, o termo antecedente t_T é inserido na lista de entidades implícitas do segmento anafórico (linha 25) e as duas listas de relevantes são devidamente ajustadas e reordenadas segundo a regra (4.1) (linha 26).

Concluídas as iterações definidas nas linhas 1 e 2, é realizada uma outra varredura nos termos t_A candidatos à anáfora, para os quais não foi possível identificar um antecedente (linha 30). Nesses casos, o atributo "Antecedente" de t_A é atribuído para nulo (linha 31), e a relação entre t_A e t_T é atualizada para acomodação (linha 32).

Segundo a proposta original, as relações podem ser contabilizadas de forma conjunta em uma só iteração sobre os termos t_A, da seguinte forma: para cada termo t_A, são realizados os testes relativos a pronomes e a SNDs, tal qual foi definido no Algoritmo 3, porém verificando a possibilidade de relacionar t_A com o possível antecedente por correferência **ou** por "membro de" (na teoria de Freitas, também considerando "parte de" e "subcategorizado por"). Em outras palavras, nesse caso a função "Relaciona" retornaria verdadeiro se conseguisse relacionar os termos por qualquer uma das relações, sem especificá-la. Neste trabalho, entretanto, além de desconsiderar as relações "parte de" e "subcategorizado por" por motivos já relacionados, houve uma simplificação da regra de "membro de", de modo que ela considera somente casos particulares dessa situação. Sendo assim, a regra para correferência é mais relevante na busca pelo relacionamento – motivo pelo qual foi

decidido neste trabalho realizar o processo de determinação da relação da forma como foi apresentada no algoritmo: percorrer todos os termos t_A primeiramente analisando somente a relação de correferência e, para os termos restantes (para os quais não foi possível estabelecer a relação), analisando a relação de "membro de".

O Algoritmo 4 é o responsável pela tentativa de relacionar um termo anafórico t_A com um suposto antecedente t_T sob uma dada relação r. Para cada valor de r, são feitas as verificações se existe uma relação de correferência entre esses termos (linha 1) ou uma relação de "membro de" (linha 5), casos em que o atributo "Antecedente" de t_A é devidamente ajustado e a relação entre t_A e t_T é atualizada para seu valor respectivo, finalizando com o retorno verdadeiro para a função. Se não foi possível estabelecer o relacionamento r, então a função simplesmente retorna falso (linha 10).

Algoritmo 4 Verifica se é possível ou não relacionar um termo anafórico e um termo antecedente, segundo uma dada relação.

Função Relaciona(Termo t_A, Termo t_T, Relacao r)
Pré-condição: Termo t_A: termo do segmento anafórico;
Termo t_T: termo do segmento antecedente;
Relacao r: relação a ser verificada
Pós-condição: Termos t_A e t_T relacionados entre si (no caso de positivo)
1: **Se** r = CORREFERENCIA **e** EhCorreferencia(t_A, t_T) **então**
2: t_A.Antecedente $\leftarrow t_T$
3: t_A.Relacao $\leftarrow t_T$.Relacao $\leftarrow$ CORREFERENCIA
4: **Retorne** Verdadeiro
5: **Senão Se** r = MEMBRO_DE **e** EhMembroDe(t_A, t_T) **então**
6: t_A.Antecedente $\leftarrow t_T$
7: t_A.Relacao $\leftarrow t_T$.Relacao $\leftarrow$ MEMBRO_DE
8: **Retorne** Verdadeiro
9: **Senão**
10: **Retorne** Falso
11: **Fim Se**

As funções fornecidas pelo Algoritmo 5 verificam se um termo anafórico t_A está relacionado através das relações de correferência e "membro de" com um termo supostamente antecedente t_T. Para isso, são consideradas as regras para identificação das relações de correferência, definida na seção 3.2.2.1, e de "membro de", definida na seção 3.2.2.2. Como já mencionado anteriormente, este trabalho desconsidera a utilização da informação de coletividade nos algoritmos, devido à sua dificuldade de obtenção para um conjunto de dados genéricos. Sendo assim, as regras tiveram que ser simplificadas, de modo a considerar somente as informações disponíveis. A relação de correferência assumida nos algoritmos pode ser reescrita da seguinte forma:

Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de um pronome ou SND e $\mathcal{A}$ e $\mathcal{T}$ concordam em número e gênero, então $\mathcal{R}$ pode ser uma relação de correferência.

Já a relação de "membro de" pode ser simplificada para:

Se $\mathcal{A}$ tiver sido introduzido no discurso por meio de SND e $\mathcal{T}$ está no plural, $\mathcal{A}$ está no singular e $\mathcal{A}$ e $\mathcal{T}$ concordam em gênero, então pode-se assumir que $\mathcal{R}$ é uma relação de "membro de".

A primeira função do Algoritmo 5 retorna verdadeiro, caso os atributos de t_A e t_T satisfaçam as condições da regra simplificada para a relação de correferência apresentada acima, e falso, caso contrário. A segunda função definida neste mesmo algoritmo aplica a lógica para a determinação de uma relação de "membro de" do modo como foi estipulado em sua forma simplificada, retornando verdadeiro se atender às condições e falso, caso contrário.

Algoritmo 5 Funções que verificam se dois termos estão relacionados por relação de correferência e "membro de", respectivamente.

Função EhCorreferencia(Termo t_A, Termo t_T)
Pré-condição: Termo t_A: termo do segmento anafórico;
Termo t_T: termo do segmento antecedente
Pós-condição: Retorna se t_A e t_T estão relacionados por correferência
1: **Se** (t_A.ClasseGramatical = PRONOME **ou** t_A.ClasseGramatical = SND) **e** (t_A.Genero = t_T.Genero) **e** (t_A.Numero = t_T.Numero) **então**
2: **Retorne** Verdadeiro
3: **Senão**
4: **Retorne** Falso
5: **Fim Se**

Função EhMembroDe(Termo t_A, Termo t_T)
Pré-condição: Termo t_A: termo do segmento anafórico;
Termo t_T: termo do segmento antecedente
Pós-condição: Retorna se t_A e t_T estão relacionados por "membro de"
1: **Se** (t_A.ClasseGramatical = SND) **e** (t_A.Genero = t_T.Genero) **e** (t_A.Numero = Singular) **e** (t_T.Numero = Plural) **então**
2: **Retorne** Verdadeiro
3: **Senão**
4: **Retorne** Falso
5: **Fim Se**

Voltando ao Algoritmo 2, uma vez estabelecido o relacionamento entre o segmento anafórico S_A e seu antecedente S_T, o segmento S_A deve ser então anexado à estrutura, tendo como ponto de interpretação o segmento S_T. Para esse processo, como mostrado na

seção 3.3, um novo segmento composto é criado no ponto de ancoragem, herdando atributos dos seus filhos – S_T à esquerda e S_A à direita. Pereira assumiu em seu trabalho uma única forma de determinação para essa herança: o segmento composto contém atributos idênticos aos do segmento antecedente S_T. Todavia, segundo a teoria de Freitas, a criação de um segmento depende da relação existente entre os focos de cada segmento-filho. Esta relação é determinada a partir da mudança ou manutenção do foco implícito, sinalizando uma mudança ou continuação do assunto, ou através da mudança ou manutenção do foco explícito, indicando a mudança ou manutenção do tópico de uma frase (efeito local). As tabelas 3.1 e 3.2 mostram, respectivamente, esses relacionamentos e os valores correspondentes para cada tipo de segmento. Este trabalho opta pela utilização dessas informações, visando inserir maior semântica ao processo de criação da END.

Para a anexação de segmento anafórico à estrutura, o Algoritmo 6 inicialmente cria um novo segmento S_{novo} (linha 1) e, para a definição dos valores dos atributos de S_{novo}, considera os seguintes tipos: elaboração, manutenção de tópico, mudança de tópico e mudança de assunto. Nas linhas 2 a 30 são tratados os casos relativos a cada tipo de segmento composto, da forma como foi apresentado na seção 3.3, atribuindo os devidos valores aos focos e às listas de relevantes de S_{novo}. O segmento S_T é então atribuído como filho à esquerda de S_{novo} na linha 31 e S_A, como filho à direita na linha 32. Nas linhas 33 e 34 são ajustados os atributos "Pai" dos segmentos S_T e S_A, respectivamente.

Com isso, a interpretação em contexto de um novo segmento está concluída. O segmento básico criado, resultante da interpretação fora de contexto, agora encontra-se devidamente anexado à estrutra, tendo todos os relacionamentos estabelecidos. O processo de criação da Estrutura Nominal do Discurso, portanto, está completo. Na próxima seção, será mostrada a especificação do processo geral do categorizador proposto, envolvendo a fase de criação da END, assim como o método de *bootstrapping* e a utilização do categorizador supervisionado.

4.3 Categorização

O Algoritmo 7 apresenta o processo geral do categorizador proposto neste trabalho. São fornecidos como entrada o conjunto de documentos do corpus em questão e a lista contendo os nomes (rótulos) das categorias consideradas. Com isso, são geradas as ENDs para cada documento, a partir das quais é estabelecido o processo de *bootstrapping*, que define uma rotulação para os documentos considerados. Utilizando essa rotulação o cate-

Algoritmo 6 Anexa o segmento anafórico à estrutura dado o segmento antecedente como ponto de interpretação.

Procedimento AnexaSegmento(Segmento S_A, Segmento S_T, END E)

Pré-condição: Segmento S_A: segmento anafórico; Segmento S_T: segmento antecedente; END E: estrutura corrente

Pós-condição: Segmento S_A anexado à estrutura E no ponto de interpretação S_T

1: Segmento $S_{novo} \leftarrow$ **novo** Segmento()
2: **Se** S_T.FocoExplicito = S_A.FocoExplicito **então**
3: **Se** S_T.FocoImplicito = $\perp$ **ou** S_T.FocoImplicito = S_A.FocoImplicito **então**
4: {*Segmento tipo elaboração*}
5: S_{novo}.FocoExplicito $\leftarrow$ S_T.FocoExplicito
6: S_{novo}.FocoImplicito $\leftarrow$ S_A.FocoImplicito
7: S_{novo}.ListaExplicita $\leftarrow$ [S_{novo}.FocoExplicito]
8: S_{novo}.ListaImplicita $\leftarrow$ [S_{novo}.FocoImplicito]
9: **Senão**
10: {*Segmento tipo manutenção de tópico*}
11: S_{novo}.FocoExplicito $\leftarrow \perp$
12: S_{novo}.FocoImplicito $\leftarrow$ S_T.FocoImplicito
13: S_{novo}.ListaExplicita $\leftarrow$ [ ]
14: S_{novo}.ListaImplicita $\leftarrow$ S_T.ListaImplicita
15: **Fim Se**
16: **Senão**
17: **Se** S_T.FocoImplicito = $\perp$ **ou** S_T.FocoImplicito = S_A.FocoImplicito **então**
18: {*Segmento tipo mudança de tópico*}
19: S_{novo}.FocoExplicito $\leftarrow$ S_T.FocoExplicito
20: S_{novo}.FocoImplicito $\leftarrow$ S_A.FocoImplicito
21: S_{novo}.ListaExplicita $\leftarrow$ S_T.ListaExplicita
22: S_{novo}.ListaImplicita $\leftarrow$ [S_{novo}.FocoImplicito]
23: **Senão**
24: {*Segmento tipo mudança de assunto*}
25: S_{novo}.FocoExplicito $\leftarrow$ S_T.FocoExplicito
26: S_{novo}.FocoImplicito $\leftarrow$ S_T.FocoImplicito
27: S_{novo}.ListaExplicita $\leftarrow$ S_T.ListaExplicita
28: S_{novo}.ListaImplicita $\leftarrow$ S_T.ListaImplicita
29: **Fim Se**
30: **Fim Se**
31: S_{novo}.FilhoAEsquerda $\leftarrow$ S_T
32: S_{novo}.FilhoADireita $\leftarrow$ S_A
33: S_T.Pai $\leftarrow$ S_{novo}
34: S_A.Pai $\leftarrow$ S_{novo}

gorizador, por fim, é definido.

Seguindo o processo passo-a-passo, o algoritmo parte da inicialização da variável $\mathcal{D}$, que representa a lista das ENDs relativas aos documentos (linha 1). As linhas 2 a 4 criam a END para cada documento, atualizando $\mathcal{D}$, iterativamente, através da função CriaEND() definida no Algoritmo 1. Com a lista $\mathcal{D}$ estabelecida, o algoritmo parte para a definição da lista de categorias $\mathcal{C}$, nas linhas 5 a 12. Na linha 5 ela é inicializada e, da 6 à 12, é gerada uma nova categoria para as quais são definidos os atributos: rótulo (linha 8), palavras relacionadas (linha 9) e conjunto expandido de palavras relacionadas (linha 10), e na linha 11 a lista $\mathcal{C}$ é atualizada, tudo isso seguindo um processo iterativo para cada um dos rótulos fornecidos como entrada. O atributo "PalavrasRelacionadas" (linha 9) estabelece manualmente um conjunto de entidades que definem a categoria relativa ao rótulo em questão. O atributo "PalavrasRelacionadasExpandido" (linha 10) parte do conjunto de termos definidos na linha 9 e utiliza um dicionário de sinônimos para a expansão desse conjunto, visando representá-lo melhor.

