Sumarização multidocumento com base em aspectos informativos

Alessandro Yovan Bokan Garay

Alessandro Yovan Bokan Garay

Sumarização multidocumento com base em aspectos informativos

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

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
Julho de 2015

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

Bokan Garay, Alessandro Yovan
B686s Sumarização multidocumento com base em aspectos informativos / Alessandro Yovan Bokan Garay; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2015.
154 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

1. Sumarização Automática Multidocumento. 2. Aspectos Informativos. 3. Anotação de Papéis Semânticos. 4. Aprendizado de Maquina. 5. Processamento de Língua Natural. I. Salgueiro Pardo, Thiago Alexandre, orient. II. Título.

Alessandro Yovan Bokan Garay

Multi-document summarization based on informative aspects

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
July 2015

Agradecimentos

À Deus, pela força para realizar este trabalho.

À minha família, pelo seu amor e carinho incondicional, sua confiança em mim e seu apoio mesmo eu estando longe.

A meu orientador, Prof. Thiago Alexandre Salgueiro Pardo, pela confiança depositada em mim e por sua orientação e dedicação ao longo deste projeto.

Aos integrantes do NILC, pelos grandes momentos de convivência juntos e por terem compartilhado comigo a sua amizade e conhecimentos da área de Processamento de Língua Natural (PLN).

À SAMSUNG Eletrônica da Amazônia LTDA, junto com a Fundação de Apoio à Física e à Química (FAFQ), por fomentar o projeto de pesquisa no qual meu mestrado está inserido e pelo apoio financeiro prestado ao meu mestrado. Também à CAPES pelo apoio financeiro logo no início do meu mestrado.

Resumo

A sumarização multidocumento consiste na produção de um sumário/resumo a partir de uma coleção de textos sobre um mesmo assunto. Devido à grande quantidade de informação disponível na Web, esta tarefa é de grande relevância já que pode facilitar a leitura dos usuários. Os aspectos informativos representam as unidades básicas de informação presentes nos textos. Por exemplo, em textos jornalísticos em que se relata um fato/acontecimento, os aspectos podem representar a seguintes informações: o que aconteceu, onde aconteceu, quando aconteceu, como aconteceu, e por que aconteceu. Conhecendo-se esses aspectos e as estratégias de produção e organização de sumários, é possível automatizar a tarefa de sumarização. No entanto, para o Português do Brasil, não há pesquisa feita sobre sumarização com base em aspectos. Portanto, neste trabalho de mestrado, investigaram-se métodos de sumarização multidocumento com base em aspectos informativos, pertencente à abordagem profunda para a sumarização, em que se busca interpretar o texto para se produzir sumários mais informativos. Em particular, implementaram-se duas etapas relacionadas: (i) identificação automática de aspectos os aspectos informativos e (ii) desenvolvimento e avaliação de dois métodos de sumarização com base em padrões de aspectos (ou *templates*) em sumários. Na etapa (i), criaram-se classificadores de aspectos com base em anotador de papéis semânticos, reconhecedor de entidades mencionadas, regras manuais e técnicas de aprendizado de máquina. Avaliaram-se os classificadores sobre o córpus CSTNews (Rassi et al., 2013; Felippo et al., 2014). Os resultados foram satisfatórios, demostrando que alguns aspec-

tos podem ser identificados automaticamente em textos jornalísticos com um desempenho razoável. Já na etapa (ii), elaboraram-se dois métodos inéditos de sumarização multidocumento com base em aspectos. Os resultados obtidos mostram que os métodos propostos neste trabalho são competitivos com os métodos da literatura. Salienta-se que esta abordagem para sumarização tem recebido grande destaque ultimamente. Além disso, é inédita nos trabalhos desenvolvidos no Brasil, podendo trazer contribuições importantes para a área.

Palavras-chave: Sumarização Automática Multidocumento, Aspectos Informativos, Anotação de Papéis Semânticos, Reconhecimento de Entidades Nomeadas, Regras Manuais, Aprendizado de Maquina, Processamento de Língua Natural, Português Brasileiro.

Abstract

Multi-document summarization is the task of automatically producing a unique summary from a group of texts on the same topic. With the huge amount of available information in the web, this task is very relevant because it can facilitate the reading of the users. Informative aspects, in particular, represent the basic information units in texts and summaries, e.g., in news texts there should be the following information: what happened, when it happened, where it happened, how it happened and why it happened. Knowing these aspects and the strategies to produce and organize summaries, it is possible to automate the aspect-based summarization. However, there is no research about aspect-based multi-document summarization for Brazilian Portuguese. This research work investigates multi-document summarization methods based on informative aspects, which follows the deep approach for summarization, in which it aims at interpreting the texts to produce more informative summaries. In particular, two main stages are developed: (i) the automatic identification of informative aspects and (ii) and the development and evaluation of two summarization methods based on aspects patterns (or templates). In the step (i) classifiers were created based on semantic role labeling, named entity recognition, handcrafted rules and machine learning techniques. Classifiers were evaluated on the CSTNews annotated corpus (Rassi et al., 2013; Felippo et al., 2014). The results were satisfactory, demonstrating that some aspects can be automatically identified in the news with a reasonable performance. In the step (ii) two novels aspect-based multi-document summarization methods are elaborated. The

results show that the proposed methods in this work are competitive with the classical methods. It should be noted that this approach has lately received a lot of attention. Furthermore, it is unprecedented in the summarization task developed in Brazil, with the potential to bring important contributions to the area.

Keywords: Multi-document Summarization, Informative Aspects, Semantic Role Labeling, Named Entity Recognition, Handcrafted Rules, Machine Learning, Natural Language Processing, Brazilian Portuguese.

Publicações

Como resultado da pesquisa feita neste trabalho de mestrado, como primeiro autor, até o momento, foram publicados 1 artigo em evento internacional e dois relatórios técnicos:

- Bokan, A. and T., Pardo. (2015). Automatic Microaspect Identification. In *Proceedings of the XVI International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2015)*, Cairo, Egypt, pp. 1-12. To appear in a special issue of the journal Research in Computing Science (RCS)[1].

- Bokan, A. and T., Pardo. (2015). Identificação Automática de **Microaspectos** em Textos Jornalísticos. Technical Report NILC-TR-15-01, Série de Relatórios do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP)[2].

- Bokan, A. and T., Pardo. (2015). Identificação Automática de **Macroaspectos** em Textos Jornalísticos. Technical Report NILC-TR-15-02, Série de Relatórios do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP)[3].

[1] http://rcs.cic.ipn.mx

[2] http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/ESTAGIO-BIBLIO_171_RT%20406.pdf

[3] http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_158_RT_407.pdf

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AM	Aprendizado de Máquina
APR	Anotação de Papéis Retóricos
APS	Anotação de Papéis Semânticos
CST	*Cross-document Structure Theory*
DRE	Detecção de Relações entre Entidades
EM	Entidade Mencionada
PLN	Processamento de Língua Natural
POS	*Part-of-speech*
REM	Reconhecimento de Entidades Mencionadas
RST	*Rhetoric Structure Theory*
SA	Sumarização Automática
TAC	*Text Analysis Conference*
TF-IDF	*Term Frequency–Inverse Document Frequency*

CAPÍTULO

1

Introdução

1.1 Contextualização e Motivação

Nas últimas décadas, muitas tecnologias têm surgido, trazendo com isso um crescente aumento no volume de informação. Com o amplo uso da Web hoje em dia, meios de comunicação (como agências de notícias online, blogs, microblogs e redes sociais) têm tornado accessível uma enorme quantidade de informação textual e, em consequência, seu processamento tem se tornado cada vez mais difícil, tanto para humanos quanto para máquinas. Enquanto humanos não têm capacidade e tempo de ler e apreender todas as informações disponíveis de seu interesse, máquinas perdem em precisão e desempenho ao lidar com tamanha quantidade e variedade de documentos.

O informe publicado em 2012 pela *International Data Corporation* (IDC) (Gantz e Reinsel, 2012), mostrou que, nesse ano, a Web foi responsável pela disponibilização de 2.8 zettabytes de informação, uma quantidade catorze vezes maior do que a produzida cinco anos atrás, e também se afirmou que a quantidade ainda crescerá para 8.0 zettabytes em 2015. Nesse cenário, a Sumarização Automática (SA) mostra-se como uma tarefa que pode auxiliar significativamente no fornecimento de informações para simplificar a leitura das pessoas.

A tarefa de SA consiste na produção automática de uma versão mais curta de um texto-fonte, chamada de sumário ou resumo (Mani, 2001). O objetivo da SA é gerar sumários similares aos sumários humanos. O sumário produzido a partir de um único texto-fonte é denominado sumário monodocumento. Com a grande quantidade de informação armazenada na Web, surgiu a área de SA multidocumento (McKeown e Radev, 1995; Radev e McKeown, 1998). A SA multidocumento consiste na produção de um único sumário a partir de um conjunto de textos-fonte/documentos sobre um mesmo assunto.

Além de selecionar as informações mais relevantes, a sumarização multidocumento deve lidar com os fenômenos multidocumento, como a presença de informações redundantes, complementares e contraditórias, estilos de escrita variados, ordenação temporal dos eventos/fatos, perspectivas e focos diferentes, assim como a própria questão da coerência e coesão do sumário. Tais fenômenos ocorrem porque os documentos a serem sumarizados têm origem diversificada e são escritos em diferentes momentos. A Fig. 1 ilustra um sumário multidocumento produzido a partir de 3 notícias jornalísticas (publicadas pelas agências de notícias online Estadão, O Globo e Jornal do Brasil) sobre a "vitória da seleção brasileira de Vôlei". Tal sumário é um texto coeso (mantém uma relação sequencial entre sentenças) e coerente (texto compreensível), formado pelas sentenças mais importantes extraídas dos textos-fonte.

A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06. Amanhã as equipes voltarão a se enfrentar, no mesmo local. Com o resultado, o Brasil está na liderança do grupo B, perto da classificação para a próxima fase do campeonato. A seleção brasileira ainda enfrentará portugueses e finlandeses na fase de classificação. A equipe brasileira já conquistou cinco vezes a Liga Mundial. A fase final deste ano acontecerá na Rússia.

Figura 1: Sumário multidocumento sobre vitória da seleção brasileira de Vôlei

As primeiras pesquisas em sumarização multidocumento datam dos anos 90 (McKeown e Radev, 1995; Carbonell e Goldstein, 1998) e as investigações são mais intensas nestes dias. Tais pesquisas são motivadas pela relevância da aplicação da sumarização

em importantes sistemas de recuperação e extração de informação, como buscadores de notícias (p.ex., Google News[1] e Wiki News[2]), sintetizadores de informação (p.ex., WolframAlpha[3] e Qwiki[4]) e repositórios digitais (por exemplo, CiteSeer[5] e DBLP[6]). Sua aplicabilidade no cenário atual pode ser facilmente ilustrada, basta imaginar um usuário que deseje conhecer algum tópico em particular. Diante da grande quantidade de notícias com que se defrontaria e a impossibilidade de lidar com ela, um sumário poderia ser de enorme valia.

Segundo Mani (2001), existem duas abordagens tradicionais para a SA em geral: a superficial (ou empírico/estatística) e a profunda (ou fundamental). A primeira faz uso de pouco conhecimento linguístico (nível léxico e morfossintático), produzindo sumários formados por extratos[7] do texto por meio de frequência de palavras (ou termos), sendo suas vantagens a escalabilidade e a robustez. A segunda faz uso de mais conhecimento linguístico, atingindo o nível semântico e discursivo, produzindo melhores resultados, entretanto, mais caros.

A diferença entre um sumário superficial e um sumário profundo[8] é ilustrada nas Figs. 2 e 3. Pode-se observar a presença de informação redundante entre as sentenças em **negrito** da Fig. 2, problema comum em uma abordagem superficial. Por outro lado, a Fig. 3 mostra um sumário sem presença de sentenças redundantes.