Algoritmo 7 Processo geral do categorizador proposto.

Procedimento Categorizador(Lista<Documentos> D, Lista<Rotulo> R)
Pré-condição: Lista<Documento> D: lista dos documentos do corpus considerado; Lista<Rotulo> R: lista dos rótulos das categorias consideradas
Pós-condição: Categorizador gerado
1: Lista<END> $\mathcal{D} \leftarrow [\,]$
2: **Para todo** Documento d em D **faça**
3: $\quad \mathcal{D} \leftarrow \mathcal{D} \cup$ CriaEND(d)
4: **Fim Para**
5: Lista<Categoria> $\mathcal{C} \leftarrow [\,]$
6: **Para todo** Rotulo r em R **faça**
7: $\quad$ Categoria $c \leftarrow$ **novo** Categoria()
8: $\quad c$.Rotulo $\leftarrow r$
9: $\quad c$.PalavrasRelacionadas $\leftarrow$ Define palavras relacionadas para r
10: $\quad c$.PalavrasRelacionadasExpandido $\leftarrow$ Define conjunto expandido de palavras relacionadas para r
11: $\quad \mathcal{C} \leftarrow \mathcal{C} \cup c$
12: **Fim Para**
13: {*O processo de* bootstrapping *define uma rotulação, a partir da atualização do atributo* CategoriaAssociada *de cada documento.*}
14: Bootstrapping($\mathcal{D}$, $\mathcal{C}$)
15: {*Com essa rotulação, aplica-se qualquer categorizador supervisionado da literatura.*}
16: CategorizadorSupervisionado($\mathcal{D}$, $\mathcal{C}$)

Tendo os conjuntos $\mathcal{D}$ e $\mathcal{C}$ gerados, o Algoritmo 7 parte para o processo de *bootstrapping* (linha 14), definido no Algoritmo 8 e detalhado na seção 3.5. Ao fim desse processo, o atributo "CategoriaAssociada" de cada documento estará estabelecido, e será utilizado

para a criação do categorizador supervisionado (linha 16). O método de categorização não está definido, uma vez que o foco deste trabalho não está nessa questão. Portanto, considera-se que a linha 16 está referindo a um método qualquer de categorização supervisionada de respaldo na literatura.

Como discutido no capítulo 3, os documentos que servem de entrada para o categorizador supervisionado apresentam duas características principais: contêm somente termos considerados de alto valor semântico e as anáforas existentes em seu conteúdo estão resolvidas. Ambas informações são obtidas durante a fase de construção da END, mais especificamente no Algoritmo 3. Nesse algoritmo, o conteúdo do documento relativo à END em execução no momento é atualizado para conter somente as entidades t_A pertencentes ao segmento anafórico S_A, além de ter suas anáforas resolvidas pelos termos t_T determinados como antecedentes no algoritmo. As entidades $t_A \in S_A$ foram previamente inicializadas no processo de interpretação fora de contexto do segmento, no qual os termos de interesse lidos da frase são armazenados na lista de entidades relevantes explícita de S_A. Com isso, o conteúdo dos documentos está devidamente tratado, bastando determinar a rotulação dos mesmos através do Algoritmo 8, onde é estabelecido o atributo "CategoriaAssociada" para cada documento.

Para o processo de *bootstrapping*, inicialmente deve ser definido o conjunto de palavras características para cada categoria, o que representa o ponto-chave do método. Para isso, são necessários três conjuntos interdependentes: conjunto de palavras relacionadas – definidas por um operador humano –, conjunto expandido de palavras relacionadas – obtido com a utilização de um dicionário de sinônimos para a expansão do conjunto inicial – e o conjunto de palavras-chave – utilizando informações de coocorrência entre o conjunto expandido e os documentos. Os dois primeiros já foram definidos no Algoritmo 7, restando somente o conjunto de palavras-chave. Na seção 3.5.1, foram apresentadas fórmulas propostas para a obtenção desse conjunto.

No Algoritmo 8, é detalhado o processo de *bootstrapping*. Nas linhas 2 a 5, são definidos o conjunto de palavras-chave e o conjunto mais geral de palavras características para cada categoria, iterativamente. A linha 3 simplesmente aplica a proposta definida na seção 3.5.1 para a obtenção das palavras-chave, e a linha 4 atribui ao conjunto de palavras características: o rótulo da categoria juntamente com a união dos conjuntos de palavras relacionadas, conjunto expandido e de palavras-chave recém-obtido. Com as palavras características de cada categoria definidas, o algoritmo então realiza a associação de cada documento à categoria com a qual apresentar maior valor de relevância (linha 7),

Algoritmo 8 Processo de *bootstrapping*.

Procedimento Bootstrapping(Lista<END> $\mathcal{D}$, Lista<Categoria> $\mathcal{C}$)

Pré-condição: Lista<END> $\mathcal{D}$: lista das ENDs de todos os documento do corpus; Lista<Categoria> $\mathcal{C}$: categorias consideradas, com o conjunto expandido de palavras relacionadas já definido

Pós-condição: Documentos das ENDs de $\mathcal{D}$ rotulados nas categorias de $\mathcal{C}$

1: {*Define o conjunto de palavras características para cada categoria.*}
2: **Para todo** Categoria c em $\mathcal{C}$ **faça**
3: c.PalavrasChave ← Define as palavras-chave de c, de acordo com a seção 3.5.1
4: c.PalavrasCaracteristicas ← [c.Rotulo] ∪ c.PalavrasRelacionadas ∪ c.PalavrasRelacionadasExpandido ∪ c.PalavrasChave
5: **Fim Para**
6: **Para todo** END E em $\mathcal{D}$ **faça**
7: E.CategoriaAssociada ← $\max_{c \in \mathcal{C}} \left\{ \frac{\sum_{p \in c.\mathrm{PC}} \mathrm{VR}(p, E)}{|c.\mathrm{PC}|} \right\}$
8: {PC *é um pseudônimo para* PalavrasCaracteristicas.}
9: {VR() *retorna o valor de relevância definido em [Pereira 2009].*}
10: **Fim Para**

seguindo o processo iterativo das linhas 6 a 10.

4.4 Considerações finais

Este capítulo apresentou uma descrição detalhada sobre os algoritmos desenvolvidos para a obtenção do método de categorização proposto. Toda a metodologia é baseada na Estrutura Nominal do Discurso proposta por Freitas [Freitas 2005] e posteriormente adaptada para uma linguagem de programação que não utiliza paradigma lógico, por Pereira [Pereira 2009]. Este autor estipulou algumas considerações e simplificações para a implementação da estrutura, sem muitas das quais seria inviável o processamento em termos computacionais.

Para este trabalho existe uma certa dependência entre a boa caracterização da END, juntamente com a detecção dos antecedentes para as anáforas, e o bom resultado da categorização. Devido a isso, foi dado um foco grande à fase de criação da END, explicitando as melhorias em relação aos algoritmos de Pereira, buscando aproximar ao máximo – não fugindo do limite computacional – da teoria de Freitas. Nesse sentido, é dada uma atenção especial aos procedimentos "EstabeleceRelacao" (Algoritmo 3) e "AnexaSegmento" (Algoritmo 6), para os quais foi embutido conhecimento a partir da utilização de informações intrínsecas da estrutura, visando uma melhor representação da END.

Além disso, foram introduzidos os algoritmos contendo o processo geral do método de categorização proposto, apresentado no capítulo 3.

O capítulo a seguir apresenta as experimentações executadas para a avaliação do método, e os respectivos resultados obtidos.

5 *Experimentações e resultados*

Este capítulo disserta sobre as características do protótipo desenvolvido, mostra um exemplo de execução do processo de criação da END e apresenta os experimentos realizados e os resultados obtidos.

5.1 Introdução

Este capítulo apresenta maiores detalhes sobre o protótipo desenvolvido, assim como a metodologia utilizada para a realização dos experimentos. Na seção 5.2, são mostrados detalhes do processo geral do protótipo. A seção seguinte, 5.3, apresenta um exemplo de execução, mostrando passo-a-passo a geração da END. Por fim, na seção 5.4.2, são expostos os experimentos realizados para a obtenção e a avaliação do categorizador proposto.

5.2 Protótipo

Para a realização dos experimentos, foi implementado um sistema baseado na metodologia de categorização proposta; para isso, foram utilizadas as linguagens de programação Java, para a construção do sistema principal, e C e Python, para a obtenção de *scripts* independentes, principalmente durante a realização dos testes. O diagrama da Figura 5.1 mostra, de maneira mais detalhada do que a apresentada na Figura 3.5, a organização do sistema concebido por diferentes módulos responsáveis por cada uma das etapas do processo. Observam-se três fases em destaque: criação das ENDs, *bootstrapping* e categorização.

Como pode ser observado no diagrama, a etapa de processamento do texto a partir da utilização do *tagger* foi implementada externamente ao protótipo. Nessa primeira etapa, o sistema requer que o texto a ser recebido como entrada tenha sua classe gramatical identificada – cujo procedimento é realizado por sistemas denominados de *Part-Of-Speech tagger* ou *POS tagger*. O protótipo desenvolvido não possui um módulo para a realização deste procedimento, devido à complexidade da tarefa de construção de um *tagger* próprio para o sistema e pela disponibilidade de uma coleção já marcada pelo PALAVRAS [Bick 2000].

De posse das classes gramaticais das palavras existentes no texto, o sistema inicialmente realiza a construção da estrutura do índice do documento, que consiste na representação da Estrutura Nominal obtida pela interpretação do texto do documento. Nessa etapa, o módulo indexador aplica os algoritmos para a construção da END sobre a entrada processada do texto, e apresenta como saída: o texto estruturado do documento – formando o índice –, e o conteúdo do mesmo apresentando somente entidades de interesse e com as anáforas existentes resolvidas. O índice criado pelo sistema tem como propósito permitir uma manipulação eficaz da Estrutura Nominal do Discurso de cada documento,

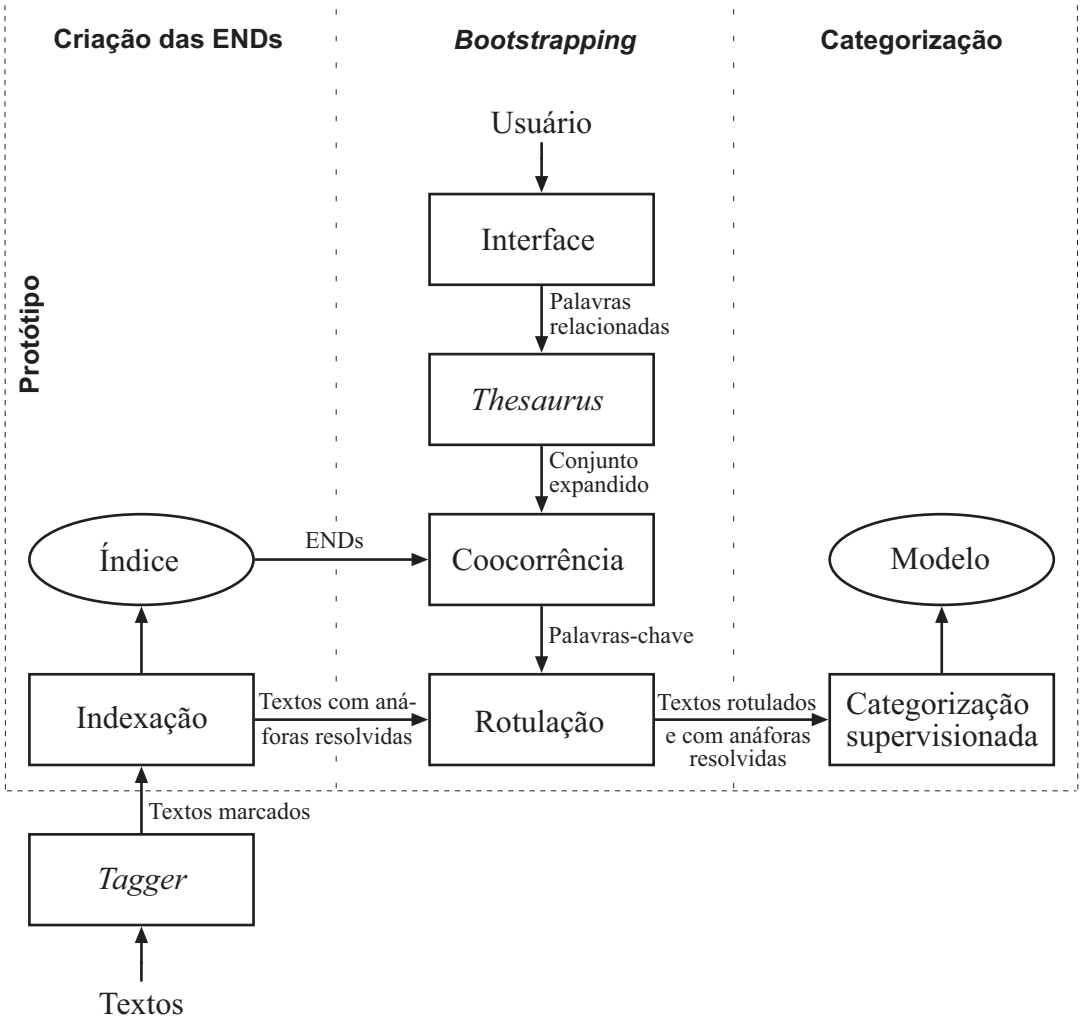


Figura 5.1: Diagrama representativo do sistema.

além de possibilitar a depuração da estrutura obtida. Por motivos equivalentes, Pereira optou em seu trabalho pela persistência dessas informações em arquivos no formato XML, escolha esta mantida no sistema de categorização aqui proposto.

A segunda fase do protótipo trata do método de *bootstrapping* proposto. Como ponto de partida, há uma interação com o usuário, que deve fornecer palavras relacionadas a cada categoria. Internamente, o sistema simplesmente lê de um determinado arquivo essas informações, não sendo necessária uma interface com o usuário propriamente dita. Contudo, é importante ressaltar no diagrama que este momento é o único em todo o sistema em que há requisição ao usuário; em outras palavras, a atuação de um especialista só é exigida para o fornecimento desses termos, cuja quantidade pode variar entre dois e cinco, de acordo com a recomendação de Ko e Seo em [Ko e Seo 2009] – ou seja, essa tarefa demanda um esforço mínimo para o especialista do domínio.

Em seguida, o sistema utiliza um dicionário de sinônimos – *thesaurus* [Maziero Thiago A. S. Pardo 2008] – para ampliar o conjunto de palavras relacionadas para cada categoria, a partir da utilização de termos sinônimos das mesmas, gerando, assim, o conjunto expandido de palavras relacionadas. Com isso, o sistema parte para a definição das palavras-chave de cada categoria. Para tal, são utilizados padrões de coocorrência entre as entidades do conjunto expandido e os termos presentes nos documentos já indexados e representados por suas respectivas ENDs. A união dos termos relacionados com as palavras-chave forma o conjunto de palavras características de cada categoria.

Como último passo da fase de *bootstrapping* surge o módulo de rotulação. Este módulo é alimentado pelo conjunto de palavras características de cada categoria e pelos documentos com as anáforas dos seus textos resolvidas – obtidos durante o processo de indexação. Com isso, o módulo de rotulação retorna os documentos associados às categorias consideradas, tendo já em seu conteúdo as anáforas resolvidas.

Por fim, a última fase do protótipo utiliza a rotulação e os documentos retornados da fase de *bootstrapping* para a geração do modelo de categorização por meio de um categorizador supervisionado.

5.3 Exemplo de execução

Esta seção apresenta, por meio de um exemplo, uma descrição passo-a-passo do processo de criação da END e definição dos antecedentes das anáforas contidas no texto. Além disso, é mostrado um exemplo abstrato de como seria o comportamento do processo

de associação do documento a uma categoria de interesse.

Considere que o texto de um documento D é dado pelo seguinte trecho da obra "Triste Fim de Policarpo Quaresma" de Lima Barreto[1]:

> "*Quaresma jantava e almoçava ali mesmo. (...) As refeições eram-lhe fornecidas por um frege próximo (...).* A refeição principal sempre era carne. *Porque a casa em que se acantonara o destacamento, era o pavilhão do imperador (...). Ficavam nela também a estação da estrada de ferro do Rio Douro e uma grande e bulhenta serraria.*"
> (Triste Fim de Policarpo Quaresma, Lima Barreto)

Após submeter o texto ao PALAVRAS, obtém-se seu conteúdo gramaticalmente classificado. Abaixo encontra-se a saída obtida pelo *tagger* para as três primeiras frases de D:

```
<s>
Quaresma [Quaresma] <hum> PROP M S @SUBJ>
jantava [jantar] <vi> <fmc> V IMPF 3S IND VFIN @FMV
e [e] KC &CO
almoçava [almoçar] <vi> <fmc> V IMPF 3S IND VFIN @FMV
ali [ali] <aloc> ADV @<ADVL
mesmo [mesmo] <quant> ADV @<ADVL
$.  [$.] PU <<<
$¶  [$¶] PU <<<
</s>
<s>
as [o] <artd> DET F P @>N
refeições [refeição] <occ> N F P @SUBJ>
eram- [ser] <fmc> V IMPF 3P IND VFIN @FAUX
lhe [ele] PERS M/F 3S DAT @<DAT
fornecidas [fornecer] V PCP F P @IMV @#ICL-AUX<
por [por] PRP @<ADVL
um [um] <arti> DET M S @>N
frege [frege] <sit> <build> N M S @P<
próximo [próximo] ADJ M S @N<
```

[1] A frase realçada no trecho foi inserida manualmente para que o exemplo também considere o caso específico de relacionamento entre segmentos: "membro de".

```
$.  [$.] PU <<<
$¶  [$¶] PU <<<
</s>
<s>
a [o] <artd> DET F S @>N
refeição [refeição] <occ> N F S @SUBJ>
principal [principal] <SUP> ADJ F S @N<
sempre [sempre] <atemp> ADV @ADVL>
era [ser] <vK> <fmc> V IMPF 3S IND VFIN @FMV
carne [carne] <food> N F S @<SC
$.  [$.] PU <<<
$¶  [$¶] PU <<<
</s>
```

As frases são delimitadas pelas *tags* `<s>` e `</s>`. Cada termo presente no texto é apresentado pelo seu radical entre colchetes (*stem* [Orengo e Huyck 2001]). As demais marcações fornecem as informações necessárias para a construção da Estrutura Nominal do texto, como função sintática, gênero e número[2].

Para a construção da END relativa ao documento D, a interpretação é realizada frase a frase, gerando um segmento para cada uma por meio do processo de interpretação fora de contexto, e o inserindo na estrutura através do processo de interpretação em contexto. Para a primeira frase, o segmento originado é adicionado à estrutura vazia, formando a END dada pela Figura 5.2.

SB_1	*Básico*
$foco^{exp}$ = Quaresma LR^{exp} = [ ]	
$foco^{imp}$ = null LR^{imp} = [ ]	

Figura 5.2: Estrutura após a interpretação da primeira frase.

A interpretação da primeira frase resultou na indexação de somente um termo de interesse: *"Quaresma"*, que é o foco explícito do segmento básico SB_1 na Figura 5.2. Prosseguindo, com a interpretação da segunda frase – Figura 5.3 –, é gerado o segmento SB_2, tendo como foco explícito o termo *"ele"* e como entidades relevantes explícitas:

[2]Para maiores informações sobre o significado dos símbolos utilizados no *tagger*: [PALAVRAS].

"refeição" e *"frege"*. Esses atributos são estipulados dessa forma pois, durante a interpretação fora de contexto, os três termos do segmento são arranjados seguindo as regras de ordenação definidas na seção 3.2.3 (e adaptadas na seção 4.2), e é selecionado como foco explícito a entidade mais bem colocada.

Para a introdução de SB_2 à estrutura, é identificado como ponto de interpretação o segmento SB_1 (que até o momento é o único da END) e, nesse ponto, é anexado um segmento composto SC_1 do tipo mudança de tópico (cujo conteúdo é obtido pelas informações da Tabela 3.2). A anáfora dada pelo pronome *"ele"* teve seu antecedente corretamente identificado, apresentando uma relação de correferência com o termo *"Quaresma"*. Devido a esse relacionamento, a lista de relevantes implícita do segmento SB_2 é atualizada para conter esse termo correferenciado[3].

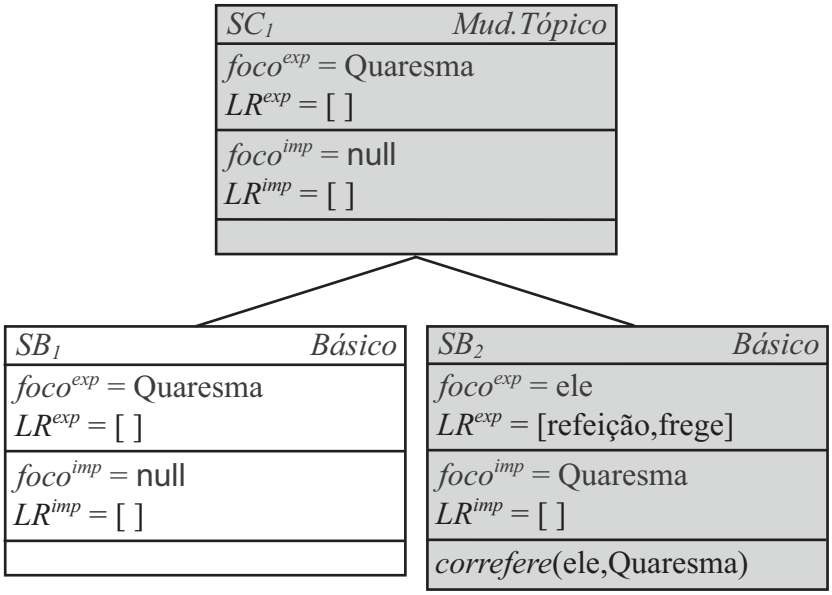


Figura 5.3: Estrutura após a interpretação da segunda frase.

A frase seguinte *"A refeição principal sempre era carne."* introduz uma relação de "membro de" entre o termo *"refeição"* e o termo *"refeições"* da frase anterior. No processo de interpretação, esse relacionamento é identificado e duas consequências se seguem: o termo *"refeição"* é inserido na lista de entidades implícita do segmento SB_3 relativo à terceira frase, e há uma troca de foco explícito do segmento antecedente SB_2 para esse mesmo termo, sinalizando a mudança de atenção das entidades. Com essas modificações, o segmento composto SC_2 é criado, com o tipo manutenção de tópico, e é anexado à END. A estrutura formada após a inserção do segmento SB_3 é mostrada na Figura 5.4.

[3]Nas figuras desta seção, a cor cinza é utilizada para diferenciar os segmentos visíveis da Estrutura Nominal do Discurso dos demais segmentos.

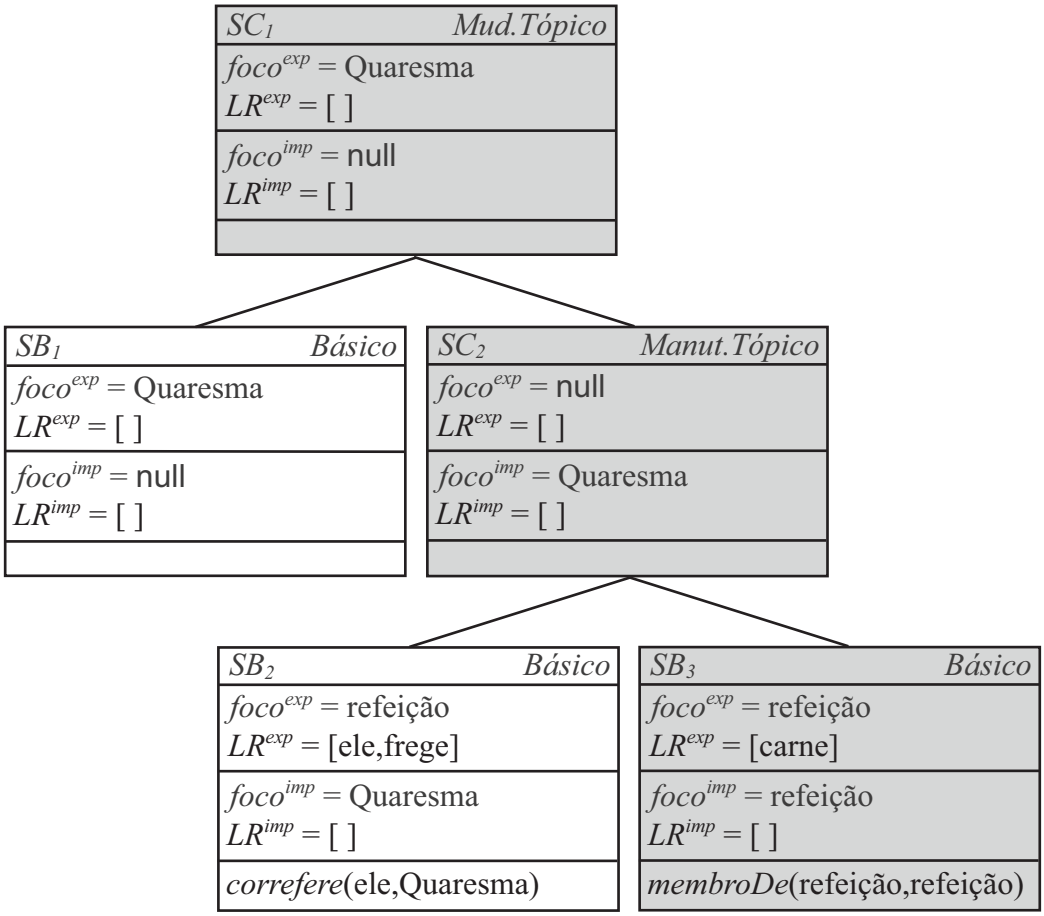


Figura 5.4: Estrutura após a interpretação da terceira frase.