[1] https://news.google.com.br/
[2] https://pt.wikinews.org
[3] http://www.wolframalpha.com/
[4] http://www.qwiki.com/
[5] http://citeseerx.ist.psu.edu/
[6] http://dblp.uni-trier.de
[7] Referem-se a palavras, frases, sentenças ou parágrafos extraídos do texto-fonte.
[8] Tais sumários foram extraídos da seção de demonstrações do projeto Sucinto (http://www.icmc.usp.br/pessoas/taspardo/sucinto/)

A seleção brasileira masculina de vôlei conseguiu, nesta sexta-feira, **a sétima vitória consecutiva na Liga Mundial ao derrotar a Finlândia por 3 sets a 0 - parciais de 25/17, 25/22 e 25/21 -, em jogo realizado na cidade de Tampere, na Finlândia. A seleção brasileira masculina de vôlei, que é treinada por Bernardinho, venceu a Finlândia por 3 sets a 0, parciais de 25/17, 25/22 e 25/21, nesta sexta-feira, em Tampere (FIN)**, e manteve sua invencibilidade na Liga Mundial-06. O resultado de hoje deixou o Brasil perto de conquistar a única vaga do Grupo B da Liga Mundial. **O Brasil arrasou a Finlândia no primeiro confronto entre as seleções, nesta sexta-feira, na cidade de Tampere, pela Liga Mundial de vôlei 2006, por 3 sets a 0, com parciais de 25/17, 25/22 e 25/21.**

Figura 2: Sumário automático com redundância produzido pela ferramenta GistSumm (Pardo, 2005)

A seleção brasileira masculina de vôlei conseguiu, nesta sexta-feira, a sétima vitória consecutiva na Liga Mundial ao derrotar a Finlândia por 3 sets a 0 - parciais de 25/17, 25/22 e 25/21 -, em jogo realizado na cidade de Tampere, na Finlândia. O resultado de hoje deixou o Brasil perto de conquistar a única vaga do Grupo B da Liga Mundial, que classifica o melhor de cada uma das quatro chaves, a Rússia (país-sede) e mais um time convidado pela Federação Internacional de Vôlei, para a fase final, de 23 a 27 de agosto, em Moscou (Rússia).

Figura 3: Sumário automático sem redundância produzido pela ferramenta CSTSumm (Castro Jorge, 2010)

Independentemente da abordagem seguida, Mani (1999) sugere uma arquitetura genérica para a SA, a qual é mostrada na Fig. 4. A arquitetura tem três etapas. Na etapa de *análise*, os textos-fonte são processados e seu conteúdo é representado em um ou mais níveis de análise linguística: morfológico, sintático, semântico e/ou discursivo. Na etapa de *transformação*, o conteúdo é simplificado, selecionando-se as informações relevantes por meio de um método de seleção de conteúdo e combinação de informações. Finalmente, na etapa de *síntese*, são usados métodos de geração de texto para organizar

e apresentar o conteúdo selecionado em língua natural. Neste trabalho de Mestrado, seguem-se as três etapas da arquitetura apresentada.

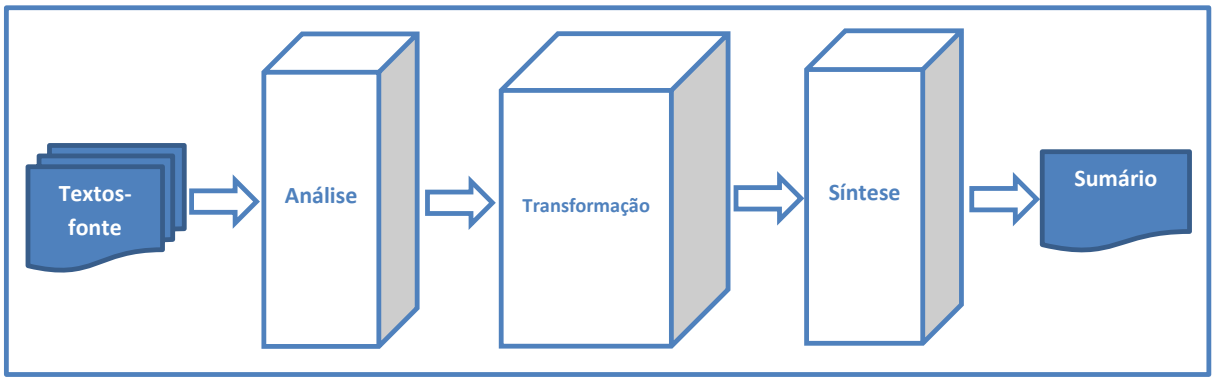

Figura 4: Arquitetura de um sistema SA

No Brasil, as pesquisas em sumarização multidocumento para a língua portuguesa são mais recentes, iniciando-se oficialmente em 2005 com uma extensão do sistema simples chamado GistSumm (Pardo, 2005). Até o momento, foram produzidos vários recursos, ferramentas e sistemas, como o córpus de referência CSTNews (Aleixo e Pardo, 2008; Cardoso et al., 2011), o parser discursivo CSTParser (Maziero e Pardo, 2011) e sistemas de sumarização do estado da arte das linhas superficial (Ribaldo et al., 2012) e profunda (Castro Jorge e Pardo, 2010; Castro Jorge, 2010; Castro Jorge e Pardo, 2011; Cardoso et al., 2011; Cardoso, 2014; Castro Jorge, 2015).

Na atualidade, no âmbito da *Text Analysis Conference* (TAC)[9] , a principal conferência e competição cietífica dedicada à SA, Owczarzak e Dang (2011) propuseram a utilização de "aspectos informativos" como uma abordagem profunda para a produção de sumários multidocumento. Segundo os autores, como há grande variabilidade na produção de sumários e muitos desafios na sumarização multidocumento, os aspectos podem ser úteis para a produção de sumários coerentes e mais direcionados para o gênero (p.ex., jornalístico, científico, opinião, literário, etc.) e categoria textual em foco. As categorias, segundo a definição dos autores, indicam o assunto ou domínio do texto (p.ex., política, economia, esporte, etc.).

Os aspectos informativos representam componentes semântico-discursivos presentes em textos, que correspondem às unidades básicas de informação. Com relação à

[9] http://www.nist.gov/tac/

semântica, os aspectos podem representar o sentido de uma sentença. Quanto ao nível discursivo, os aspectos podem representar o sentido ou organização do texto como um todo, baseado na relação entre as sentenças anotadas com aspectos. Por exemplo, sabe-se que notícias devem conter pelo menos os aspectos "o que aconteceu", "quando aconteceu" e "onde aconteceu". A identificação de aspectos informativos pode ser útil tanto para a determinação de informações relevantes dos textos-fonte, quanto para a identificação de restrições estruturais durante a construção dos sumários (Genest et al., 2009).

Na TAC, os aspectos são propostos em função da categoria textual (ou domínio) para a qual se quer produzir sumários. Como ilustração, propõe-se na TAC que sumários da categoria "Acidentes e desastres naturais" podem conter os aspectos WHAT, WHEN, WHERE, WHY, DAMAGES, WHO_AFFECTED e COUNTERMEASURES[10]; já os sumários da categoria "Ataques", por sua vez, podem conter os aspectos WHAT, WHEN, WHERE, DAMAGES, WHO_AFFECTED, PERPETRATORS e COUNTERMEASURES. Na Fig. 5, ilustra-se um sumário multidocumento da categoria "Ataque" anotado manualmente com aspectos. A primeira sentença do sumário informa sobre uma série de ataques criminosos ocorridos (WHAT) na cidade de São Paulo (WHERE) no dia 7 (WHEN). A segunda sentença descreve as entidades afetadas pelos ataques ocorridos em São Paulo (WHO_AFFECTED). Por fim, a última sentença narra quem são os criminosos (PERPETRATORS). Os aspectos podem ser específicos para cada categoria ou genéricos. Os aspectos textuais serão detalhados na Seção 2.2.

[Uma nova série de ataques criminosos foi registrada na madrugada desta segunda -feira, dia 7, em São Paulo e municípios do interior paulista.] **WHAT/WHEN/WHERE**
[Os bandidos atacaram agências bancárias, bases policiais e prédios públicos com bombas e tiros.] **WHO_AFFECTED**
[As ações são atribuídas à facção criminosa Primeiro Comando da Capital (PCC), que já comandou outros ataques em duas ocasiões.] **PERPETRATOR**

Figura 5: Sumário multidocumento sobre ataques criminosos em São Paulo

[10]A terminologia foi mantida em inglês, como a proposta original da TAC.

A partir de sua adoção na TAC, os aspectos foram utilizados em vários trabalhos da literatura para auxiliar a tarefa de sumarização (Steinberger et al., 2010; Li et al., 2011; Genest e Lapalme, 2012). Porém, o uso de aspectos não é novidade em sumarização e nem em outras áreas da Linguística e do Processamento de Linguagem Natural (PLN). Por exemplo, Swales (1990) propõe o uso de aspectos como componentes semântico-discursivos aplicados no modelo CARS (*Create a Research Space*) na forma de estruturas esquemáticas para construir/estruturar textos científicos. Tal modelo serviu de base para vários outros e já foi empregado para o português (Aluísio e Oliveira Jr., 1996; Feltrim, 2004; Souza e Feltrim, 2013). Alguns trabalhos pioneiros em sumarização que usaram o conceito de aspectos informativos são os trabalhos de Teufel e Moens (1999, 2002) e White et al. (2001), os quais se focaram em textos científicos e notícias sobre desastres naturais. Acredita-se também que os aspectos possam auxiliar outras tarefas relacionadas, como Mineração de Texto, por exemplo.

Diante desse cenário, neste projeto de pesquisa, explora-se o uso de aspectos na sumarização multidocumento de textos escritos em português do Brasil, usando os aspectos informativos como fonte de conhecimento para gerar sumários. Além de ser uma estratégia de sumarização que ganhou destaque ultimamente, o trabalho com aspectos pertence à abordagem profunda e dá continuidade natural aos trabalhos desenvolvidos até o momento, sendo sua aplicação inédita para a sumarização no Brasil. Cabe ressaltar, que neste trabalho, os aspectos informativos estão definidos especificamente para o **gênero jornalístico**, com base na tarefa de sumarização promovida pela TAC. A seguir, especifica-se a lacuna tratada e apresentam-se os objetivos e hipóteses deste trabalho.

1.2 Lacuna, Objetivos e Hipóteses

A SA multidocumento para o português do Brasil tem sido relativamente pouco investigada diante do que ocorre para o inglês. Existem algumas pesquisas na abordagem superficial (Pardo, 2005; Ribaldo et al., 2012), mas com resultados insatisfatórios desde a perspectiva humana por causa da falta de conhecimento linguístico empregado. Por outro lado, na abordagem profunda, existem os trabalhos de Castro Jorge e Pardo (2010); Castro Jorge (2010); Cardoso et al. (2011); Maziero e Pardo (2011); Cardoso

(2014); Castro Jorge (2015), fazendo uso de teorias discursivas como *Cross-document Structure Theory* (CST) (Mann e Thompson, 1987) e *Rhetoric Structure Theory* (RST) (Radev, 2000), mas, até então, nenhuma fez uma analise semântico-discursiva com base em aspectos informativos. Há, portanto, lacunas importantes a serem exploradas na área.

O objetivo principal deste trabalho consiste em investigar métodos automáticos de sumarização multidocumento com base em aspectos que possam gerar sumários mais informativos. Os objetivos específicos são:

- Identificar automaticamente aspectos informativos utilizando papéis semânticos, entidades nomeadas, regras manuais e técnicas de aprendizado de máquina e, assim, criar um classificador multirrótulo de aspectos;
- Desenvolver e avaliar alguns métodos de seleção de conteúdo para sumarização com base em padrões de ocorrência de aspectos em textos e sumários;
- Avaliar os sumários gerados.

As seguintes hipóteses permearam este trabalho:

- É possível identificar automaticamente aspectos informativos, assim como afirmar que existe um conjunto sistemático/recorrente de aspectos para cada domínio ou categoria textual específica.
- Existe uma ou mais estruturas típicas de aspectos (ou organizações aspectuais) em sumários e existem métodos eficazes para selecionar os aspectos que produzirão o sumário.

Portanto, identificando-se os aspectos e conhecendo-se as estratégias de produção e organização de sumários, é possível automatizar a tarefa de sumarização.

1.3 Metodologia de Trabalho

Para atingir os objetivos e verificar as hipóteses deste trabalho, em primeiro lugar, estudaram-se os conceitos relevantes da tarefa de SA multidocumento e revisaram-se os

principais trabalhos de sumarização com base em aspectos informativos. Em seguida, revisaram-se alguns recuros e ferramentas para o Português, que foram utilizadas ao longo deste trabalho. Um dos principais recursos é o corpus CSTNews (Aleixo e Pardo, 2008; Cardoso et al., 2011), formado por 50 coleções de textos jornalísticos e seus sumários, as quais contém textos que versam sobre um mesmo assunto. O CSTNews contém diversas anotações linguísticas de nível semântico-discursivo (ver Seção 3.1), sendo que, atualmente, seus sumários foram anotados com aspectos informativos em nível sentencial (Rassi et al., 2013; Felippo et al., 2014). Esse córpus subsidia o processo de identificação de aspectos, o levantamento da estrutura típica de sumários (em função dos aspectos), e a avaliação dos sumários produzidos automaticamente.

Com a finalidade de criar um classificador que possa identificar automaticamente aspectos informativos, utilizaram-se algumas ferramentas como anotador de papéis semânticos e reconhecedor de entidades mencionadas. Também se utilizaram técnicas de Aprendizado de Máquina e regras manuais com base em padrões linguísticos identificados nos textos. Avaliaram-se os classificadores sobre as sentenças anotadas com aspectos do córpus CSTNews.

Com base nas estruturas de típicas de aspectos identificadas na anotação do córpus CSTNews (Rassi et al., 2013), elaboraram-se dois **inovadores** métodos para selecionar e organizar o conteúdo que formará o sumário final. Para avaliar o desempenho dos métodos implementados, mediu-se a informatividade dos sumários gerados sobre os sumários manuais do córpus CSTNews. Também se comparou o desempenho dos métodos propostos contra dois dos melhores métodos de sumarização multidocumento para o Português.

Neste trabalho, realizou-se uma pesquisa exaustiva da SA multidocumento com base em aspectos informativos, sendo um trabalho inédito para a sumarização no Brasil. Já que a sumarização com base em aspectos à abordagem profunda, contribuiu-se com a exploração de vários recursos e a construção de ferramentas para beneficio da tarefa de sumarização multidocumento (ver Seção 6.1.

Este trabalho foi desenvolvido dentro do Núcleo Interinstitucional de Linguística Computacional (NILC)[11], que é um dos maiores grupo de pesquisa científica em PLN no

[11] http://nilc.icmc.usp.br/

Brasil e é pioneiro nas pesquisa de SA para o Português. Cabe ressaltar que este trabalho se beneficia de um projeto intitulado "Processamento Semântico de Textos em Português Brasileiro"[12], do convênio entre o Instituto de Ciências Matemáticas e de Computação (ICMC-USP) e a empresa SAMSUNG Eletrônica da Amazônia Ltda., cujo objetivo geral é avançar o estado da arte em processamento semântico de textos/documentos escritos em português brasileiro, via realização de pesquisa de base para o desenvolvimento futuro de ferramentas e produtos de PLN.

1.4 Organização do Trabalho

O restante do trabalho está organizado da seguinte forma: no Capítulo 2 , apresentam-se os principais conceitos da SA e os aspectos informativos, seguido de uma revisão literária sobre os métodos de sumarização com base em aspectos. No Capítulo 3, descrevem-se os recursos e ferramentas desenvolvidos para a língua portuguesa que foram utilizados neste trabalho de pesquisa. No Capítulo 4, explica-se detalhadamente o processo de identificação de aspectos e os métodos de sumarização. No Capítulo 5, apresentam-se os experimentos realizados ao longo deste trabalho e os resultados obtidos com as suas respectivas discussões. Por último, no Capítulo 6, apresentam-se as considerações finais do presente trabalho de mestrado.

[12] http://nilc.icmc.usp.br/semanticnlp/

CAPÍTULO

2

Identificação de Aspectos e Sumarização

Este capítulo se inicia com uma revisão dos conceitos envolvidos na tarefa de Sumarização Automática (SA) para, em seguida, definir os aspectos informativos que irão auxiliar o processo de sumarização. Os aspectos visam auxiliar na construção de sumários curtos e informativos segundo categorias pré-definidas (Owczarzak e Dang, 2011), promovendo uma análise linguística profunda (sintática, semântica e discursiva) dos textos-fonte. Além disso, serão apresentados alguns trabalhos da literatura que utilizaram aspectos para auxiliar a tarefa de sumarização.

Neste capítulo, apresentam-se conceitos gerais relacionados à SA (Seção 2.1). A seguir, descrevem-se detalhadamente os aspectos informativos usados como guia para sumarizar (Seção 2.2). Também são descritos alguns métodos de sumarização com base em aspectos informativos (Seçao 2.4). Por último, apresentam-se as considerações finais deste capítulo (Seção 2.5).

2.1 Sumarização Automática

O objetivo da Sumarização Automática (SA) é recuperar o conteúdo mais importante de um texto-fonte e apresenta-lo para o usuário final. A diferença entre sumarização

e outras tarefas de PLN é a condensação de conteúdo da informação de um ou mais documentos para o benefício do leitor (Mani, 2001). É sabido que os computadores podem examinar minuciosamente grandes quantidades de dados, mas os humanos são melhores fazendo inferências baseadas no contexto e no conhecimento do mundo. A tarefa de SA tenta modelar esse conhecimento, tentando gerar sumários automáticos similares aos sumários produzidos por humanos.

Sumários podem ser de 3 tipos: indicativos, informativos e críticos (Mani, 2001; Nenkova e McKeown, 2011). Sumários indicativos contêm apenas tópicos essenciais dos textos-fonte, não necessariamente contendo detalhes de resultados, argumentos e conclusões (p.ex., índices). Sumários informativos são considerados substitutos dos textos, devendo conter todos os aspectos principais (p.ex., *abstracts* de artigos científicos). Sumários críticos, além de sumarizar o conteúdo dos textos, adicionam crítica em relação ao conteúdo (p.ex., sinopses dos filmes).

Em termos de formação, sumários podem ser classificados como extrativos (extratos) ou abstrativos (em inglês, *abstracts*). Extratos são sumários compostos por segmentos inalterados dos textos-fonte, sendo construídos por operações de cópia/extração e cola/justaposição de segmentos integrais dos textos (normalmente, sentenças). *abstracts* apresentam partes reescritas, ou paráfrases[1], do texto-fonte. Métodos superficiais normalmente produzem extratos, enquanto métodos profundos têm capacidade para produzir *abstracts*, embora, na maioria dos casos, também produzam extratos, dadas as dificuldades de se produzir *abstracts*.

Existe uma grande variedade de requerimentos por parte do usuário para a tarefa de sumarização. Assim, podem existir sumários de diferentes tamanhos ou de distintas formas e organizações, etc. Também há sumários que são formados de acordo com uma determinada consulta. Por exemplo, é muito comum produzir sumários focados em um determinado tópico ou palavra-chave. Por outro lado, a quantidade de informação que se deseja ter em um sumário deve ser limitada para que este possa cumprir seu propósito. Logo, o usuário deve poder delimitar certo valor, chamado *taxa de compressão*, que determina o tamanho do sumário final em relação aos textos de origem, normalmente, em número de palavras. Usualmente, nos trabalhos para o Português,

[1]Reafirmação das ideias de um texto usando outras palavras.

um sumário deve ter 70% de compressão com respeito ao maior texto-fonte (ou texto com maior quantidade de palavras).

A tarefa que produz um único sumário automaticamente a partir de um único texto-fonte é denominada SA monodocumento. Esta tarefa tem sido bastante explorada e discutida por vários autores (Luhn, 1958; Edmundson, 1969; O'Donnell, 1997; Salton et al., 1997; Marcu, 2000; Conroy e O'leary, 2001; Pardo e Rino, 2002; Rino et al., 2004; Svore, 2007; Uzêda et al., 2010; Clarke e Lapata, 2010; Louis et al., 2010; Contractor et al., 2012).

Por outro lado, a SA multidocumento visa produzir um único sumário automaticamente a partir de uma coleção de textos-fonte sobre um mesmo assunto. A SA multidocumento representa uma área mais nova que tem adquirido relevância nos últimos anos. Como citado na Seção 1.1, as primeiras pesquisas datam dos anos 90 (McKeown e Radev, 1995; Radev e McKeown, 1998; Carbonell e Goldstein, 1998) e as investigações se intensificam nos dias atuais (Radev, 2000; Zhang et al., 2002; Otterbacher et al., 2002; McKeown et al., 2005; Wan e Yang, 2006; Nenkova, 2005b,a; Afantenos et al., 2008; Wan, 2008; Haghighi e Vanderwende, 2009; Castro Jorge e Pardo, 2010, 2011; Celikyilmaz e Hakkani-Tür, 2011; Ribaldo et al., 2012; Cardoso, 2014; Castro Jorge, 2015). Além disso, a área tem se especializado para lidar com determinados tipos de tarefas, como a sumarização de blogs e microblogs, sumarização de reuniões e eventos (em inglês, *meeting summarization*), sumarização de atualização (em inglês, *update summarization*) e sumarização de sentimentos/opiniões (em inglês, *sentiment/opinion summarization*), dentre outras.

Neste trabalho de pesquisa, visa-se gerar sumários extrativos de tipo informativo sobre um conjunto de documentos (multidocumento) do gênero jornalístico sobre um mesmo assunto. Escolheu-se esse tipo de textos, já que são escritos em uma linguagem formal, objetiva e imparcial, para compreensão do leitor. A seguir, descreve-se o contexto em que os aspectos foram utilizados como guia ou auxílio na produção de sumários mais informativos.

2.2 Aspectos Informativos

Como já foi dito no início (ver Seção 1.1), os aspectos informativos representam componentes semântico-discursivos que correspondem às unidades básicas de informação presentes nas sentenças dos textos do gênero jornalístico. Os aspectos podem representar componentes locais da sentença, indicando informações tais como um local específico ou uma data determinada; também podem representar o sentido ou organização do texto como um todo, baseado na relação entre as sentenças anotadas com aspectos.

Os aspectos foram propostos no âmbito da *Text Analysis Conference* (TAC) por Owczarzak e Dang (2011) como uma abordagem profunda para a produção de sumários. A TAC é a principal conferência e competição científica dedicada à SA, fornecendo uma grande coleção de dados de teste, procedimentos de avaliação e um fórum para compartilhar resultados. No ano 2010[2], a TAC propôs a tarefa de Sumarização Guiada[3] (em inglês, *Guided Summarization*), com a finalidade de produzir um sumário de 100 palavras a partir de um conjunto de 10 notícias jornalísticas *on-line*[4] para um tópico dado. Todos os participantes disponibilizavam uma lista de aspectos para cada categoria textual, e cada sumário produzido deveria conter todos os aspectos designados para cada categoria, sendo que as categorias indicam o assunto ou domínio do texto. As categorias e seus aspectos definidos pela TAC incluem:

- **Acidentes e desastres naturais:** descrição do fato (WHAT), data (WHEN), localização (WHERE), razões do acidente/desastre (WHY), danos (DAMAGES), entidade afetada (WHO_AFFECTED), esforços de resgate/contramedidas (COUNTERMEASURES).

- **Ataques:** descrição do fato (WHAT), data (WHEN), localização (WHERE), entidade afetada (WHO_AFFECTED), danos (DAMAGES), criminosos (PERPETRA-

[2] http://www.nist.gov/tac/2010/Summarization/Guided-Summ.2010.guidelines.html

[3] Modalidade da SA multidocumento "'assistida'" por aspectos informativos que visam construir sumários orientados pelo significado.

[4] Notícias recentes (*up-to-the-minute*), geralmente eletronicamente, para a mídia e muitas vezes para o público.

TORS), esforços de resgate/contramedidas (COUNTERMEASURES).

- **Saúde e segurança:** qual é o problema (WHAT), como foi afetado (HOW), quem foi afetado (WHO_AFFECTED), contramedidas (COUNTERMEASURES), por que isso acontece (WHY).

- **Recursos naturais ameaçados:** descrição do recurso (WHAT), importância do recurso (IMPORTANCE), ameaças (THREATS), contramedidas (COUNTERMEASURES).