A interpretação fora de contexto da quarta frase resulta nas seguintes entidades explícitas indexadas, considerando a ordenação dos termos: *"se"*, *"casa"*, *"destacamento"*, *"pavilhão"* e *"imperador"*, sendo *"se"* o foco explícito do segmento SB_4 relativo à frase. De acordo com a regra de identificação de um candidato à anáfora, todos os termos listados são possíveis referências – *"se"* é um pronome e os demais são substantivos precedidos por artigos definidos, remetendo a um SND. Sendo assim, durante o processo de interpretação em contexto, deve ser encontrado o segmento que permita o maior número de resoluções para essas entidades. A busca considera somente os segmentos visíveis da Estrutura Nominal e sempre parte do nó mais profundo. Observando a Figura 5.4, o primeiro segmento analisado é o SB_3, que fornece 1 (uma) possibilidade de resolução, através do relacionamento de correferência entre o termo anafórico *"casa"* e a entidade *"refeição"*. O segundo segmento visível, SC_2, permite quatro possibilidades de relações de correferência: dos termos *"se"*, *"destacamento"*, *"pavilhão"* e *"imperador"* (do segmento SB_4) com o foco implícito de SC_2 *"Quaresma"*. Analogamente, o último segmento visível SC_1 (a raiz da árvore) apresenta as mesmas possibilidades de relacionamentos, com a diferença de que o termo *"Quaresma"* no segmento antecedente é o foco explícito ao invés do implícito. O foco explícito possui prioridade sobre o foco implícito e, portanto, o segmento SC_1 é identificado como o ponto de interpretação para o segmento anafórico SB_4. A Figura 5.5 mostra a END obtida com a criação de um segmento composto SC_3, do tipo mudança de tópico, no ponto de interpretação determinado e com sua posterior anexação à estrutura.

Por fim, para a interpretação da última frase, foi escolhido o segmento SB_4 como antecedente. A Figura 5.6 apresenta a Estrutura Nominal do Discurso obtida após a inserção do último segmento SB_5, para os termos do qual foram encontradas relações de correferência. Vale ressaltar que nem todos os relacionamentos determinados pelo sistema entre termos anafóricos e antecedentes são condizentes com a semântica do texto. Na frase relativa ao segmento SB_5, por exemplo: *"Ficavam nela também a estação da estrada de ferro do Rio Douro e uma grande e bulhenta serraria."*, sabe-se que o termo *"ela"* (retirado da forma contraída *"nela"*) está referenciando a *"casa"* da frase anterior, fato que foi devidamente alcançado na estrutura obtida. Por outro lado, ao contrário do que a END da Figura 5.6 indica, os sintagmas nominais definidos *"estação"*, *"estrada de ferro"*, *"Rio Douro"* e *"serraria"* não são anafóricos nesses casos – estão inserindo informação ao texto, ao invés de estarem diretamente vinculados a conteúdos já apresentados.

Por outro lado, ao analisar as entidades referenciadas a partir de uma relação considerada errônea, observa-se que geralmente se tratam de termos relevantes no contexto e, mesmo que apresentem resultados negativos para a tarefa específica de resolução de

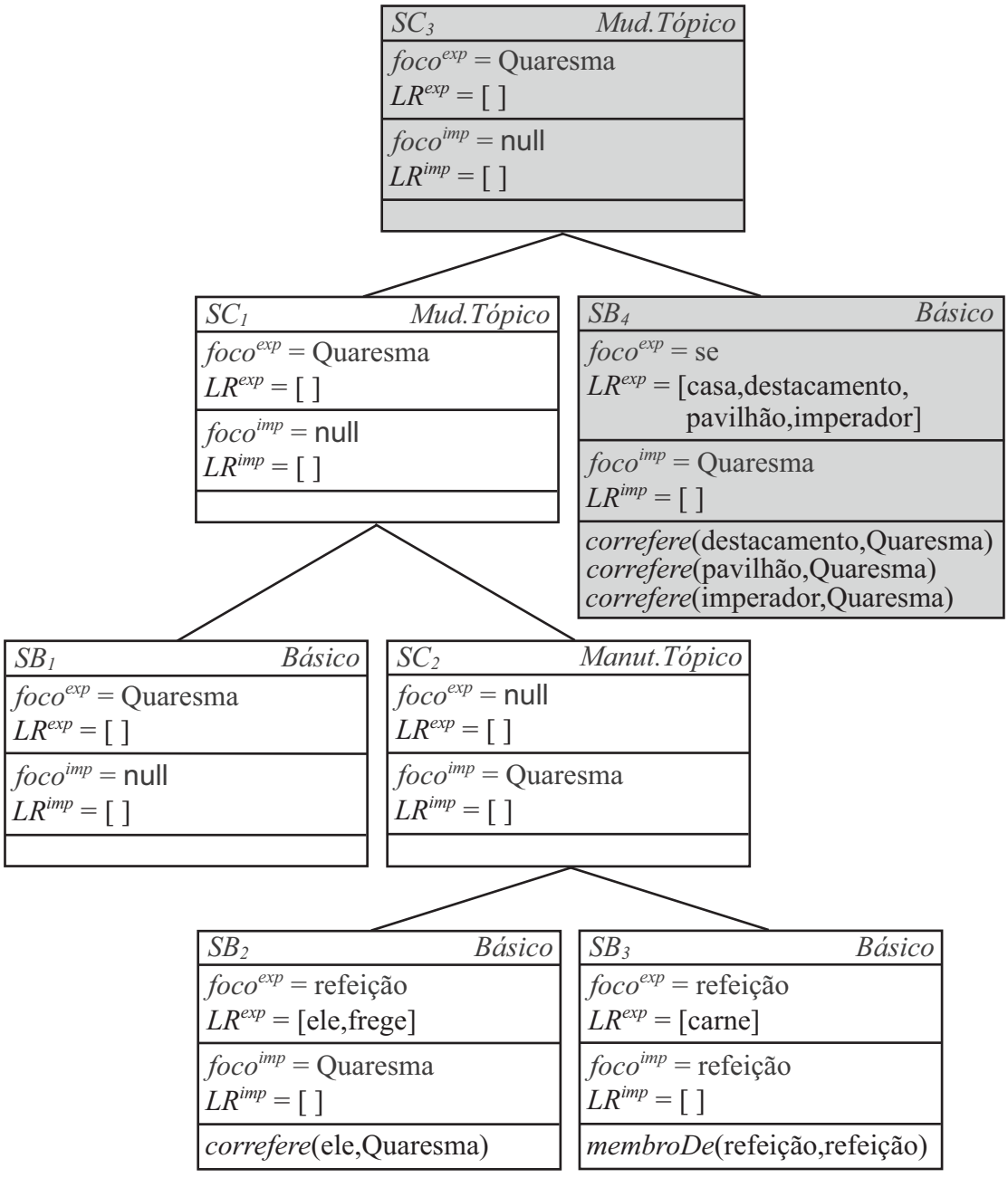


Figura 5.5: Estrutura após a interpretação da quarta frase.

anáforas, podem melhorar a performance qualitativa de um sistema de categorização. Isso ocorre porque, nesses casos, um assunto de interesse para o texto de uma forma geral está sendo reforçado ao longo dessas referências, mesmo que, em um contexto mais específico, esse relacionamento não faça sentido. Com isso, é embutido maior valor semântico ao texto para a identificação da classificação desse documento.

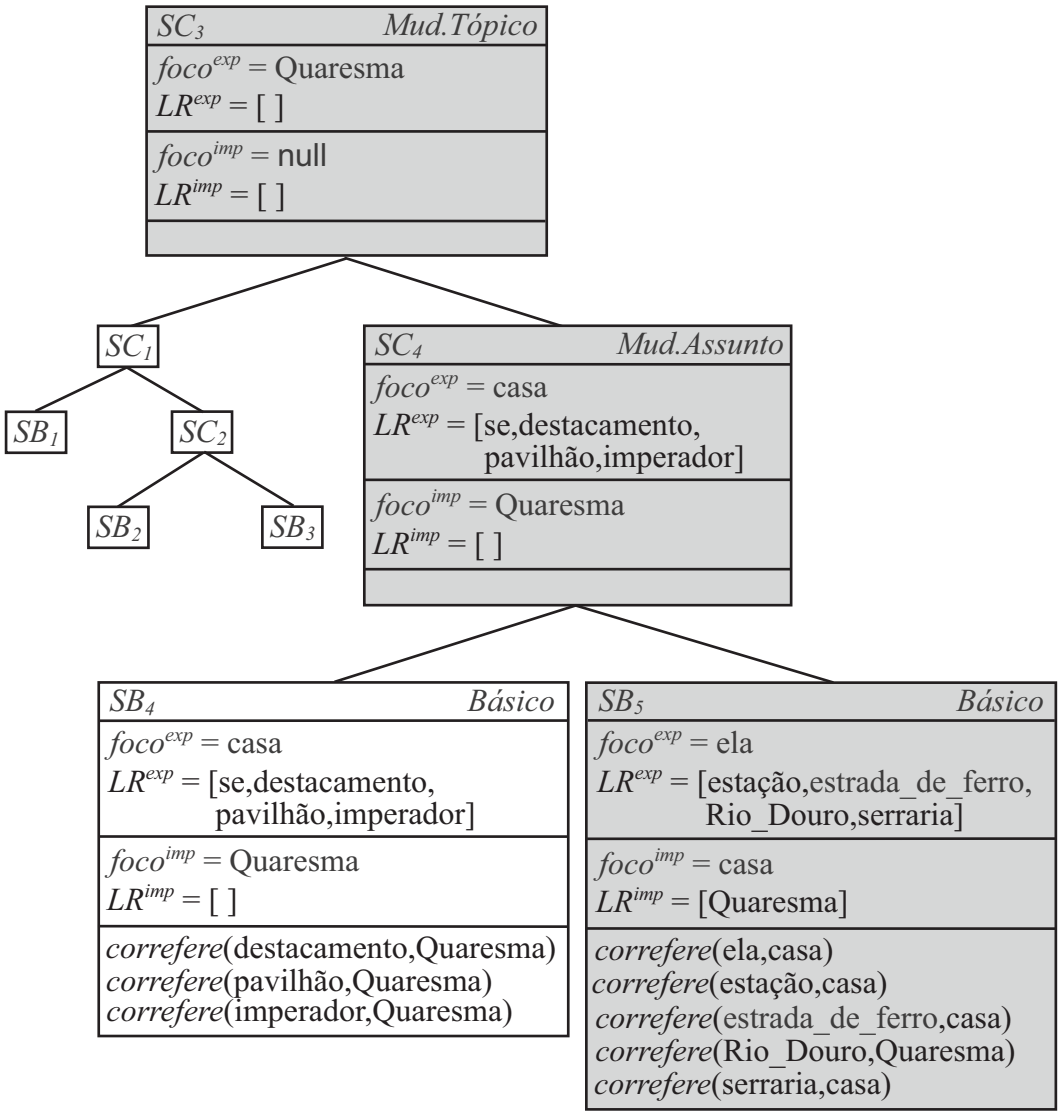


Figura 5.6: Estrutura após a interpretação da quinta e última frase.

Mantendo o exemplo do segmento SB_5, é sabido que os termos *"estação"*, *"estrada de ferro"*, *"Rio Douro"* e *"serraria"* não estabelecem referência a nenhum dos termos *"casa"* e *"Quaresma"*. Todavia, esses termos identificados como antecedentes apresentam grande valor semântico para a frase, uma vez que: *"Quaresma"* é, sem dúvida, o personagem principal do texto e deve ter sua relevância ressaltada; e *"casa"* pode ser considerada como um assunto secundário, mas também de suma importância, pois nas duas últimas frases do texto ela é utilizada para descrever o ambiente. No segmento obtido, essas

informações são capturadas e são dadas às entidades mais relevantes do texto sua devida importância.

Observando a END final mostrada na Figura 5.6, vale ressaltar que as anáforas do texto: *"ele"* (obtida do pronome *"lhe"* da segunda frase), *"refeição"* (presente na terceira frase) e *"ela"* (obtida da contração *"nela"* da última frase) foram resolvidas de forma correta. Na quarta frase *"Porque a casa em que se acantonara o destacamento (...)"*, sabe-se que o pronome *"se"* referencia ao termo *"destacamento"*. Essa relação, entretanto, não foi identificada pelo sistema, pelo seguinte motivo: o uso passivo do pronome *"se"* normalmente não faz sentido se utilizado como referência a um termo de uma frase já apresentada e, como anáforas intrafrases não são abordadas nesta proposta, esse tipo de pronome foi desconsiderado no processo de resolução de anáforas.

Além disso, pode-se observar que o termo *"Quaresma"* está presente em todos os segmentos da estrutura (seja em focos ou nas listas de relevantes), à exceção de um. Com isso, a conclusão de que esse termo é o principal do documento D é reiterada. Em segundo lugar, vem o termo *"casa"*, que aparece em três segmentos, seguido do termo *"refeição"*, presente em dois segmentos.

Ou seja, no exemplo do trecho da obra "Triste Fim de Policarpo Quaresma", a Estrutura Nominal gerada apresenta de forma correta as resoluções das anáforas presentes, e reforça a utilização de termos relevantes no contexto, inserindo maior valor semântico à sua representação. Essa característica tende a beneficiar a tarefa de categorização. Para exemplificar essa vantagem, assuma que exista uma categoria de nome "Lima Barreto", cujo objetivo é organizar documentos que tratem dos romances desse autor. Tendo conhecimento das seguintes obras de sua autoria: "Triste fim de Policarpo Quaresma", "Recordações do Escrivão Isaías Caminha", "Vida e Morte de M. J. Gonzaga de Sá" e "Clara dos Anjos", por exemplo, seria interessante definir termos presentes nesses títulos para o conjunto de palavras relacionadas dessa categoria, como: "Isaías", "Caminha", "Policarpo", "Quaresma", "Gonzaga de Sá", "Clara", "Anjos".