- **Julgamentos e investigações:** o quem está sob investigação (WHO), quem está investigando ou processando (WHO_INV), o por quê (WHY), acusações específicas (CHARGES), sentença/consequência (SENTENCE), como é que reagiram às acusações (PLEAD).

Os aspectos podem ser específicos para cada categoria ou genéricos, indicando sua validade ou abrangência para um leque maior de domínios. Por exemplo, os aspectos WHY e DAMAGES da categoria "Acidentes e desastres naturais" são diferentes dos aspectos IMPORTANCE, THREATS e COUNTERMEASURES da categoria "Recursos naturais ameaçados". Por outro lado, o aspecto WHAT é geral e se aplica a quase todas as categorias (da mesma forma ocorre com WHEN e WHERE para as categorias "Acidentes e desastres naturais" e "Ataques"). Por esse motivo, os aspectos devem ser devidamente analisados e definidos para cada categoria de um dado córpus.

Muitos estudos foram desenvolvidos seguindo os princípios da TAC 2010. Por exemplo, Steinberger et al. (2010) realizaram análises semânticas profundas para a modelagem de aspectos visando a SA multilíngue. Makino et al. (2012) e Li et al. (2011), compilaram aspectos de sumários da Wikipédia. Barrera et al. (2011) criaram um sistema de perguntas e respostas com base na identificação de aspectos para diferentes categorias. Mesmo antes da TAC, alguns trabalhos já apresentavam abordagens semelhantes, por exemplo, White et al. (2001) propuseram *templates* com base em aspectos para sumários de textos de desastres naturais, e Zhou et al. (2005) estudaram os aspectos presentes em sumários biográficos.

Recentemente, Bing et al. (2015) propuseram um método de SA multidocumento que gera sumários do tipo abstrativo com base nos sintagmas nominais e verbais da sen-

tença. Os autores acreditam na ideia que os leitores armazenam conceitos chaves e fatos que ocorrem nos documentos alocados nos sintagmas da sentença. Por exemplo, no sintagma nominal "Um homem armado" e no sintagma verbal "entrou na escola pública" se expressam dois fatos importantes. A ideia é tentar reorganizar tais fatos/conceitos para formar novas as sentenças que irão ao sumário final. Os sumários foram testados sobre o córpus de teste da TAC 2011. Para avaliar os sumários gerados, utilizou-se a medida ROUGE [5] (Lin, 2004). Os resultados da ROUGE em relação à medida F1 foram baixos: ROUGE-2=0.1170 e ROUGE-SU4=0.1480.

Na seguinte seção, relatam-se alguns trabalhos que identificam aspectos informativos. Em seguida, descrevem-se alguns dos principais trabalhos de sumarização com base em aspectos, apresentados na competição da TAC 2011.

2.3 Identificação Automática de Aspectos

Os papéis retóricos indicam funções argumentativas e informativas dos segmentos textuais. Eles podem ser sinalizados por padrões linguísticos presentes na sentença. Assim, os alguns aspectos são similares aos papéis retóricos por emergirem da informação contida nas sentenças em seus contextos.

Dayrell et al. (2012) propuseram um sistema que detecta padrões linguísticos particulares em *abstracts* de artigos científicos escritos em língua inglesa, denominado MAZEA (***M**ulti-label **A**rgumentative **Z**oning for **E**nglish **A**bstracts*). O sistema tenta identificar papéis retóricos, também chamados de zonas argumentativas, nas sentenças dos textos de gênero científico: BACKGROUNG (contexto), GAP (lacuna), PURPOSE (objetivo), METHOD (metodologia), RESULT (resultados) e CONCLUSION (conclusão).

Devido ao fato de uma sentença poder ser anotada com mais de uma zona argumentativa, o problema de classificação torna-se multirrótulo. Os algoritmos de classificação resultaram da combinação dos algoritmos das bibliotecas Mulan[6], tais como *Classifier*

[5] Medida de avaliação da informatividade dos sumários automáticos. Basicamente, computa-se a co-ocorrência de n-gramas entre o sumário automático e um ou mais sumários de referência humanos (ver Seção 5.4).

[6] http://sourceforge.net/projects/mulan/

Chain (Read et al., 2011) e Rakel (Tsoumakas e Katakis, 2007), e da biblioteca WEKA[7], como *Sequential Minimal Optimization* (SMO) (Platt, 1998), otimização do método *Support Vector Machine* (SVM) (Vapnik, 1995); e J48, implementação *open source* do algoritmo C4.5 (Quinlan, 1993). Os melhores resultados foram obtidos pela combinação dos classificadores *Chain* + SMO com 69.00% de acurácia, a comparação do Rakel + J48 que obteve 65.00% de acurácia.

O sistema MAZEA está baseado no sistema AZEA (***A**rgumentative **Z**oning for **E**nglish **A**bstracts*) (Genoves Jr. et al., 2007), em que foi criado um classificador binário para identificar papéis retóricos nas sentenças. Utilizaram-se os algoritmos J48, SMO e Naïve Bayes (Russell e Norvig, 2003). Os melhores resultados foram obtidos pelo Naïve Bayes com 93.10% de acurácia, seguido do SMO com 92.9% de acurácia. Tanto o sistema MAZEA quanto o sistema AZEA baseiam-se na tentativa de identificar movimentos retóricos em textos científicos proposta por Teufel e Moens (2002) e Feltrim et al. (2006), adotando um enfoque linguístico profundo. Outros sistemas são independentes da categoria textual e usam um enfoque superficial com base na estatística somente.

A extração de atributos é um passo importante na hora de se criar um classificador de papéis retóricos. Teufel (1999) define um total de 12 tipos de atributos (ver Tab. 1). Tanto o MAZEA quanto o AZEA se baseiam na extração de 6 dos 12 atributos na criação dos classificadores multirrótulo e binário, respectivamente: *tamanho*, *posição*, *tempo*, *voz*, *modal* e *expressão padrão*.

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

Atributo	Descrição	Valores
TF-IDF	A sentença contém termos significativos determinados pela medida TD-IDF?	Sim ou não
Título	A sentença contém palavras que ocorrem no título?	Sim ou não
Tamanho	A sentença possui um maior número de palavras que um limiar definido?	Sim ou não
Posição Texto	Posição da sentença no texto em relação a 10 segmentos	Faixas de A a J
Posição Parágrafo	Posição da sentença dentro de um paragrafo	Começo, meio e fim
Tempo	Tempo do primeiro verbo flexionado	9 tempos verbais
Voz	Voz do primeiro verbo flexionado	Ativa ou passiva
Modal	Presença de verbos auxiliares modais	Sim ou não
Citação 1	A sentença contém alguma citação ou nome de autor?	Citação, autor, nenhum
Citação 2	A sentença contém uma autocitação?	Sim, não, nenhum
Citação 3	Posição da citação na sentença	Começo, meio, fim, nenhum
Expressão Padrão	Primeira Expressão Padrão (EP) na sentença	20 tipos de EP

Tabela 1: Atributos definidos por Teufel (1999)

O atributo *TF-IDF* visa identificar termos significativos que são frequentes numa sentença, mas que são raros nas outras sentenças do documento. Tal atributo é bastante usado na tarefa de Recuperação de Informação (Salton e McGill, 1986). Assim, por exemplo, uma sentença que contenha termos específicos relacionados à metodologia de pesquisa pode representar o papel retórico METHOD. As palavras que ocorrem no *título* também são boas candidatas para identificar papéis retóricos específicos no texto. Por exemplo, em PURPOSE, costuma-se colocar palavras relacionadas ao título. De mesma forma, as sentenças que contêm termos significativos ou palavras do título podem ser bons indicadores do *macroaspecto* WHAT.

O *tamanho* da sentença é um atributo trivial, já que não está relacionado diretamente aos papéis retóricos. Mas, mesmo assim, deve ser considerado porque indica a complexidade da sentença. A *posição* da sentença no texto e nos parágrafos é muito importante. Normalmente, as primeiras sentenças de um *abstract* científico descrevem BACKGROUND e PURPOSE; já as últimas descrevem RESULT e CONCLUSION. Da

mesma forma, existem aspectos que sempre ocorrem na primeira sentença dos textos, como WHAT, WHEN e WHERE.

Os atributos *tempo*, *voz* e *modal* representam a morfossintaxe do verbo. Os textos científicos seguem um padrão de escrita bem definido, portanto, a análise do primeiro verbo da sentença é de muita importância. Por exemplo, nas sentenças do tipo BACKGROUND, o primeiro verbo costuma estar no tempo presente. Em textos jornalísticos, o padrão de escrita é diferente, sendo que todos os verbos da sentença possuem a mesma importância.

As *citações* são referências que indicam o trabalho de outras pessoas. Tais trabalhos representam o estado da arte ou os trabalhos relacionados ao foco da pesquisa. Esse atributo é um claro identificador do papel retórico BACKGROUND. No gênero jornalístico, praticamente não existem esses tipos de citações, portanto, esse atributo não será útil para identificar aspectos. Já as *expressões padrão* são combinações de palavras frequentes, as quais estão relacionadas aos papéis retóricos. Por exemplo, as expressões "acredita-se que" e "visa-se" comumente são sinalizadoras de PURPOSE.

Souza e Feltrim (2013) desenvolveram o modulo de análise de coerência do sistema de auxílio à escrita científica Scipo[8] (*Scientific Portuguese*) (Feltrim, 2004). Tal módulo visa identificar aspectos de coerência em *abstracts* por meio de classificadores de relações semânticas entre sentenças, com a finalidade de fornecer sugestões ao usuário para que este possa melhorar a sua escrita.

Compilaram-se um total de 385 *abstracts* científicos escritos em língua portuguesa por estudantes de graduação. Tais *abstracts* foram anotados com mesmos papéis retóricos descritos no trabalho de Dayrell et al. (2012) (Contexto, Lacuna, Objetivo, Metodologia, Resultado e Conclusão). Na criação dos classificadores, no total, extraíram-se 13 atributos, sendo alguns deles definidos por Teufel (1999): papel retórico da sentença (atual, anterior e posterior), presença de anáfora, *posição* da sentença, *expressões padrão*, *tamanho* da sentença, tamanho do título e 5 atributos de similaridade semântica fornecidos pelo LSA (*Latent Semantic Analisys*) (Landauer et al., 1998). O algoritmo de classificação foi o SMO (Platt, 1998). A avaliação do modulo de coerência foi feita pelos mesmos estudantes, sendo que 62.00% deles consideraram relevantes as sugestões

[8] http://www.nilc.icmc.usp.br/scipo-farmacia/

oferecidas pelo módulo e 13.00% consideraram irrelevantes.

2.4 Sumarização com base em Aspectos

Até o momento, os principais trabalhos de SA multidocumento com base em aspectos foram feitos para a língua inglesa. Tais trabalhos estão disponíveis nas publicações oficiais da TAC 2010[9] e 2011[10]. A seguir, descrevem-se alguns dos principais métodos de sumarização multidocumento que utilizaram aspectos informativos como guia para sumarizar. Tais métodos foram apresentados na TAC 2011[11], já que foi o último ano em que se realizou a tarefa de sumarização[12]. No total, participaram 50 equipes. Cabe ressaltar que cada trabalho é identificado pelo ID da equipe participante.

Para avaliar os sumários automáticos gerados pelos participantes, a TAC 2011 forneceu um conjunto de dados de teste (em inglês, *Test Data*). O conjunto de dados é composto por aproximadamente 44 tópicos, divididos nas cinco categorias definidas pela TAC: "Acidentes e desastres naturais", "Ataques", "Saúde e segurança", "Recursos naturais ameaçados" e "Julgamentos e investigações". Cada tópico possui um ID, a categoria respectiva, um título e 20 documentos divididos em 2 conjuntos: *A* (10 documentos) e *B* (10 documentos). Os documentos do conjunto *A* precedem cronologicamente aos documentos do conjunto *B*. Nota-se que o conjunto *B* também foi criado para avaliar a tarefa Sumarização de Atualização. Os sumários de referência foram criados por anotadores humanos usando o critério de balanceamento de cobertura de aspectos para cada categoria específica. O tamanho de cada sumário é de 100 palavras aproximadamente.