A partir dessas palavras relacionadas, o objetivo é avaliar a probabilidade, em termos genéricos, de o seguinte trecho – o mesmo considerado na construção da END – ser classificado na categoria "Lima Barreto":

> "*Quaresma jantava e almoçava ali mesmo. (...) As refeições eram-lhe fornecidas por um frege próximo (...).* A refeição principal sempre era carne. *Porque a casa em que se acantonara o destacamento, era o pavilhão do imperador*

(...). Ficavam nela também a estação da estrada de ferro do Rio Douro e uma grande e bulhenta serraria."

Avaliando o trecho puro, sem tratamento, observa-se que há uma única coincidência de entidades presentes no conjunto de palavras relacionadas da categoria com as palavras encontradas em seu conteúdo: o termo *"Quaresma"*, na primeira frase. Nesse caso, esse trecho apresentará uma relação bem reduzida com a categoria "Lima Barreto". Se, por outro lado, for considerada uma estrutura de representação para esse trecho, tal qual a definida na Figura 5.6, haverá uma similaridade muito maior com a categoria considerada, uma vez que o termo de maior destaque da END é *"Quaresma"*. Assim, com esse exemplo abstrato, é possível observar as vantagem das características da Estrutura Nominal do Discurso para categorização.

5.4 Experimentos em corpora

Esta seção mostra os testes realizados para a avaliação da metodologia de categorização proposta. Para a execução do método, foram utilizadas algumas ferramentas e dados externos, os quais são apresentados na seção 5.4.1. Nessa mesma seção, são definidos os corpora de documentos considerados nos experimentos e todas as assunções realizadas para a obtenção dos testes. Na seção 5.4.2, são apresentados, então, os experimentos executados.

5.4.1 Configuração do experimento

Para a execução dos testes apresentados neste capítulo, foram estipulados três corpus de documentos, retirados da coleção CHAVES (seção 5.4.1.1). São eles:

- D_1: composto de 30 documentos.
- D_2: composto de 68 documentos.
- D_3: composto de 148 documentos.

Os conjuntos de testes anteriores foram definidos dessa forma devido ao fato de o sistema trabalhar com o armazenamento do índice em memória, o que tornou-se um fator restritivo para o número de documentos que pôde ser utilizado no experimento. Essa limitação ocorre principalmente na fase de *bootstrapping* do categorizador, que define

uma série de operações sobre cada termo presente na base de dados e cada documento considerado.

Porém, em comparação com um sistema de recuperação de informação, como o desenvolvido por Pereira [Pereira 2009], a performance computacional em sistemas de categorização é menos restritiva. Isso porque o objetivo é a criação de um modelo de categorização, a partir do qual a classificação de novos documentos é realizada. O processo mais custoso é justamente a geração do modelo, mas essa etapa normalmente é executada somente uma vez ou, dependendo da aplicação, de tempos em tempos. Fornecido o modelo, a associação de novos documentos é executada de forma mais eficiente. Evidentemente, o sistema desenvolvido neste trabalho é um protótipo e o objetivo é avaliar especificamente a etapa de geração do modelo; devido a isso, para tornar factível a execução dos vários experimentos (apresentados na seção 5.4.2), foram assumidos os conjuntos de documentos citados.

Os próximos tópicos desta seção apresentam as ferramentas e metodologias externas utilizadas para a obtenção do categorizador proposto neste trabalho. São mostradas as definições, particularidades e assunções relativas à sua utilização. A seção 5.4.1.1 apresenta a coleção de documentos utilizada para os testes, juntamente com o *tagger* considerado; a seção 5.4.1.2 mostra informações sobre o dicionário de sinônimos utilizado para a expansão do conjunto de palavras relacionadas; e a seção 5.4.1.3 disserta sobre as principais características do sistema de categorização usado para a geração do modelo final.

5.4.1.1 A coleção CHAVE e o *tagger* PALAVRAS

O corpus CHAVE [CHAVE] contém textos jornalísticos do jornal português Público e da Folha de São Paulo dos anos de 1994 e 1995. A coleção disponibiliza o texto integral de cada uma das matérias do jornal, um identificador para o texto e uma categorização referente aos cadernos existentes no jornal (à exceção da coleção da Folha de 1994, que não apresenta uma categoria relacionada para cada texto). Além do texto das matérias disponibilizado em texto plano, a coleção também fornece uma versão dos textos marcada pelo PALAVRAS.

O PALAVRAS, desenvolvido por Bick [Bick 2000], é um *POS-tagger* (*Part-Of-Speech tagger*) baseado em uma gramática restritiva e capaz de identificar a estrutura sintática de uma frase. Ele atribui etiquetas morfológicas para palavras e sinais de pontuação, além de determinar o lema de cada palavra. O lema de uma palavra é dado pela sua forma

canônica: no caso de verbos, o lema é representado pela sua forma infinitiva, e no caso de um termo nominal, é dado por sua forma singular e masculina (quando existente). Essas informações são necessárias para a construção da Estrutura Nominal do Discurso, conforme discutido no capítulo 4. Em [PALAVRAS] são apresentados os significados de cada símbolo utilizado no *tagger*, para os quais são mostrados exemplos de aplicação em frases escritas em português.

Para a realização do experimento, foi utilizada a coleção do jornal Folha de 1995. Os fatores levados em consideração para essa decisão são os seguintes:

- a coleção da Folha está disponibilizada em português brasileiro, e não em português de Portugal – assim, os testes ficam mais próximos da nossa realidade;
- os textos são apresentados em uma versão já marcada, o que agiliza o desenvolvimento do trabalho, pois elimina a necessidade da construção de um *tagger* para a execução do experimento;
- o corpus apresenta a informação da classe de cada documento, com base nas categorias referentes aos cadernos do jornal, tornando possível a avaliação do categorizador obtido neste trabalho.

As categorias existentes no corpus utilizado são: "Agrofolha", "Brasil", "Caderno Especial – Anos FHC", "Cotidiano", "Dinheiro", "Empregos", "Esporte", "Folhateen", "Folhinha", "Fovest", "Ilustrada", "Imóveis", "Informática", "Mais!", "Mundo", "Opinião", "Primeira Página", "Revista da Folha", "Tudo", "Turismo", "TV Folha", "Veículos", além de uma série de outros cadernos especiais. Todavia, muitas dessas categorias apresentam um sentido ambíguo ou vago, como por exemplo a categoria "Primeira Página", que pode abranger conteúdo originalmente de qualquer outra categoria, ou a categoria de nome "Mais!", cujo significado não está claro. Sendo assim, essas categorias consideradas imprecisas foram descartadas para os testes.

Assumindo que as categorias "Caderno Especial – Anos FHC" e "TV Folha" podem ser reescritas como somente "FHC" e "TV", respectivamente, e considerando a forma canônica dos nomes das mesmas, ao fim, as categorias de interesse se resumem a nove: "Brasil", "Dinheiro", "Emprego", "Esporte", "FHC", "Imóvel", "Informática", "TV" e "Veículo".

5.4.1.2 O dicionário de sinônimos TeP 2.0

Para a expansão do conjunto de palavras relacionadas, de acordo com o que foi proposto no capítulo 3, é utilizado um dicionário de sinônimos. Para isso, foi estipulado o *Thesaurus*[4] Eletrônico Básico para o Português do Brasil, em sua segunda versão – TeP 2.0 [Maziero Thiago A. S. Pardo 2008] –, desenvolvido segundo as diretrizes da WordNet de Princeton [Stark e Riesenfeld 1998]. Ele é estruturado em entradas indexadas, cada uma contendo um conjunto de sinônimos e um rótulo que explicita a classe gramatical de seus elementos. Uma entrada pode conter, ainda, o índice de outra entrada, caso seus conjuntos expressem sentidos distintos.

Uma palavra em sua forma canônica, se pertencente à base léxica do *thesaurus*, irá possuir tantas acepções quanto conjuntos de sinônimos que a contêm. Os demais elementos do conjunto serão seus sinônimos naquela acepção. Atualmente, o TeP 2.0 contém 19888 conjuntos de sinônimos e 44678 unidades lexicais, tendo a média de 2.5 unidades por conjunto de sinônimos.

5.4.1.3 O categorizador *Rainbow*

Para a geração do modelo de categorização a partir de um categorizador supervisionado, foi utilizada uma ferramenta desenvolvida por McCallum e distribuída livremente pela Internet. *Bow* [McCallum] é um *kit* de ferramentas implementado na linguagem de programação C, visando problemas de linguagem de modelagem estatística, recuperação de informação, classificação e *clustering*. A distribuição inclui a biblioteca e os códigos-fonte para três abordagens distintas:

- *Rainbow*, para categorização de documentos;
- *Arrow*, para recuperação de informação; e
- *Crossbow*, para *clustering* de documentos.

Rainbow foi a ferramenta utilizada para a execução dos experimentos deste trabalho. O padrão geral para sua utilização pode ser definido em dois passos: (1) ler os documentos de entrada e escrever em disco um modelo contendo suas estatísticas, e (2) usar o modelo gerado para realizar a classificação e/ou diagnósticos sobre os resultados. Do segundo

[4] *Thesaurus* é uma palavra latina que significa "tesouro" e foi empregada, a partir de 1500, para indicar um acervo ordenado de informações e conhecimentos. Aqui, é utilizada para designar um dicionário de sinônimos.

passo, conclui-se que o próprio *Rainbow* fornece meios para o cálculo de métricas capazes de avaliar o categorizador obtido. Todos os resultados apresentados nos experimentos são baseados nessa medida de precisão.

O algoritmo de categorização padrão do *Rainbow* é o *naive Bayes*; além desse, várias outras implementações estão disponíveis, como: uma versão simplificada do *naive Bayes*, *Support Vector Machines*, *k-Nearest Neighbor*, TFIDF/*Rocchio*, *Maximum Entropy*, *Fuhr's Probabilistic Indexing*, entre outros.

Neste ponto, vale uma observação. O *Rainbow* requer como entrada os documentos rotulados que, em um processo de categorização tradicional, representa a rotulação definida manualmente por um especialista (chamada de *verdadeira*). Essa rotulação é a mesma utilizada como base para o cálculo da métrica de avaliação na fase de teste do categorizador. Entretanto, considerando o problema atual, a rotulação inicial dos documentos passada para o *Rainbow* não equivale à verdadeira, e sim, ao resultado do processo de *bootstrapping* do método. Devido a isso, foram necessárias algumas alterações nos fontes do *Rainbow* visando tratar essa questão. Foi definido um novo parâmetro de execução para a definição da rotulação verdadeira, de forma que o método seja treinado a partir dos documentos de entrada (marcados pelo processo de *bootstrapping*), mas seja avaliado a partir da rotulação verdadeira passada como parâmetro.

A ferramenta possui alguns parâmetros para sua execução. A seguir, os principais são mostrados, juntamente com a escolha estipulada para os testes deste trabalho:

- método de categorização: foi escolhido o categorizador *naive Bayes* para os experimentos, pois se trata de um método simples, que apresenta bons resultados;

- parcela de treino/teste: deve ser fornecida à ferramenta a porcentagem de documentos a serem considerados como conjunto de teste. O *Rainbow* define aleatoriamente esse conjunto, fixando o restante dos documentos para o conjunto de treino. Nos testes, é considerada a taxa de 30% dos documentos para o conjunto de teste.

- número de execuções: como a escolha do conjunto de teste é obtida aleatoriamente, pode haver alterações nos resultados de uma execução para outra. Assim, foi definido um número de 100 execuções para cada experimento, sendo o resultado final apresentado como uma média dos resultados obtidos em cada execução.

5.4.2 Avaliação empírica

Para a avaliação do método de categorização não supervisionada proposto neste trabalho, foram realizados experimentos visando a estimativa de parâmetros internos do algoritmo, comparação com um método supervisionado de respaldo na literatura e avaliação no que diz respeito à utilização do processo de resolução de anáforas na categorização. Os próximos itens desta seção apresentam os resultados e considerações a respeito de cada um desses testes.

Como os três conjuntos de documentos considerados nos testes são pertencentes à mesma base, as categorias de D_1, D_2 e D_3 são as mesmas. Assim, independente do experimento, as palavras relacionadas e o conjunto expandido para cada categoria já podem ser definidos. A Tabela 5.1 mostra as cinco palavras relacionadas que foram estipuladas manualmente para cada categoria.

Categoria	Palavras relacionadas
Brasil	brasília, presidente, governo, rio de janeiro, são paulo
Dinheiro	preço, comércio, empresa, bolsa, valor
Emprego	trabalho, mercado, funcionário, currículo, profissão
Esporte	futebol, vôlei, técnico, clube, jogador
FHC	fernando henrique cardoso, presidente, governo, governador, posse
Imóvel	casa, apartamento, proprietário, prédio, aluguel
Informática	computador, internet, dispositivo, programa, sistema
TV	televisão, emissora, novela, jornal, ator
Veículo	carro, pneu, motor, concessionária, autopeça

Tabela 5.1: Palavras relacionadas estipuladas para cada categoria.

Na Tabela 5.2 são mostrados cinco termos extraídos do conjunto expandido de palavras relacionadas. Não foram listados todos os termos do conjunto, pois esse número passou de 20 para algumas categorias.

A maioria dos termos considerados no conjunto expandido de palavras relacionadas apresenta conteúdo semântico de interesse para a categoria. No entanto, há casos em que os sinônimos obtidos para uma dada palavra apresentam acepções diferentes do sentido pretendido. Considere como exemplo o termo *"governo"*, que foi estipulado como palavra relacionada para as categorias "Brasil" e "FHC". No TeP 2.0, foram identificadas as seguintes acepções para esse termo:

1. {estado, governo}
2. {administração, governança, governo, regência}
3. {governação, governo, regime, regímen, regimento}

4. {direção, governo, norteamento, norteio, orientação}
5. {controle, domínio, governo}

Categoria	Conjunto expandido de palavras relacionadas
Brasil	estado, presidência, administração, controle, regime
Dinheiro	negócio, mercado, moeda, capital, custo
Emprego	serviço, carreira, contratação, responsabilidade, tarefa
Esporte	desporto, grêmio, associação, agremiação, sociedade
FHC	estado, orientação, administração, controle, regime
Imóvel	edifício, locação, moradia, estabelecimento, construção
Informática	aplicativo, código, empresa, norma, método
TV	televisor, artista, noticiário, gazeta, salário
Veículo	automóvel, máquina, transporte, condutor, móvel

Tabela 5.2: Termos pertencentes ao conjunto expandido de palavras relacionadas para cada categoria.

As acepções 1, 2 e 3 são diretamente relacionadas ao contexto das categorias "Brasil" e "FHC". Por outro lado, as acepções 4 e 5 apresentam um sentido que foge à ideia de estado e poder executivo que ambas categorias sugerem. Infelizmente, para filtrar somente as acepções de interesse, seria necessária uma estruturação mais robusta no processo de definição das palavras relacionadas, como por exemplo, estipular não só um termo relacionado, mas também o sentido (acepção) pretendido. Isso requereria uma atuação mais elaborada do engenheiro de conhecimento nessa fase, o que foge ao objetivo deste trabalho.

As seções seguintes apresentam os experimentos realizados visando avaliar a performance do método de categorização proposto.

5.4.2.1 Variação do número de palavras-chave

Na seção 3.5.1, foi apresentado um método para a obtenção de palavras-chave baseado em padrões de coocorrência entre as palavras relacionadas e os termos presentes nos documentos. Ao final do processo o conjunto de palavras-chave encontra-se ordenado, estando mais bem colocados os termos considerados mais relevantes, de acordo com a métrica utilizada. Entretanto, para cada categoria é gerado um número grande de termos; para o menor corpus considerado (D_1), por exemplo, foi gerada uma média de aproximadamente 714 palavras-chave para cada classe. Obviamente, a utilização desse conjunto completo é inviável. Assim, devem ser selecionadas k palavras-chave para cada categoria, sendo esse valor de k um parâmetro para o sistema.

Para a definição desse valor, foi realizado o seguinte teste. Variou-se k de 0 a 20 para os três conjuntos de documentos D_1, D_2 e D_3, permitindo a comparação entre os resultados obtidos. A Figura 5.7 mostra essa configuração.

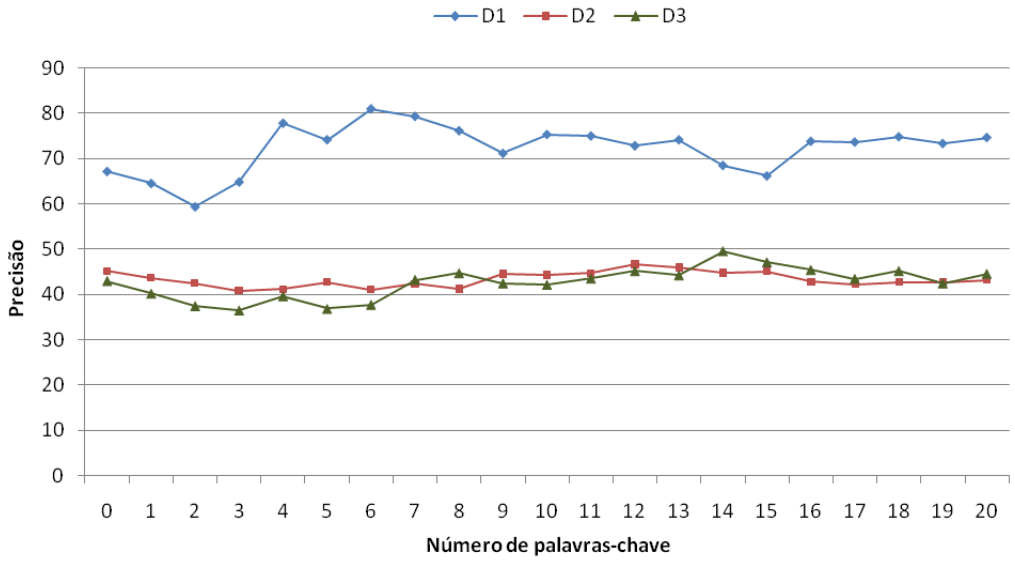


Figura 5.7: Comparação da performance de acordo com o número de palavras-chave.

Como pode ser visto na Figura 5.7, os melhores valores para k alcançados pelos corpus D_1, D_2 e D_3 foram, respectivamente: 6, 12 e 14. Valores de k posteriores a esse máximo geram precisão inferior devido aos ruídos causados na classificação. Porém, esses ruídos tornam-se menos impactantes à medida que o número de documentos cresce – para corpus maiores, a utilização de mais palavras-chave torna-se cada vez mais benéfica.

As cinco primeiras palavras-chave geradas para cada categoria, considerando o corpus D_3, são mostradas na Tabela 5.3.

Categoria	Palavras-chave
Brasil	fernando henrique cardoso, sucursal de brasília, R$, país, itamar franco
Dinheiro	ação, CUB, mês, programa, renda
Emprego	indústria, legislação, tecnologia bancária, hora, empresa
Esporte	patriarca, ação, londrina, matsubara, fórmula
FHC	país, ação, ano, sucursal de brasília, PSDB
Imóvel	inquilino, valor, mês, empresa, contrato
Informática	usuário, empresa, microsoft, sistema operacional, SCSI
TV	globo, audiência, janete clair, surpresa, irmãos coragem
Veículo	motivo, análise, goodyear, desgaste, fiat

Tabela 5.3: Cinco primeiras palavras-chave determinadas para cada categoria.

O que se pode notar pela Tabela 5.3 é que, de forma geral, os termos são específi-

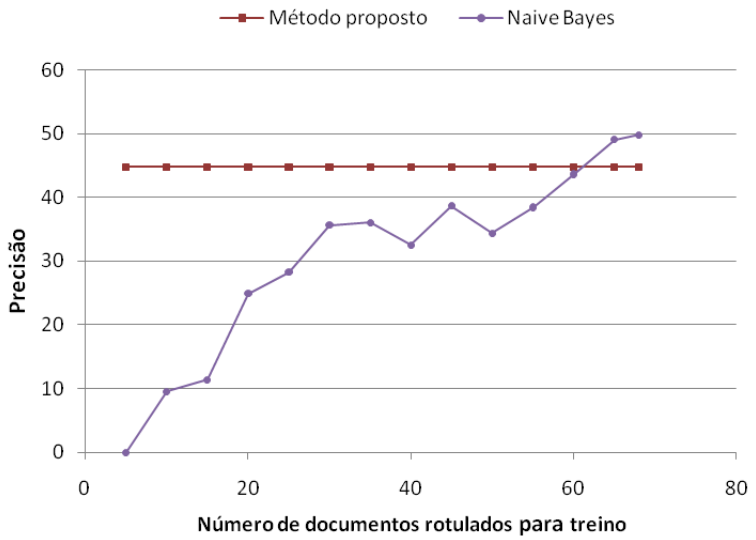
cos do corpus considerado, como já era previsto. Para o conjunto de documentos D_2, por exemplo, as cinco primeiras palavras-chave obtidas para a categoria "Veículo" foram: combustível, máquina, pacote, injeção e mão – nenhum termo repetido do corpus D_3. Essa especificidade possui dois pontos de vista no que diz respeito ao efeito na categorização. Pode ser benéfica quando o termo, apesar de ser específico da base, apresenta um sentido mais amplo, como acontece com *"globo"* e *"audiência"*, na categoria "TV", por exemplo. Nessa mesma categoria, entretanto, há dois termos que provavelmente têm grande ocorrência, mas em poucos documentos: *"janete clair"* e *"irmãos coragem"*. Nesse caso, esses termos não apresentam poder discriminativo para a escolha da categoria à qual um dado documento pertence, à exceção das poucas instâncias onde ocorrem.

5.4.2.2 Comparação com um método supervisionado

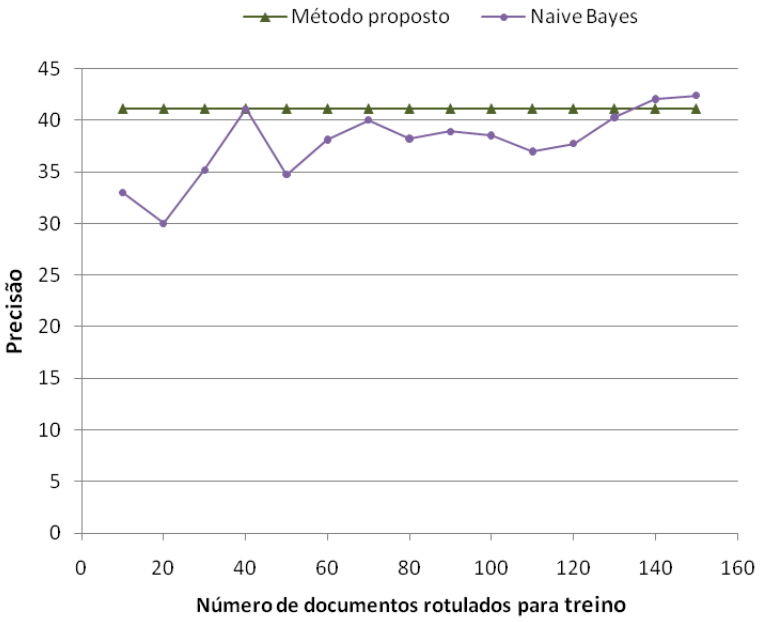
Um algoritmo de classificação supervisionada, considerando uma base completa de treino, certamente apresenta uma categorização mais precisa do que o método de *bootstrapping* proposto. Entretanto, quanto menor o número disponível de documentos rotulados para o treino do categorizador, a performance desse algoritmo tende a diminuir. O objetivo desta seção é observar a quantidade de documentos rotulados necessária para que o método supervisionado atinja a mesma performance do método aqui proposto.

A Figura 5.8 mostra os resultados do algoritmo *naive Bayes*, obtidos através do *Rainbow*, quando o número de documentos de treino manualmente rotulados é variado. Os valores do método proposto estão dispostos em uma reta horizontal, porque ele independe do tamanho do conjunto de treino, uma vez que não são considerados documentos rotulados para sua concepção. A Figura 5.8(a) apresenta a configuração para o corpus D_2 e a Figura 5.8(b), para o corups D_3. Não foram apresentados os resultados para o corpus D_1, pois esse teste perderia o sentido para um número reduzido de documentos.

Analisando a Figura 5.8, observa-se que, de forma geral (à exceção de alguns pontos), quanto maior o número de documentos rotulados, melhor a performance do categorizador supervisionado. O algoritmo *naive Bayes* atinge uma precisão maior do que o método proposto entre 60 e 65 documentos rotulados para o corpus D_2, e entre 130 e 140 para o corpus D_3. É necessária, portanto, uma média de rotulação de aproximadamente 91% dos documentos do corpus considerado para que o categorizador supervisionado alcance o mesmo resultado que o método de *bootstrapping* – qualquer percentual acima desse resulta em uma melhor classificação. Contudo, essa taxa ainda é muito alta; o esforço requerido para a rotulação manual de 91% de um dado conjunto de documentos continua sendo um



(a) Conjunto de documentos D_2.



(b) Conjunto de documentos D_3.

Figura 5.8: Performance do categorizador *naive Bayes* para diferentes tamanhos do conjunto de treino, em comparação com o método proposto, para os corpus (a) D_2 e (b) D_3.

fator restritivo para muitas aplicações.

Além disso, observa-se que a precisão do algoritmo supervisionado para o conjunto completo de treino não apresenta muita diferença em relação ao método proposto. Para o corpus D_2, foi observada uma melhora de aproximadamente 11% nos resultados, e para o D_3, de somente 3%. Dependendo da aplicação, esse ganho é desprezível se comparado com a vantagem da não necessidade de um corpus rotulado.