2.4.1 lrlab2011

A equipe *lrlab2011* composta por Makino et al. (2011) realizou um trabalho visando gerar sumários multidocumento que envolvam todos os aspectos possíveis para uma categoria/domínio específico. Por exemplo, um sumário da categoria "Desastres Naturais" deve conter no mínimo os aspectos WHAT, WHEN, WHERE, WHO_AFFECTED, DAMA-

[9] http://www.nist.gov/tac/publications/2010/index.html
[10] http://www.nist.gov/tac/publications/2011/index.html
[11] http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html
[12] http://www.nist.gov/tac/tracks/index.html

GES e COUNTERMEASURES. Na fase de identificação de aspectos, os autores adotaram uma abordagem de Aprendizado de Máquina (AM) utilizando um classificador de Máxima Entropia (também chamado de Regressão Logística) (Shannon, 2001) para calcular a cobertura de um aspecto na sentença e, assim, predizer se a sentença reflete o aspecto ou não. A ideia principal é de maximizar os aspectos com as pontuações mais baixas para que o sumário possa abranger todos os aspectos possíveis.

O problema de cobertura máxima (em inglês, *maximum coverage*) é um problema NP-completo em que, dados vários conjuntos, visa-se selecionar os *k* conjuntos que possam cobrir o máximo número de elementos em comum (Hochbaum, 1997). Em teoria da decisão, o método *max-min* é utilizado tanto para maximizar o ganho mínimo quanto para minimizar a perda máxima possível. Assim, utiliza-se o método *max-min* como solução do problema de cobertura máxima de aspectos. Por exemplo, na categoria "Acidentes", o aspecto COUNTERMEASURES é muito pouco referenciado pelas sentenças dos textos-fonte. Para solucionar essa questão, o método *max-min* realiza um balanceamento da cobertura dos aspectos, fazendo que a sentença que contenha o aspecto COUNTERMEASURES possa ser considerada parte do sumário final.

Para a identificação de aspectos, o córpus utilizado, tanto para treino quanto para teste, é composto por artigos da Wikipedia anotados manualmente com aspectos informativos. Foram extraídos vários atributos (em inglês, *features*). Uma representação dos atributos é a co-ocorrência de n-gramas no córpus, dada pela seguinte equação:

$$\phi_k(j, y) = \begin{cases} 1, & \text{se o n-grama } k \text{ ocorre em } j, \text{ e } y = a; \\ 0, & \text{caso contrário.} \end{cases} \quad (1)$$

Nesse caso, os atributos são todos os unigramas e bigramas do córpus, sendo que o valor do atributo, denominado $\phi_k(j, y)$, é 1, se a sentença j contém o n-grama k e se o rótulo y é o aspecto a a ser anotado, caso contrário, o valor é 0. Outros atributos utilizados são a *localização* da sentença no texto e o *tamanho* da sentença. Assim, obtêm-se os atributos que representarão o conjunto de instâncias de treinamento e de teste na criação do classificador de aspectos.

O classificador de Máxima Entropia determina a probabilidade condicional de um aspecto ser refletido ou não na sentença. Cada aspecto recebe uma pontuação dada

pela soma de todas as pontuações desse aspecto para cada sentença dos textos-fonte. O sumário final é criado com base no algoritmo *max-min*, em que se realiza um balanceamento das pontuações dos aspectos, selecionando as sentenças que contenham aspectos definidos para a categoria em questão.

Os resultados com respeito à identificação de aspectos não foram mostrados no artigo. Já na avaliação da informatividade dos sumários automáticos, utilizou-se a medida ROUGE. Embora os autores não especifiquem a medida de avaliação utilizada (precisão, cobertura, acurácia ou F1), eles afirmam que, para o conjunto A, obtiveram uma informatividade de ROUGE-2=0.1188 e ROUGE-SU4=0.1479. Já para o conjunto B, obtiveram uma informatividade de ROUGE-2=0.0850 e ROUGE-SU4=0.1172. Os resultados obtidos tanto para o conjunto A quanto para o conjunto B superam a média dos resultados de todos os participantes: ROUGE-2=0.0930 e ROUGE-SU4=0.1270 para o conjunto A, e ROUGE-2=0.0700 e ROUGE-SU4=0.1094 para o conjunto B. Observa-se que, para o conjunto B, os resultados diminuíram.

2.4.2 PolyCom

A equipe *PolyCom* formada por Zhang et al. (2011) propôs um sistema de SA multidocumento com base em aspectos composto de três módulos: (i) identificação dos aspectos nas sentenças, (ii) modelagem dos aspectos com o *Hidden Markov Model* (HMM) (Baum e Petrie, 1966), e (iii) seleção e ordenação das sentenças anotadas. O primeiro módulo trata o problema de classificação como multirrótulo, extraindo os atributos que formarão as instâncias de treinamento e teste. O segundo módulo visa modelar a coerência textual por meio do HMM. Após a modelagem, selecionam-se e organizam-se as sentenças com a melhor pontuação dada pelo HMM, que formarão o sumário final.

Na identificação de aspectos, os autores se basearam nos trabalhos de Teufel e Moens (1999, 2002). Nesses trabalhos, narra-se a criação de um classificador em nível sentencial que possa identificar relações retóricas em textos de gênero científico (p.ex., introdução, proposta, resultados, etc.). Para extração de atributos, foram definidas as denominadas meta-frases (em inglês, *meta-phrases*), as quais representam uma 2-tupla (m_1, m_2), em que m_1 é uma palavra ou uma frase. Cada frase pode estar formada tanto por categorias sintáticas quanto por entidades nomeadas. As categorias sintáticas

representam os atributos das palavras numa sentença, incluindo constituintes lógicos (p.ex., /PRED/ para predicado, /ARG/ para argumento) e classes gramaticais (p.ex., /dobj/ para objeto direto, /agent/ para agente, /nn/ para substantivo). Já as entidades nomeadas representam características semânticas (p.ex., /PER/ para pessoa, /ORG/ para organização). Caso não tiver nenhuma entidade nomeada, é atribuído um valor /NULL/.

No momento da classificação, cada meta-frase identificada nas sentenças terá um relacionamento com um determinado aspecto. Por exemplo, o segmento "o presidente disse", representado pela meta-frase (/agent/, "disse") ou (/PER/, "disse"), pode ser classificado com o aspecto WHO_AGENT. A seguir, mostram-se exemplos em inglês citados pelos autores, das combinações de meta-frases, tanto para categorias sintáticas quanto para entidades nomeadas:

$$\textit{linked fen-phen} = \begin{cases} (\text{/PRED/}, \text{/dobj/}), \\ (\text{/PRED/}, \text{'fen-phen'}), \\ (\text{'linked'}, \text{/dobj/}), \\ (\text{'linked'}, \text{'fen-phen'}) \end{cases} \tag{2}$$

$$\textit{Mayo Clinic study} = \begin{cases} (\text{/ORG/}, \text{'study'}), \\ (\text{/ORG/}, \text{/NULL/}), \\ (\text{'Mayo Clinic'}, \text{'study'}), \\ (\text{'Mayo Clinic'}, \text{/NULL/}) \end{cases} \tag{3}$$

Devido ao fato de uma sentença poder estar associada a vários aspectos, a classificação torna-se um problema multirrótulo. Para solucionar esse problema, utilizaram-se duas técnicas de transformação de problemas: Combinação de Rótulos (em inglês, *Label Combination*) e Decomposição Binária (em inglês, *Binary Descomposition*) (Boutell et al., 2004; Tsoumakas e Katakis, 2007). Em seguida, utilizaram-se dois algoritmos de aprendizado semissupervisionado: SVM transdutivo e SVM indutivo (Vapnik, 1995). Ambos os algoritmos predizem novas etiquetas de teste a partir do conhecimento fornecido pelos próprios dados de treinamento e de teste.

A coerência textual foi modelada por meio do HMM, em que os tópicos dos textos-fonte representam os estados e as sentenças representam as sequências já observadas. Nesse cenário, utilizaram-se aspectos informativos como componentes semânticos inseridos dentro de cada tópico. Com o modelo HMM já aprendido, é possível determinar o ordenamento das sentenças e selecionar as sentenças mais prováveis entre todas as possíveis permutações geradas pelo modelo. Logo, calcula-se a pontuação de cada aspecto, dada pela função objetivo do classificador SVM transdutivo ou indutivo. Além disso, foi calculada a pontuação de frequência da sentença sobre um conjunto de documentos. A pontuação final da sentença é computada pela combinação das pontuações anteriores. Por último, as sentenças selecionadas com as melhores pontuações são organizadas, escolhendo a maior probabilidade de uma permutação de sentenças.

Para avaliar a etapa de identificação, utilizou-se o córpus da anterior competição da TAC, do ano 2010. Em seguida, criou-se uma lista pré-definida de aspectos para cada categoria. No total, anotaram-se manualmente 2000 sentenças: 90% para treino e 10% para teste. Os resultados para a etapa de identificação de aspectos foram bons em comparação aos resultados dos outros participantes. O classificador SVM transdutivo com Decomposição Binária obteve o melhor resultado com uma medida macro-média F de 0.287, nas cinco categorias definidas pela TAC.

Para avaliar os sumários gerados pelo sistema, usou-se o córpus de teste da TAC 2011. Os resultados da avaliação deram uma informatividade de ROUGE-2=0.1231 e ROUGE-SU4=0.1598. Observa-se que os autores não explicam para qual conjunto (A ou B) foi obtido o melhor resultado. Porém, os autores afirmam que os resultados foram melhores que a média total dos participantes. Os resultados foram melhores que os resultados obtidos pela equipe *Irlab2011* (ver Seção 2.4.1).

2.4.3 Txsumm

A equipe *Txsumm* formada por Barrera et al. (2011) propôs um sistema de perguntas e respostas para auxiliar a tarefa de sumarização multidocumento. Basicamente, através dos aspectos informativos, o sistema gera um sumário orientado ao usuário respondendo perguntas tais como “o que aconteceu”, “quando aconteceu”, “onde aconteceu”, etc. O sistema segue uma metodologia de três fases: (i) limpeza da informação

desnecessária e identificação linguística, (ii) pontuação de sentenças, e (iii) extração de sentenças.

A primeira fase visa remover todas as etiquetas HTML e as manchetes inseridas nos textos-fontes do córpus de teste da TAC 2011. Em seguida, passa-se a etiquetar o córpus com aspectos linguísticos. Para isso, utiliza-se um sistema que possui um reconhecedor de entidades nomeadas (p.ex., pessoa, organização, local, data, dinheiro, etc.) denominado Jigsaw (Stasko et al., 2008). Também se utiliza a WordNet (Fellbaum, 1998), que oferece um conjunto de palavras que indicam os 5 níveis de sinonímia para uma lista de palavras relacionadas aos aspectos de uma categoria em particular. Por exemplo, na Fig. 6, mostra-se uma lista de palavras-chave definidas para os aspectos WHAT, WHO_AFFECTED, HOW, WHY e COUNTERMEASURES, da categoria "Saúde e segurança", alocadas no nível 0 de sinonímia na WordNet. Já no nível 1, os sinônimos dos hipônimos de cada palavra da lista serão adicionados ao conjunto final de palavras, e assim por diante para os outros níveis.

afetar, prevenção, vacinação, enfermidade, doença, vírus, demografia

Figura 6: Lista de palavras no nível 0, para a categoria "'Saúde e segurança"'

Na segunda fase, passa-se a dar uma pontuação para cada sentença. A seguir, apresentam-se os tipos de pontuação:

- **Penalidade de pronome:** a presença de sentenças que contêm pronomes (p.ex., nós, ele, ela, eles, etc.) sem uma referência direta é um problema na hora de extrair sentenças, já que o leitor não entende a quem ou a que estão se referindo as sentenças. Assim, uma penalidade é atribuída para cada sentença S, dada pela fórmula a seguir:

$$PronScore(S) = \frac{TotalPronounCount}{|S|} \tag{4}$$

  em que o numerador $TotalPronounCount$ representa a quantidade de pronomes identificados na sentença e o denominador $|S|$ representa o total de palavras da sentença.

- **Pontuação de entidades nomeadas:** segundo os autores, acredita-se que as sentenças que contenham uma grande quantidade de entidades nomeadas são mais prováveis de responder à maioria das perguntas feitas pelos aspectos. Portanto, uma pontuação será outorgada a cada sentença S dada pela formula a seguir:

$$NEWeight(S) = \sum_{\substack{n \epsilon S \\ n \epsilon C(D)}} \frac{nFrequencyCount(D)}{|D|} \tag{5}$$

em que n é a entidade presente na sentença S, e $C(D)$ são as entidades definidas para o conjunto de documentos D de uma categoria específica. Logo, o numerador $nFrequencyCount(D)$ representa o número de documentos do conjunto D em que ocorre a entidade n, enquanto o denominador $|D|$ representa o total de documentos. Por exemplo, a categoria "Saúde e segurança" dispõe das entidades pessoa/organização e dinheiro, associadas aos aspectos WHO_AFFECTED e COUNTERMEASURES, respectivamente. Porém, os aspectos WHAT, HOW e WHY não têm entidades nomeadas que possam identificá-los.

- **Pontuação da WordNet:** utilizou-se a WordNet (Fellbaum, 1998) para determinar os 5 níveis de sinonímia das palavras-chave relacionadas aos aspectos (ver Fig. 6). O objetivo é fornecer uma pontuação àquelas sentenças que contenham o conteúdo mais relevante para cada aspecto dado pela identificação dos sinônimos. Assim, por exemplo, a palavra-chave "vacinação" e seus sinônimos, nos distintos níveis, podem representar o aspecto COUNTERMEASURES. A pontuação da WordNet para uma sentença S é dada pela fórmula a seguir:

$$WNScore(S) = \sum_{\substack{w \epsilon S \\ w \epsilon L(C)}} \frac{1}{2^l} \tag{6}$$

em que w é a palavra contida tanto na sentença S quanto no conjunto total de sinônimos de cada categoria $L(C)$. O valor l indica o nível entre [0-4] em que a palavra w foi encontrada dentro do conjunto $L(C)$.

- **Pontuação M-SynSem:** baseado no sumarizador monodocumento de tipo extrativo SynSem (Barrera e Verma, 2011), cria-se um sumarizador multidocumento

chamando M-SynSem, que combina informações sintáticas e semânticas dos textos-fonte, tais como etiquetação morfossintática (POS), entidades nomeadas, remoção de *stopwords*, ranque de popularidade de palavras, desambiguação de palavras, *parser* sintático e uso da WordNet para análise semântica. O objetivo de atribuir uma pontuação M-SynSem para todas as sentenças é de explorar a eficácia do sumarizador SynSem para vários documentos. A pontuação dada pelo sistema para cada sentença S é definida como $MSynSemScore(S)$.

Por último, na terceira fase, computa-se a pontuação final de cada sentença por meio das fórmulas apresentadas anteriormente e extraem-se as sentenças melhor pontuadas. Nesse passo, o sistema faz dois cálculos dependendo de um limiar chamado $NamedEntityBox$. Tal limiar visa abranger a maior quantidade de aspectos definidos para uma categoria específica com base nas entidades nomeadas relacionadas àqueles aspectos, como já foi explicado na "Pontuação de entidades nomeadas". Basicamente, o $NamedEntityBox$ auxilia na seleção das sentenças que contenham a maioria das entidades nomeadas para formar parte do sumário extrativo, até chegar as 100 palavras do sumário final. O valor do $NamedEntityBox$ é definido para cada categoria, como o total de entidades nomeadas sobre o total de aspectos definidos para a categoria em questão. Por exemplo, a categoria "Ataques" possui 5 entidades nomeadas (data, local, pessoa, pessoa/organização e dinheiro) sobre 8 aspectos definidos (WHAT, WHEN, WHERE, PERPETRATORS, WHY, WHO_AFFECTED, DAMAGES e COUNTERMEASURES), sendo que a entidade data está relacionada com o aspecto WHEN, local com WHERE, pessoa com PERPETRATORS, pessoa/organização com WHO_AFFECTED e dinheiro com COUNTERMEASURES. Assim, o limiar seria $NamedEntityBox = 5/8$. Os cálculos para pontuar as sentenças são descritos a seguir:

$$SentScore_1(S) = (WNScore(S)) * NEWeight(S)) - PronScore(S) \quad (7)$$

$$SentScore_2(S) = (WNScore(S)) * MSynSemScore(S)) - PronScore(S) \quad (8)$$

A ideia da pontuação $SentScore_1$ é priorizar a presença de entidades nomeadas na sentença S combinando as pontuações WordNet e NEWeight. O primeiro passo é selecionar as sentenças que cumpram com o limiar $NamedEntityBox$. Uma vez que o

sumário extrativo E até então contenha a maioria das entidades nomeadas, a pontuação $SentScore_2$ prioriza as próximas sentenças resultantes da combinação das pontuações M-SynSem e WordNet, para, em seguida, adicioná-las ao resto do sumário E. A seguir descrevem-se os critérios de pontuação de sentenças:

$$FinalScore(S) = \begin{cases} SentScore1, & \text{se } |E| \leq NamedEntityBox; \\ SentScore2, & \text{se } |E| > NamedEntityBox \end{cases} \quad (9)$$

em que $|E|$ representa o total de palavras do sumário extrativo E (até o momento) sobre o total de palavras do sumário final (100 palavras, aproximadamente). Além disso, utilizou-se o algoritmo *Maximal Marginal Relevance* (MMR) (Carbonell e Goldstein, 1998) para remover as sentenças redundantes. Por último, o sumário é gerado pela ordenação das sentenças conforme as pontuações finais de cada sentença ($FinalScore(S)$).

Apesar dos autores não mostrarem os resultados exatos da ROUGE-2 e ROUGE-SU4, eles afirmam que, em comparação com o ano passado (TAC 2010), obtiveram uma melhora de 17% para o conjunto A e 7% para o conjunto B. Também afirmam que os resultados são melhores que 70% dos sistemas dos outros participantes.

2.5 Considerações Finais

Nesse capítulo, apresentaram-se os aspectos informativos como guia para sumarizar. Relataram-se três dos principais métodos de SA multidocumento com base em aspectos apresentados na competição da TAC 2011. Esses métodos foram escolhidos por utilizarem diferentes enfoques tanto na identificação de aspectos quanto na geração de sumários. Cabe ressaltar que esses trabalhos foram desenvolvidos para a língua inglesa. Neste trabalho de pesquisa, tomaram-se algumas ideias propostas dentro desses métodos como referência para desenvolver um sumarizador multidocumento com base em aspectos para a língua portuguesa.

Como já foi dito anteriormente, alguns aspectos são parecidos com os papéis retóricos. Portanto, com base nos atributos da literatura (Teufel e Moens, 2002; Feltrim et al., 2006; Genoves Jr. et al., 2007; Dayrell et al., 2012; Souza e Feltrim, 2013), visa-se criar um classificador que identifique automaticamente aspectos. Os atributos a serem utilizados são: *TF-IDF*, *título*, *posição*, *tamanho*, *tempo*, *voz*, *modal* e *expressão padrão*. Cabe

ressaltar que os textos a serem processados são do gênero jornalístico. Dessa forma, os atributos serão **adaptados** para o gênero em foco (ver Seção 4.2.1).

Do trabalho realizado pela equipe *lrlab2011* (Makino et al., 2011), ressalta-se a extração dos atributos: co-ocorrência de unigramas e bigramas do córpus (também chamado de *bag of words*), a *localização* da sentença no texto e o *tamanho* da sentença. Mesmo que os autores não expliquem o tipo de valores que foram atribuídos aos atributos *localização* e *tamanho*, pode se ter em consideração tais atributos na criação do classificador de aspectos.

Do trabalho da equipe *PolyCom* (Zhang et al., 2011), destaca-se a extração do atributo denominado meta-frase, que representa uma 2-tupla formada por uma palavra ou uma frase, em que a frase é representada por uma categoria sintática ou uma entidade nomeada. Da mesma forma, a equipe *Txsumm* (Barrera et al., 2011) dá ênfase à relação entre as entidades nomeadas e os aspectos. Portanto, para este trabalho de pesquisa, considerou-se como referência o uso de atributos sintáticos e de entidades nomeadas para identificar aspectos informativos.

Cabe ressaltar que os três métodos foram avaliados sobre o córpus oferecido pela TAC 2011, no contexto da competição da tarefa de Sumarização Guiada. Tal córpus está formado por textos-fonte escritos em língua inglesa. Já que nosso trabalho visa desenvolver e avaliar pelo menos um método de sumarização com base em aspectos para textos-fontes escritos em língua portuguesa, não será possível utilizar o córpus proposto pela TAC 2011. Em vez disso, utilizou-se o córpus jornalístico anotado manualmente com aspectos informativos denominado CSTNews (Cardoso et al., 2011). A medida utilizada para avaliar a informatividade dos sumários automáticos é a ROUGE (Lin, 2004). No seguinte capítulo, descrevem-se os recursos e ferramentas (a maioria deles feitos para o Português Brasileiro) utilizados neste trabalho de pesquisa.

CAPÍTULO

3

Recursos e Ferramentas para o Português

Neste capítulo, descrevem-se os recursos e ferramentas desenvolvidos para a língua portuguesa que foram utilizados neste trabalho de pesquisa. Como parte dos recursos linguísticos, encontra-se o córpus CSTNews (Aleixo e Pardo, 2008; Cardoso et al., 2011) (Seção 3.1), que será utilizado como córpus de teste tanto na identificação de aspectos quanto na construção/avaliação dos sumários automáticos. Outro recurso utilizado nesta pesquisa é o repósitorio de entidades nomeadas REPENTINO (Cardoso, 2008) (Seção 3.2). Como parte das ferramentas que auxiliarão o processo de identificação de aspectos e geração de sumários, encontra-se o *parser* sintático PALAVRAS (Bick, 2000) (Seção 3.3). Para a identificação de aspectos, particularmente, descreve-se o proceso de anotação de papéis semânticos (Seção 3.4) e reconhecimento de entidades nomeadas (Seção 3.5). Apresentam-se também dois dos melhores sumarizadores multidocumento para a língua portuguesa (Seção 3.6). Por último, apresentam-se as considerações finais deste capítulo (Seção 3.7).

3.1 O córpus CSTNews

O córpus CSTNews[1] (Aleixo e Pardo, 2008; Cardoso et al., 2011) é um recurso composto por coleções de textos-fonte de gênero jornalístico, construído com vistas à investigação da SA mono e multidocumento para o português brasileiro. É o primeiro córpus anotado com base no modelo *Cross-document Structure Theory* (CST) (Radev, 2000) para esta língua. O córpus contém 50 coleções de textos jornalísticos. Cada coleção engloba de 2 a 3 textos sobre um mesmo assunto. Os textos foram compilados manualmente dos jornais online *Folha de São Paulo*, *O Globo*, *Jornal do Brasil*, *Estadão* e *Gazeta do Povo*. As coleções foram classificadas em 6 categorias textuais: "Cotidiano", "Esporte", "Mundo", "Política", "Dinheiro" e "Ciência". Cada categoria contém uma determinada quantidade de coleções de textos jornalísticos (ver Fig. 7). Assim, foram identificadas 14 coleções da categoria "Cotidiano", 10 coleções da categoria "Esporte", 14 coleções da categoria "Mundo", 10 coleções da categoria "Política", 1 coleção da categoria "Dinheiro" e 1 coleção da categoria "Ciência".