5.4.2.3 Utilização das anáforas resolvidas na categorização

Na seção 5.2 foi mostrado o processo geral do método proposto, no qual foi especificado que, durante a construção da Estrutura Nominal, são gerados os documentos tendo as anáforas presentes em seu conteúdo resolvidas. Para isso, podem ser consideradas duas possíveis formas para sua representação, no que diz respeito à permanência ou não dos termos anafóricos no conteúdo resultante. Por exemplo: se a anáfora é pronominal, muito provavelmente o termo anafórico não apresenta valor semântico de interesse e, portanto, nesse caso sua utilização pode ser tranquilamente descartada; se, por outro lado, a anáfora provém de um sintagma nominal definido, pode ocorrer a situação em que uma informação relevante para o assunto seja introduzida através da entidade anafórica. Nesse último caso, seria um desperdício, em termos de semântica, desconsiderar esse termo.

Para exemplificar, considere o seguinte texto:

> "*Matheus não esquece os nomes de seus pacientes.*" (5.1a)
>
> "***Ele** é sempre muito atencioso.*" (5.1b)
>
> "***O médico** é sempre muito atencioso.*" (5.1c)

A frase (5.1a) introduz a entidade *"Matheus"*, que é referenciada nas duas opções de frases seguintes. Em (5.1b), é utilizado o pronome *"Ele"* para o relacionamento. Esse termo não apresenta conteúdo semântico suficiente para interferir na categorização. Ao contrário da entidade apresentada na frase (5.1c), que introduz uma nova informação sobre o elemento referenciado, o que pode ser útil na geração de conhecimento geral do texto e, portanto, beneficiar a tarefa de categorização.

Para avaliação, foram estipulados quatro modos de execução, incluindo as duas abordagens citadas, que diferem na forma como o conteúdo dos documentos é apresentado. São eles:

- Modo 1: texto puro em linguagem natural;
- Modo 2: texto contendo somente os termos indexados na END (anafórico ou não), sem considerar os antecedentes identificados para as anáforas;
- Modo 3: texto contendo os termos indexados na END, sendo as anáforas substituídas pelos antecedentes (e eliminadas);
- Modo 4: texto contendo os termos indexados na END, apresentando tanto as anáforas quanto os antecedentes.

Cada um desses modos apresenta diferentes termos no conteúdo dos documentos e é interpretado no categorizador supervisionado de forma distinta. Sendo assim, dependendo do algoritmo de classificação utilizado, os resultados podem apresentar comportamentos diversos. Para analisar essa variação, foram testados cinco métodos disponibilizados pelo *Rainbow*: *naive Bayes* (NB), versão simplificada do *naive Bayes* (NB-Simple), TFIDF/*Rocchio* (TFIDF), TFIDF *Words* (TFIDF-Words) e *Fuhr's Probabilistic Indexing* (Prind).

A Tabela 5.4 apresenta os resultados obtidos para os corpus D_1, D_2 e D_3, considerando os cinco algoritmos de categorização e os quatro modos de execução definidos. Os melhores resultados de cada método estão destacados em negrito.

Algoritmo	Corpus	Modo 1	Modo 2	Modo 3	Modo 4
NB	D_1	74.37	75.36	74.59	**76.90**
	D_2	44.31	43.71	45.50	**46.49**
	D_3	39.37	**40.16**	38.98	39.95
NB-Simple	D_1	75.14	**76.57**	75.14	74.59
	D_2	46.39	**46.88**	45.54	45.69
	D_3	39.47	39.47	39.67	**39.97**
TFIDF	D_1	75.03	75.03	76.13	**77.12**
	D_2	44.16	45.30	45.20	**45.50**
	D_3	39.30	**40.00**	39.58	39.53
TFIDF-Words	D_1	74.59	76.90	75.03	**77.67**
	D_2	45.50	44.11	**46.73**	45.40
	D_3	39.10	39.47	39.26	**39.53**
Prind	D_1	75.69	75.14	76.02	**76.13**
	D_2	45.54	**46.49**	45.79	45.74
	D_3	39.05	38.96	39.40	**39.53**

Tabela 5.4: Comparação entre os resultados obtidos através dos quatro modos de execução.

De acordo com a Tabela 5.4, percebe-se que a utilização de termos indexados é vantajosa em relação ao texto puro em linguagem natural – 8 das 15 combinações algoritmo-corpus obtêm melhores resultados nos modos 2, 3 e 4, em relação ao modo 1; além disso, nenhum caso apresenta como melhor resultado o modo 1. Comparando a classificação de documentos com texto puro, com a classificação com termos indexados (modo 2), observa-se uma melhora em 11 dos 15 casos.

Dentre os três modos que utilizam termos indexados, é visto que, de forma geral, os melhores resultados são obtidos através do modo 4. Em segundo lugar, o modo 2 e, por fim, o modo 3. O modo 2 não considera de forma alguma os antecedentes determinados pela resolução anafórica; comparando-o com o modo 4, percebe-se uma melhora na maioria dos casos, indicando que a utilização dos antecedentes, além dos termos anafóricos, é benéfica. Por outro lado, ao confrontar os resultados dos modos 2 e 3, repara-se que 9 dos 15 resultados apresentam valores inferiores no modo 3 – esse fato fornece o indício de que utilizar os antecedentes, em substituição aos termos anafóricos, não é uma boa abordagem.

Em justificativa a essa última informação, há uma característica do processo de resolução de anáforas que deve ser observada: muitos sintagmas nominais definidos são identificados como anafóricos erroneamente, como foi visto na seção 5.3. Para esses termos, são determinados antecedentes que, mesmo apresentando informação de interesse para o contexto geral do documento, não se referem diretamente ao SND em questão. Sabe-se que SNDs normalmente introduzem informação ao texto; descartá-los, portanto, quando não há um relacionamento direto com o antecedente reconhecido, certamente produz resultados inferiores – dessa afirmativa, compreende-se o fato de o modo 3 apresentar piores resultados que o modo 2. Apesar disso, considerando que não são descartados os termos anafóricos, a utilização dos antecedentes ainda é benéfica para o processo de categorização, como pode ser comprovado pelos melhores resultados obtidos no modo 4 em relação ao modo 2.

Portanto, foi visto que o processo de resolução de anáforas estabelece ganhos à tarefa de classificação de textos. A utilização dos termos reconhecidos como antecedentes é útil nesse processo, mas sem necessariamente desprezar a informação introduzida pelos termos identificados como anafóricos pelo sistema.

6 Conclusões e trabalhos futuros

Neste capítulo são apresentadas as conclusões deste trabalho e algumas propostas para trabalhos futuros.

6.1 Conclusões e trabalhos futuros

A maioria das metodologias de categorização de textos encontradas na literatura é apoiada no paradigma supervisionado que, para a construção do classificador, requer um conjunto de instâncias previamente rotuladas. Em casos em que essa rotulação não está disponível *a priori*, o processo para sua obtenção é altamente custoso, uma vez que essa tarefa deve ser feita manualmente por um especialista de domínio. Nessas situações, convém utilizar procedimentos que independam de tal restrição – métodos não supervisionados. Este trabalho, visando oferecer uma solução para esse tipo de aplicação, propôs uma metodologia de categorização não supervisionada, para a qual somente são fornecidos como entrada documentos não rotulados e as categorias de interesse.

O processo geral para a obtenção do método foi baseado em duas etapas: uma fase de *boostrapping* que, a partir da definição de um conjunto de palavras características para cada categoria, associou rótulos para os documentos; e a fase de geração do modelo de categorização, que utilizou a rotulação retornada pelo processo de *boostrapping* em um classificador supervisionado.

A solução proposta foi fundamentada em uma abordagem que considera fatores linguísticos em sua formação. Foi utilizada a Estrutura Nominal do Discurso desenvolvida por Freitas em [Freitas 2005], com o propósito de resolução de anáforas. A partir da END, somente foram indexados termos com alto valor semântico; além disso, no processo de categorização, foram considerados os antecedentes reconhecidos para as entidades anafóricas. Para a implementação, foram apresentadas melhorias em relação ao sistema original desenvolvindo por Pereira [Pereira 2009], visando aproximar ao máximo – sem fugir do limite computacional – da teoria de Freitas.

Com o intuito de avaliar a solução proposta, foi apresentado um exemplo de execução, mostrando passo-a-passo o processo de criação da END, e também os experimentos em corpora realizados para avaliar o método de classificação proposto. A partir da comparação com o algoritmo supervisionado *naive Bayes*, foi observado que são necessários muitos documentos de treino para que o paradigma supervisionado atinja a performance do método proposto. Além disso, mesmo utilizando o corpus de treino completo, a diferença entre os dois métodos foi desprezível, levando em consideração as vantagens do método não supervisionado. Por fim, foi feita uma avaliação no que diz respeito à utilização do processo de resolução de anáforas na categorização de textos. Chegou-se à conclusão de que considerar somente os termos indexados pela END produz melhores

resultados do que considerar o texto puro em linguagem natural; e, mais do que isso, introduzir os antecedentes reconhecidos, além dos termos anafóricos indexados, gera uma melhor classificação.

Os experimentos apresentados mostraram uma forma para a avaliação da qualidade do método de um modo geral, mas seria interessante uma comparação com algum outro método da literatura que siga o mesmo princípio de categorização não supervisionada a partir de um processo de *bootstrapping*, como em [Ko e Seo 2009], [Mccallum e Nigam 1999] e [Gliozzo, Strapparava e Dagan 2009], por exemplo. Entretanto, a maioria dos métodos encontrados foram propostos para o idioma inglês, não sendo possível uma comparação direta com os resultados apresentados – nenhum processo semelhante para o português foi identificado na literatura. Para tornar essa comparação possível, uma solução futura seria a implementação de um desses métodos que apresente bons resultados, e a aplicação do mesmo em um corpus da língua portuguesa, realizando os ajustes necessários.

Foram inseridos à geração da END novos procedimentos e características, visando uma aproximação maior com a teoria de Freitas. Alguns conceitos não foram implementados devido à sua complexidade; entretanto, sua utilização tornaria processo de resolução de anáforas mais preciso, beneficiando também o método de categorização proposto. Por exemplo, este trabalho não tratou a resolução de elipses, uma figura de linguagem recorrente no idioma português. A identificação dos seus antecedentes tende a inserir ainda mais semântica ao contexto, além do que foi obtido com as anáforas pronominais e ANDs. Uma outra informação que traria benefícios é a utilização das relações de "parte de" e "subcategorizado por", visando abranger ainda mais os relacionamentos possíveis entre anáforas e antecedentes.

O sistema de categorização desenvolvido não apresentou como foco o desempenho computacional. Como um trabalho futuro, o protótipo implementado pode ser ampliado, firmado em uma estruturação de índices mais robusta, visando bases de dados maiores. Pode ser criado um sistema automatizado de categorização, a partir do qual seja possível a identificação de categorias para novas instâncias que surjam. Com isso, o sistema pode ser aplicado para bases de documentos online, como os existentes em sites de conteúdo e bibliotecas digitais, por exemplo, ou até mesmo em bancos de dados locais que exijam uma organização em categorias de interesse.

Referências

[Adami, Avesani e Sona 2003]ADAMI, G.; AVESANI, P.; SONA, D. Bootstrapping for hierarchical document classification. In: *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*. New York, NY, USA: ACM, 2003. p. 295–302. 23, 26, 46

[Adami, Avesani e Sona 2005]ADAMI, G.; AVESANI, P.; SONA, D. Clustering documents into a web directory for bootstrapping a supervised classification. *Data Knowl. Eng.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 54, n. 3, p. 301–325, 2005. 13, 26

[Bick 2000]BICK, E. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Tese (Doutorado) — Arthus University, Arthus, 2000. 59, 75, 87

[Blum e Mitchell 1998]BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*. New York, NY, USA: ACM, 1998. p. 92–100. 25

[CHAVE]CHAVE. A colecção chave. Http://www.linguateca.pt/chave/. 87

[Chaves e Rino 2008]CHAVES, A. R.; RINO, L. H. The mitkov algorithm for anaphora resolution in portuguese. In: *PROPOR '08: Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*. Berlin, Heidelberg: Springer-Verlag, 2008. p. 51–60. 30

[Dahl e Fraurud 1996]DAHL Östen; FRAURUD, K. Animacy in grammar and discourse. In: REFERENCE; ACCESSIBILITY, R. (Ed.). *In Thorstein Fretheim and Jeanette K. Gundel (eds.)*. Amsterdam/Philadelphia: [s.n.], 1996. 37

[El-Yaniv e Souroujon 2001]EL-YANIV, R.; SOUROUJON, O. Iterative double clustering for unsupervised and semi-supervised learning. In: *In Advances in Neural Information Processing Systems (NIPS*. [S.l.: s.n.], 2001. p. 121–132. 23

[FORMA]FORMA. Ferramenta forma: etiquetagem morfológica e lematização. Http://www.inf.pucrs.br/ gonzalez/tr+/forma/. 59

[Freitas 1992]FREITAS, S. A. A. de. A utilização da drt em um sistema de representação do discurso. In: *IX Simpósio Brasileiro de Inteligencia Artificial*. São Carlos-SP: [s.n.], 1992. 39

[Freitas 2005]FREITAS, S. A. A. de. *Interpretação automatizada de textos: Processamento de Anáforas*. Tese (Doutorado) — Univerdade Federal do Espírito Santo, 2005. , 14, 16, 27, 28, 30, 33, 34, 39, 53, 54, 55, 58, 72, 100