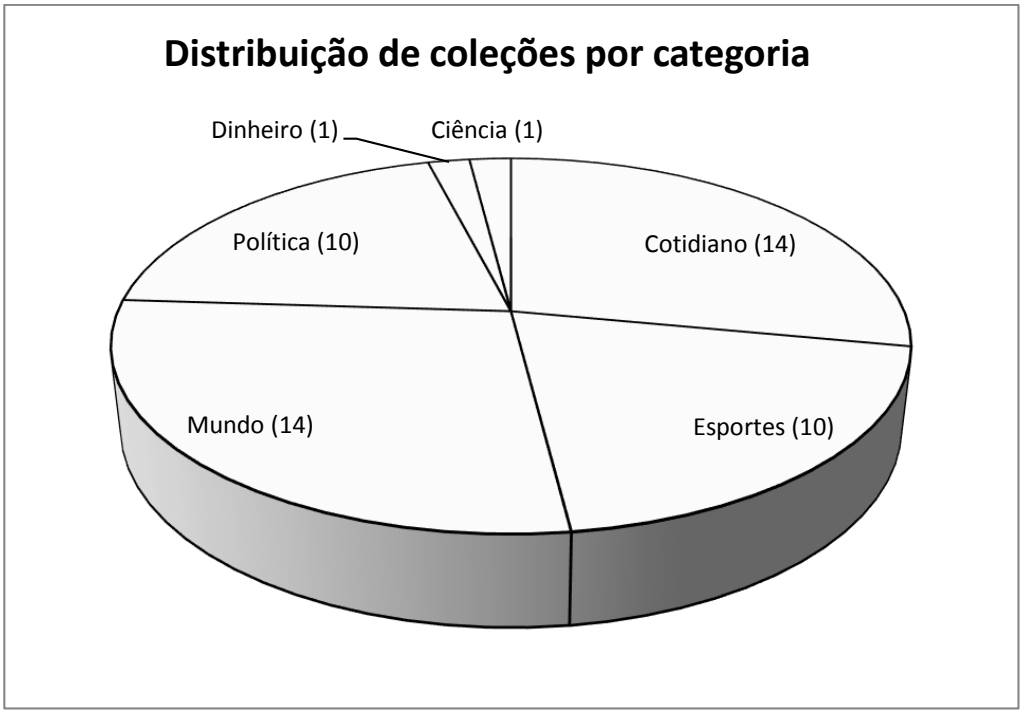

Figura 7: Distribuição das coleções por categoria

[1] http://www.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html

Além dos textos-fonte crus (ou seja, textos sem nenhum tipo de anotação), o córpus CSTNews disponibiliza sumários mono e multidocumento:

- Sumários manuais monodocumento para cada um dos textos do CSTNews, formando um total de 140 sumários.
- Um sumário automático multidocumento para cada coleção, produzido por um sistema computacional baseado em um método particular de SA, formando um total de 50 sumários automáticos.
- Um *abstract* manual multidocumento para cada coleção, formando um total de 50 *abstracts*.
- Um sumário extrativo manual multidocumento para cada coleção, formando um total de 50 sumários extrativos.
- Mais 5 sumários extrativos manuais multidocumento e mais 5 *abstracts* manuais multidocumento para cada coleção (Dias et al., 2014), formando um total de 300 sumários extrativos e 300 *abstracts*.

No córpus CSTNews, existem versões anotadas, em nível discursivo, dos textos-fonte com base na *Rhetorical Structure Theory* (RST) (Mann e Thompson, 1987) e na *Cross-document Structure Theory (CST)* (Radev, 2000), além de várias outras anotações. Na subseção seguinte, descreve-se a anotação manual de aspectos informativos sobre 50 sumários manuais multidocumento do córpus CSTNews, que consistem na anotação utilizada neste trabalho.

3.1.1 CSTNews: Aspectos

A tarefa de anotação de córpus é uma tarefa de classificação que consiste em atribuir um ou mais rótulos a uma unidade representativa do texto (palavra, sentença ou parágrafo, normalmente). A anotação de aspectos informativos foi feita por Rassi et al. (2013) em **nível sentencial** sobre sumários manuais multidocumento do córpus CSTNews. Para a tarefa de SA multidocumento, os aspectos podem indicar estruturas padrão para a modelagem de critérios de seleção e organização de conteúdo nos sumários.

As categorias no córpus CSTNews diferem das definidas originalmente na TAC 2010. Contudo, existem similaridades com as seis categorias consideradas (ver Fig. 7). Por exemplo, nas categorias “Cotidiano” ou “Mundo”, pode haver menção a “Acidentes e desastres naturais”.

A tarefa de anotação foi realizada por quatro subgrupos de anotadores compostos por 3 ou 4 linguistas computacionais, havendo um pesquisador sênior em cada subgrupo responsável pela coordenação da tarefa de anotação. Cada subgrupo ficou responsável pela anotação completa de uma das quatro categorias mais representativas, ou com maior quantidade de textos-fonte do córpus (“Cotidiano”, “Esportes”, “Mundo”, “Política”). Na **fase preliminar** de anotação, para ter uma referência consensual, foram anotados os sumários das categorias “Dinheiro” (1) e “Ciência” (1). Já na **fase final** de anotação, foram anotados os 48 sumários das categorias “Cotidiano” (14), “Esporte” (10), “Mundo” (14) e “Política” (10).

Com base na tarefa de anotação definida pela TAC, realizou-se um refinamento e definição dos aspectos em função das diferentes categorias sugeridas nos textos-fonte. Esse refinamento envolveu tanto a exclusão de algumas etiquetas originais quanto a inserção de novas etiquetas de interesse para os textos do córpus CSTNews. Assim, foram definidos 20 aspectos informativos (ver Tab. 2).

Macroaspectos	**Microaspectos**
COMMENT	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
COUNTERMEASURES	WHERE
DECLARATION	WHY
GOAL	HOW
HISTORY	SCORE
PREDICTION	SITUATION
SITUATION	GOAL
WHAT	
HOW	

Tabela 2: Aspectos gerais no córpus CSTNews

A necessidade de identificação de segmentos textuais em diversos níveis estruturais

para a determinação do aspecto correspondente resultou na classificação dos aspectos em *micro* e *macroaspectos*. Os *microaspectos* representam segmentos locais que compõem uma sentença. Os *macroaspectos* dependem do conteúdo sentencial em contexto. No total, foram identificados 11 *macroaspectos* e 9 *microaspectos*, apesar de haver alguma variação nesses conjuntos em função da categoria anotada (ver Apêndice B). Por exemplo, a categoria "Esportes" é a única que possui o *microaspecto* SCORE. Nota-se que os aspectos SITUATION, GOAL e WHO podem acontecer tanto como *macroaspectos* quanto como *microaspectos*.

Cabe ressaltar que a anotação de aspectos foi feita em **nível sentencial**, seguindo a metodologia da TAC, ou seja, os aspectos identificados são posicionados ao final da sentença. Na Fig. 8, mostra-se um exemplo de uma sentença anotada com aspectos da categoria "Mundo". Com respeito aos *macroaspectos*, descreve-se o acontecimento de um desastre natural (WHAT) e a declaração emitida pelo jornal japonês pró-Pyongyang (DECLARATION). Com respeito aos *microaspectos*, informa-se que o fato aconteceu no mês de julho (WHEN), na Coréia no Norte (WHERE), por causa das enchentes (WHY), deixando muitas pessoas mortas e outras feridas (WHO_AFFECTED).

[Ao menos 549 pessoas morreram, 3.043 ficaram feridas e outras 295 ainda estão desaparecidas em consequência das enchentes que atingiram a Coréia do Norte em julho, segundo um jornal japonês pró-Pyongyang.] **WHAT/DECLARATION/WHEN/WHERE/WHY/WHO_AFFECTED/**

Figura 8: Sentença anotada do sumário da coleção C1 do córpus CSTNews

Na anotação, foi relevante distinguir aspectos que transmitem informações principais daqueles relativos a informações secundárias. Diante disso, aos aspectos podia ser adicionado o sufixo EXTRA. Por exemplo, uma sentença é anotada como WHERE_EXTRA se possuir alguma informação de localidade que não se refere ao evento principal. Neste trabalho, não existe uma distinção entre ideias principais e secundárias. Portanto, os sufixos EXTRA foram ignorados, deixando os aspectos em suas formas originais (ver Tab. 2).

Além da anotação manual de aspectos, Rassi et al. (2013) fizeram uma análise de ocorrência de aspectos nos sumários do córpus CSTNews. Essa análise foi feita por

dois motivos: (i) a posição de um aspecto, em textos jornalísticos, pode indicar sua relevância; (ii) a ocorrência significativa de um aspecto em uma certa coleção pode indicar que seu conteúdo é relevante para a modelagem computacional.

No caso (i), é sabido, por exemplo, que as informações que remetem ao tópico principal aparecem, em geral, logo no início, nas chamadas *sentenças lead*. As colunas da Tab. 3 e 4, referem-se as posições das sentenças nos sumários. Calculou-se que os sumários da categoria "Cotidiano" contêm um total de 13 sentenças, aproximadamente. Na Tab. 3, mostra-se uma pequena amostra em que os *microaspectos* WHO_AGENT, WHEN e WHERE ocorrem com maior frequência na primeira sentença (S1) de todos os sumários da categoria "Cotidiano"; no entanto, na Tab. 4, os *macroaspectos* DECLARATION e WHAT também ocorrem com alta frequência na primeira sentença (S1). A distribuição de aspectos por sentenças pode indicar esquemas que servirão de modelos para a geração automática de sumários.

Microaspectos	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**Total**
WHO_AGENT	7	3	4	1	3	0	1	1	1	0	1	0	0	22
WHEN	12	1	0	0	0	0	0	0	0	0	0	0	0	13
WHERE	10	1	0	0	0	1	0	0	0	0	0	0	0	12

Tabela 3: *Microaspectos* por posição na categoria "Cotidiano"

Macroaspectos	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**Total**
WHO_AGENT	6	3	4	4	2	0	2	3	2	1	1	0	0	28
WHEN	0	6	2	2	1	2	2	3	2	1	1	0	0	22
WHERE	10	0	1	1	0	1	0	0	0	0	0	0	0	13

Tabela 4: *Macroaspectos* por posição na categoria "Cotidiano"

No caso (ii), as distribuições dos aspectos mais relevantes nas coleções de uma categoria específica podem indicar que esses aspectos devem ser parte do sumário final. Na Tab. 5, mostra-se a distribuição de alguns aspectos nos sumários das coleções da categoria "Esporte", sendo que os valores da tabela representam o número de ocorrencias do aspecto no sumário da coleção (p.ex: o aspecto WHO_AGENT ocorre 5 vezes na coleção C25). Obverva-se que os aspectos WHO_AGENT e WHAT têm uma distribuição considerável entre os sumários de todas as coleções. Usualmente, as notícias da

categoria “Esporte” descrevem um fato ou evento (WHAT) em que os atletas ou times (WHO_AGENT) apresentam o desempenho mencionado. O aspecto SCORE, apesar de ter uma distribuição média, é considerado exclusivo dessa categoria.

Aspectos	**C8**	**C38**	**C41**	**C24**	**C25**	**C27**	**C28**	**C31**	**C48**	**C19**	**Total**
WHO_AGENT	4	2	1	2	5	1	2	1	1	0	19
WHAT	1	1	1	1	1	1	1	1	1	1	10
SCORE	1	1	1	1	1	0	1	0	1	0	7

Tabela 5: Aspectos por coleção na categoria “Esportes”

Após a análise da anotação de aspectos informativos sob o córpus, obtiveram-se padrões estruturais (em inglês, *templates*) que servirão de guia para a construção de sumários. Basicamente, os *templates* são estruturas organizacionais, como resultado da formalização dos padrões de comportamento identificados nos sumários do córpus CSTNews, contendo regras e operadores definidos a fim de selecionar e organizar o conteúdo dos textos-fonte para gerar o sumário final.

Acredita-se que os padrões identificados se aplicam tanto à SA monodocumento quanto à SA multidocumento. Salienta-se que esses padrões são dependentes de domínios de conhecimento variados do córpus, referenciados pelas categorias textuais. A seguir, apresentam-se os *templates* identificados para as quatro categorias principais do córpus CSTNews: “Cotidiano”, “Esportes”, “Mundo” e “Política”.

Padrões da categoria “Cotidiano”

Na categoria “Cotidiano”, observou-se que não há um grupo de aspectos comum para todos os sumários, provavelmente devido à variedade de temas que eles envolvem. No entanto, há um grupo que pode caracterizar a maioria deles, que apresenta uma ordenação parcial clara, porém, não necessariamente em sequência direta, ou seja, um imediatamente após o outro. Na Tab. 6, apresentam-se os aspectos mais frequentes para a maioria dos sumários, ou seja, aqueles que ocorrem no primeiro parágrafo do sumário, e sua ordenação parcial ($X < Y$ indica que o aspecto X ocorre em uma sentença anterior à sentença em que ocorre o aspecto Y no texto). Observa-se que os *macroaspectos* COMMENT, COMPARISON, CONSEQUENCE, COUNTERMEASURES, GOAL,

HISTORY, PREDICTION e SITUATION. Assim como os *microaspectos* WHO_AFFECTED, WHY e HOW, não aparecem na Tab. 6 por serem pouco frequentes e, por isso, não representativos na formação do sumário final. Da mesma forma, os aspectos pouco frequentes não serão considerados nas categorias restantes.