[Freitas e Lopes 1993]FREITAS, S. A. A. de; LOPES, J. G. P. Um sistema de representação do discurso utilizando a drt e a teoria do foco. In: *X Simpósio Brasileiro de Inteligencia Artificial*. [S.l.: s.n.], 1993. 28, 39

[Freitas e Lopes 1994]FREITAS, S. A. A. de; LOPES, J. G. P. Discourse segmentation: Extending the centering theory. In: *XI Simpósio Brasileiro de Inteligência Artificial*. UFCE - Fortaleza - CE: [s.n.], 1994. 28, 37

[Freitas e Lopes 1995]FREITAS, S. A. A. de; LOPES, J. G. P. Improving centering to support a discourse segmentation. In: *Workshop on Focus and Natural Language Processing*. IBM Working Papers of the Institute for Logic and Linguistics: Focus and Natural Language Processing. v.3.: [s.n.], 1995. 28

[Ghahramani 2004]GHAHRAMANI, Z. *Unsupervised Learning*. London, UK, 2004. 13, 22

[Ghani 2000]GHANI, R. Using error-correcting codes for text classification. In: *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. p. 303–310. 23

[Ghani 2002]GHANI, R. Combining labeled and unlabeled data for multiclass text categorization. In: *In Proceedings of the International Conference on Machine Learning*. [S.l.: s.n.], 2002. p. 187–194. 13, 23

[Gliozzo, Strapparava e Dagan 2009]GLIOZZO, A.; STRAPPARAVA, C.; DAGAN, I. Improving text categorization bootstrapping via unsupervised learning. *ACM Trans. Speech Lang. Process.*, ACM, New York, NY, USA, v. 6, n. 1, p. 1–24, 2009. ISSN 1550-4875. 13, 26, 101

[Goldstein 1972]GOLDSTEIN, M. K-nearest neighbor classification. *IEEE Transactions On Information Theory*, IEEE Computer Society, v. 18, p. 627–630, 1972. 21

[Gomes 1990]GOMES, H. E. *Manual de elaboração de tesauros monolíngues*. Brasília, 1990. 47

[Grosz 1977]GROSZ, B. J. *The Representation and Use of Focus in a System for Understanding Dialogs*. SRI International, Menlo Park, California, 1977. 37

[Grosz, Joshi e Weinstein 1995]GROSZ, B. J.; JOSHI, A. K.; WEINSTEIN, S. Centering: A framework for modelling the local coherence of discourse. *cl*, v. 21, n. 2, p. 203–225, 1995. 37

[Hajičová, Skoumalová e Sgall 1995]HAJIČOVá, E.; SKOUMALOVá, H.; SGALL, P. An automatic procedure for topic-focus identification. *cl*, v. 21, n. 1, p. 81–94, 1995. 37

[Han, Karypis e Kumar 2001]HAN, E.-H.; KARYPIS, G.; KUMAR, V. Text categorization using weight adjusted k-nearest neighbor classification. In: *PAKDD '01: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. London, UK: Springer-Verlag, 2001. p. 53–65. 22

[Hobbs 1986]HOBBS, J. Resolving pronoun references. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 339–352, 1986. 30

[Iida, Inui e Matsumoto 2005]IIDA, R.; INUI, K.; MATSUMOTO, Y. Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processing (TALIP)*, ACM, New York, NY, USA, v. 4, n. 4, p. 417–434, 2005. ISSN 1530-0226. 30

[Jain, Murty e Flynn 1999]JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 31, n. 3, p. 264–323, 1999. 23

[Jeon e Landgrebe 1999]JEON, B.; LANDGREBE, D. A. Partially supervised classification using weighted unsupervised clustering. *IEEE Trans. on Geoscience and Remote Sensing*, v. 37, p. 1073–1079, 1999. 13, 24

[Júnior 2007]JúNIOR, H. S. *Recuperação de informações relevantes em documentos digitais baseada na resolução de anáforas*. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2007. 16, 28, 51, 58

[Joachims 2002]JOACHIMS, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 2002. 22

[Kamp e Reyle 1993]KAMP, H.; REYLE, U. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. [S.l.]: Kluwer Academic Publishers, 1993. 39

[Kaufman e Rousseeuw 1990]KAUFMAN, L.; ROUSSEEUW, P. *Finding Groups in Data An Introduction to Cluster Analysis*. New York: Wiley Interscience, 1990. 23

[Khan, Ding e Perrizo 2002]KHAN, M.; DING, Q.; PERRIZO, W. k-nearest neighbor classification on spatial data streams using p-trees. In: *PAKDD '02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. London, UK: Springer-Verlag, 2002. p. 517–518. 22

[Ko e Seo 2002]KO, Y.; SEO, J. Text categorization using feature projections. In: *Proceedings of the 19th international conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2002. p. 1–7. 25

[Ko e Seo 2004]KO, Y.; SEO, J. Learning with unlabeled data for text categorization using bootstrapping and feature projection techniques. In: *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004. p. 255. 25

[Ko e Seo 2009]KO, Y.; SEO, J. Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 45, n. 1, p. 70–83, 2009. 13, 16, 25, 48, 50, 77, 101

[Lappin e Leass 1994]LAPPIN, S.; LEASS, H. J. An algorithm for pronominal anaphora resolution. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 20, n. 4, p. 535–561, 1994. ISSN 0891-2017. 30

[Lewis et al. 1996]LEWIS, D. D. et al. Training algorithms for linear text classifiers. In: *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 1996. p. 298–306. 22

[Liu et al. 2002]LIU, B. et al. Partially supervised classification of text documents. In: *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. p. 387–394. 13, 24

[Liu et al. 2004]LIU, B. et al. Text classification by labeling words. In: *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*. [S.l.]: AAAI Press / The MIT Press, 2004. p. 425–430. 13

[Manning e Schtze 1999]MANNING, C. D.; SCHTZE, H. *Foundations of Statistical Natural Language Processing*. [S.l.]: The MIT Press, 1999. Hardcover. 25

[Maziero Thiago A. S. Pardo 2008]MAZIERO THIAGO A. S. PARDO, A. D. F. B. C. D.-d.-S. E. G. A base de dados lexical e a interface web do tep 2.0 - thesaurus eletrônico para o português do brasil. In: *TIL'2008: Proceedings of the 6th Workshop in Information and Human Language Technology*. Vila Velha-ES, Brasil: [s.n.], 2008. p. 390–392. 77, 89

[McCallum e Nigam 1998]MCCALLUM, A.; NIGAM, K. A comparison of event models for naive bayes text classification. In: *In AAAI-98 Workshop on Learning for Text Categorization*. [S.l.]: AAAI Press, 1998. p. 41–48. 21

[Mccallum e Nigam 1999]MCCALLUM, A.; NIGAM, K. Text classification by bootstrapping with keywords, em and shrinkage. In: *In ACL99 - Workshop for Unsupervised Learning in Natural Language Processing*. [S.l.: s.n.], 1999. p. 52–58. 13, 25, 30, 101

[McCallum]MCCALLUM, A. K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Http://www.cs.cmu.edu/ mccallum/bow. 89

[McCallum et al. 2000]MCCALLUM, A. K. et al. Automating the construction of internet portals with machine learning. *Inf. Retr.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 3, n. 2, p. 127–163, 2000. ISSN 1386-4564. 26

[Mitchell 1997]MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. 21

[Mitkov 2005]MITKOV, R. *The Oxford Handbook of Computational Linguistics*. [S.l.]: Oxford University Press, 2005. 21, 31

[Mitkov et al. 2007]MITKOV, R. et al. Anaphora resolution: to what extent does it help nlp applications? In: *DAARC'07: Proceedings of the 6th discourse anaphora and anaphor resolution conference on Anaphora*. Berlin, Heidelberg: Springer-Verlag, 2007. p. 179–190. 16, 26, 27, 55

[Nigam et al. 1998]NIGAM, K. et al. Learning to classify text from labeled and unlabeled documents. In: *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1998. p. 792–799. 13, 23

[Oliveira e Quental 2003]OLIVEIRA, C. M. G. M. de; QUENTAL, V. de S. T. D. B. Aplicações do processamento automático de linguagem natural na recuperação de informações. *Congresso Internacional da ABRALIN*, Anais do III ABRALIN, Rio de Janeiro, RJ, Brasil, p. 949–955, 2003. 59

[Orasan e Evans 2007]ORASAN, C.; EVANS, R. Np animacy identification for anaphora resolution. *J. Artif. Int. Res.*, AI Access Foundation, , USA, v. 29, n. 1, p. 79–103, 2007. ISSN 1076-9757. 37

[Orengo e Huyck 2001]ORENGO, V.; HUYCK, C. A stemming algorithmm for the portuguese language. *String Processing and Information Retrieval, International Symposium on*, IEEE Computer Society, Los Alamitos, CA, USA, v. 0, p. 0186, 2001. 79

[PALAVRAS]PALAVRAS. The constraint grammar category set of "palavras". Http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html. 79, 88

[Palomar et al. 2001]PALOMAR, M. et al. An algorithm for anaphora resolution in spanish texts. *Computational Linguistics*, v. 27, p. 567, 2001. 30

[Pereira 2009]PEREIRA, F. S. do C. *Uma Metodologia para a utilização do processamento de Linguagem Natural na busca de informações em documentos digitais.* Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2009. 16, 28, 51, 58, 72, 87, 100

[Pereira, Júnior e Freitas 2009]PEREIRA, F. S. do C.; JúNIOR, H. S.; FREITAS, S. A. A. de. An anaphora based information retrieval model extension. In: *CSIE*. Los Angeles, LA, USA: [s.n.], 2009. 28, 51, 58

[Pereira, Morellato e Freitas 2009]PEREIRA, F. S. do C.; MORELLATO, L. V.; FREITAS, S. A. A. de. Evaluation of an information retrieval model based in anaphora resolution. In: *IADIS International Conference WWW/Internet*. Rome, Italy: [s.n.], 2009. 28

[Polanyi 1988]POLANYI, L. A formal model of the structure of discourse. *Journal of Pragmatics*, n. 12, p. 601–638, 1988. 40

[Polanyi, Berg e Ahn 2003]POLANYI, L.; BERG, M. van den; AHN, D. Discourse structure and sentential information structure. *jolli*, v. 12, p. 337–350, 2003. 40

[Ragas e Koster 1998]RAGAS, H.; KOSTER, C. H. A. Four text classification algorithms compared on a dutch corpus. In: *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 1998. p. 369–370. 22

[Ren et al. 2009]REN, J. et al. Naive bayes classification of uncertain data. *Data Mining, IEEE International Conference on*, IEEE Computer Society, Los Alamitos, CA, USA, v. 0, p. 944–949, 2009. 21

[Riloff e Jones 1999]RILOFF, E.; JONES, R. Learning dictionaries for information extraction by multi-level bootstrapping. In: *AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1999. p. 474–479. 25

[Rochio 1971]ROCHIO, J. Relevance feedback in information retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, p. 313–323, 1971. 22

[Rogati e Yang 2002]ROGATI, M.; YANG, Y. High-performing feature selection for text classification. In: *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*. New York, NY, USA: ACM, 2002. p. 659–661. 22

[Rogova 1994]ROGOVA, G. Combining the results of several neural network classifiers. *Neural Netw.*, Elsevier Science Ltd., Oxford, UK, UK, v. 7, n. 5, p. 777–781, 1994. 22

[Sebastiani 2002]SEBASTIANI, F. Machine learning in automated text categorization. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 34, n. 1, p. 1–47, 2002. 13, 20, 53

[Sidner 1979]SIDNER, C. L. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Cambridge, MA, USA, 1979. 37

[Sidner 1981]SIDNER, C. L. Focusing for interpretation of pronouns. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 7, n. 4, p. 217–231, 1981. ISSN 0891-2017. 37

[Slonim, Friedman e Tishby 2002]SLONIM, N.; FRIEDMAN, N.; TISHBY, N. Unsupervised document classification using sequential information maximization. In: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2002. p. 129–136. 23

[Slonim e Tishby 2000]SLONIM, N.; TISHBY, N. Document clustering using word clusters via the information bottleneck method. In: *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2000. p. 208–215. 23

[Stark e Riesenfeld 1998]STARK, M. M.; RIESENFELD, R. F. Wordnet: An electronic lexical database. In: *Proceedings of 11th Eurographics Workshop on Rendering*. [S.l.]: MIT Press, 1998. 89

[Vapnik 1995]VAPNIK, V. N. *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8. 22

[Wang et al. 2007]WANG, Q. et al. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, American Society for Microbiology, v. 73, p. 5261–5267, 2007. 21

[Yang, Slattery e Ghani 2002]YANG, Y.; SLATTERY, S.; GHANI, R. A study of approaches to hypertext categorization. *J. Intell. Inf. Syst.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 18, n. 2-3, p. 219–241, 2002. 21

[Yarowsky 1995]YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1995. p. 189–196. 25

[Yeh e Chen 2003]YEH, C.-L.; CHEN, Y.-C. Using zero anaphora resolution to improve text categorization. In: *Proceedings of the 17th Pacific Asia Conference*. Taipei, Taiwan: Colips Publications, 2003. p. 423–430. 16, 27

[Zhang et al. 2009]ZHANG, C. et al. Web-scale classification with naive bayes. In: *WWW '09: Proceedings of the 18th international conference on World wide web*. New York, NY, USA: ACM, 2009. p. 1083–1084. 21