Em comum	WHAT, WHERE, WHEN, WHO_AGENT, DECLARATION
No 1ro parágrafo	WHAT, WHERE, WHEN
Ordenação parcial	WHAT < WHERE < WHEN

Tabela 6: Padrões nos sumários da categoria *Cotidiano*

Padrões da categoria "Esportes"

Na categoria "Esportes", observa-se também que alguns aspectos são mais comuns que outros e que existe uma ordenação parcial entre eles que também não obedece a uma sequência direta nos sumários. Porém, isso não ocorre para todos os casos. Por essa razão, apresentam-se separadamente na Tab. 7 os dados comuns a todos os sumários analisados e os dados comuns à maioria deles. Os sumários dessa categoria sempre apresentam eventos envolvendo aspectos WHO_AGENT e WHAT em parágrafos iniciais, além de sempre aparecerem nessa ordem. Para a maioria dos sumários, nota-se que ambos os aspectos antecedem os *macroaspectos* CONSEQUENCE e SITUATION e os *microaspectos* SCORE e WHERE. Salienta-se que alguns aspectos definidos para essa categoria não foram identificados por causa da sua baixa ocorrência nos sumários anotados.

Para todos os sumários	
Em comum	WHO_AGENT , WHAT
No 1ro parágrafo	WHO_AGENT , WHAT
Ordenação parcial	WHO_AGENT $<$ WHAT
Para a maioria dos sumários	
En comum	WHO_AGENT, WHAT, SCORE, CONSEQUENCE, SITUATION, COMMENT, WHEN, WHERE
No 1ro parágrafo	WHO_AGENT, WHAT, SCORE, CONSEQUENCE, SITUATION, COMMENT, WHEN, WHERE
Ordenação parcial	• WHO_AGENT $<$ WHAT • WHO_AGENT, WHAT $<$ SCORE • WHO_AGENT, WHAT $<$ CONSEQUENCE • WHO_AGENT, WHAT $<$ SITUATION • WHO_AGENT, WHAT $<$ WHERE • WHO_AGENT, WHAT, SCORE $<$ CONSEQUENCE

Tabela 7: Padrões nos sumários da categoria “Esportes”

Padrões da categoria “Mundo”

Na categoria “Mundo”, identificaram-se vários assuntos ou domínios de conhecimento nos sumários. Na Tab. 8, apresentam-se os padrões reconhecidos para os quatro assuntos: “Acidentes”, “Ataques”, “Decisões legais e políticas” e “Desastres naturais”. Nota-se a similaridade com as categorias definidas pela TAC (ver Seção 2.2). Observa-se, por exemplo, que, para todos os sumários sobre “Desastres naturais”, o aspecto WHAT sempre ocorre antes de WHERE, mas quando WHAT e WHERE coocorrem, eles sempre antecedem CONSEQUENCE. Sobre esse mesmo assunto, a maioria dos sumários apresenta COUNTERMEASURES e PREDICTION, sendo que não são aspectos comums nas outras categorias. É interessante notar que alguns dos aspectos identificados só aparecem em poucas subcategorias, caso de CONSEQUENCE, que só ocorre em “Desastres naturais”. Por último, a marca “—” indica que não se encontrou nenhum padrão nos sumários analisados.

	Acidentes	**Ataques**	**Decisões Legais e Políticas**	**Desastres Naturais**
Para todos os sumários				
Em comum	WHAT, WHERE, WHO_AFFECTED, WHY	WHAT, WHERE, WHO_AFFECTED, WHEN	WHAT, WHO_AGENT	WHAT, WHERE, WHO_AFFECTED, CONSEQUENCE
No 1ro parágrafo	WHAT, WHERE, WHO_AFFECTED	WHAT, WHERE, WHO_AFFECTED, WHEN	WHAT, WHO_AGENT	WHAT, WHERE
Ordenação parcial	WHAT < WHERE WHO_AFFECTED, WHAT, WHERE < WHY	WHAT < WHERE	—	WHAT < WHERE < CONSEQUENCE
Para a maioria dos sumários				
Em comum	—	WHO_AGENT, WHY, HISTORY	HISTORY, WHO_AFFECTED, WHERE, DECLARATION, GOAL	CONTERMEASURES, PREDICTION
No 1ro parágrafo	—	—	WHO_AFFECTED, WHERE	WHO_AFFECTED, CONSEQUENCE
Ordenação parcial	—	* WHO_AFFECTED < WHEN, WHERE * WHO_AGENT < HISTORY * WHEN < WHY * WHAT < WHO_AFFECTED	* WHO_AFFECTED, WHAT, WHO_AGENT < HISTORY * WHAT, WHO_AGENT < GOAL * WHO_AGENT, WHAT, WHERE < DECLARATION * WHO_AGENT < WHAT	* WHAT, WHERE < WHO_AFFECTED * WHAT, WHERE, CONSEQUENCE, WHO_AFFECTED < PREDICTION

Tabela 8: Padrões nos sumários da categoria “Mundo”

Padrões da categoria “Política”

A categoria “Política” abarca três diferentes estruturas textuais: a) textos sobre pesquisas eleitorais, em que se divulgam intenções de voto, usualmente por meio da comparação entre os candidatos, razão da alta frequência da etiqueta COMPARISON; b) textos que mencionam trocas de agressões verbais entre políticos ou candidatos, os quais possuem uma estrutura retórica diferente daquela geralmente presente em textos jornalísticos, devido às trocas de turnos e, como consequência, à grande frequência da etiqueta DECLARATION; c) textos que noticiam fatos ou eventos mais gerais relacionados aos personagens da vida política. Essas estruturas textuais, por trazerem suas

próprias cargas semânticas, indicam padrões distintos de ocorrência e ordenação de aspectos. Na Tab. 9, mostram-se os padrões identificados para essa categoria. Nota-se a falta do aspecto WHERE, pois não existe **nenhuma** anotação desse aspecto para essa categoria.

Para todos os sumários	
Em comum	WHAT, WHO_AGENT, WHAT
No 1ro parágrafo	WHAT
Ordenação parcial	—
Para a maioria dos sumários	
En comum	WHO_AGENT, WHAT, WHEN, WHY
No 1ro parágrafo	WHO_AGENT, WHAT, WHO_AFFECTED, WHEN
Ordenação parcial	• WHO_AGENT, WHAT, WHEN < PREDICTION • WHO_AGENT, WHAT, WHEN < COUNTERMEASURES • WHO_AGENT, WHAT, WHEN < DECLARATION • WHO_AGENT, WHAT, WHEN < HISTORY • PREDICTION < CONSEQUENCE

Tabela 9: Padrões nos sumários da categoria "Política"

3.2 O repositório REPENTINO

O REPENTINO[2] (**REP**ositório para reconhecimento de **ENTI**dades **NO**meadas) (Sarmento et al., 2006) é um repositório que contém exemplos de entidades nomeadas, ou seja, de entidades concretas ou abstratas que possuem um nome próprio. Esses exemplos encontram-se divididos em várias categorias conceituais organizadas em uma estrutura de árvore no formato XML. Na atualidade, o REPENTINO possui um total de 450000 exemplos, aproximadamente.

Até o momento, foram definidas 10 categorias conceituais: "Abstrações", "Arte/Mídia", "Natureza", "Eventos", "Documentos", "Locais", "Organizações", "Produtos", "Seres" e "Substâncias". Cada uma dessas categorias contêm diversas subcategorias[3]. Por exem-

[2] http://labclup.letras.up.pt/repentino/

[3] http://labclup.letras.up.pt/repentino/docs/docs.html

plo, a categoria “Seres” contém as subcategorias “Humano” (p.ex: “Othelo”, “Branca de Neve”, “Brad Pitt”, etc.) e “Coletivo Humano” (p.ex: “The Rolling Stones”, “Bonnie and Clyde”, etc.), entre outras. Já a categoria “Locais” contém as subcategorias “País/Estado” (p.ex: “Portugal”, “Estados Unidos”, “São Paulo”, etc.) e “Espacial” (p.ex: “Mercúrio”, “Via Láctea”, “Constelação de Capricórnio”, etc.), entre outras. O REPENTINO é também um recurso de construção coletiva que vai crescendo com ajuda de colaboradores especializados.

É importante dizer que o REPENTINO reúne exemplos de entidades nomeadas, ignorando informações acerca do contexto em que essa entidade é mencionada. As “entidades mencionadas” são aquelas entidades referenciadas a um determinado contexto, podendo assumir papéis semânticos diferentes em função desse contexto. Por exemplo, a entidade nomeada “Restaurante Universitário” normalmente é vista como organização. No entanto, pode ser mencionada em diferentes contextos, por exemplo, “nós comemos no Restaurante Universitário” (local) e “eu sou funcionário do Restaurante Universitário” (organização). Existem sistemas reconhecedores de entidades mencionadas, como o REMBRANDT (Cardoso, 2008), que será descrito na Seção 3.5.

Como parte da metodologia deste trabalho de pesquisa, foram criadas regras manuais para identificar aspectos informativos (ver Seção 4.1.2). O REPENTINO irá a auxiliar as regras que visam identificar expressões de localização referentes ao aspecto WHERE (ver Fig. 16). Portanto, somente se escolheu a categoria “Locais” com as suas subcategorias correspondentes.

3.3 O parser PALAVRAS

O parser PALAVRAS é um analisador sintático de textos em língua portuguesa baseado em regras, desenvolvido por Bick (2000). O PALAVRAS segue a metodologia da Gramática de Constituintes (em inglês, *Constraint Grammar*) introduzido por Karlsson (1990), a fim de resolver problemas de ambiguidade morfológica e mapear funções sintáticas por meio da dependência de contexto.

O *parser* pode transformar uma notação de Gramática de Constituintes (formato *flat*) em uma estrutura de árvore sintática tradicional (formato *tree*). Na Fig. 9, ilustra-

se um exemplo de anotação simples da sentença “O menino nada na piscina”. Dentro dos colchetes ([]), encontra-se a palavra na forma lematizada. Em seguida, aparecem os rótulos semânticos[4] entre os símbolos “<” e “>”. Logo depois, são anotadas as classes gramaticais, como substantivo (N), verbo (V), determinante (DET) e preposição (PREP). Junto com as classes gramaticais, estão as informações morfossintáticas indicando, por exemplo, que o verbo “nadar” está no tempo presente (PR), na terceira pessoa do singular (3S), do modo indicativo (IND), flexionado (VFIN). Por último, após o símbolo “@”, indicam-se as funções sintáticas. Por exemplo, a palavra “menino” foi marcada com @SUBJ, que indica o sujeito da oração.

```
O [o] <artd> DET M S @>N
menino [menino] <H> N M S @SUBJ>
nada [nadar] <fmc> <mv> V PR 3S IND VFIN @FS-STA
em [em] <sam-> PRP @<ADVL
a [o] <artd> <-sam> DET F S @>N
piscina [piscina] <Lh> N F S @P<
$.
```

Figura 9: Anotação de Gramática de Constituintes simples (*flat*)

Segundo seu autor, usando um conjunto de etiquetas gramaticais bastante diversificado, o *parser* alcança um nível de correção (ou exatidão) de 99% em termos de morfossintaxe (classe gramatical e flexão), e 97-98% em termos de sintaxe. Na prática, tem se verificado desempenho inferior a esse relatado. Neste trabalho de pesquisa, utilizou-se o *parser* PALAVRAS como fornecedor de informações léxicas, morfossintáticas e semânticas, tanto para a criação do classificador de aspectos usando técnicas de Aprendizado de Máquina (ver Seção 4.1.4 e 4.2.1), quanto para a criação de regras manuais (ver Seção 4.1.2 e 4.2.2).

[4] O PALAVRAS conta com um total de 157 rótulos semânticos, divididos em várias categorias: humano, organização, animal, lugar, etc. (http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html)

3.4 Anotação de Papéis Semânticos

Uma das formas de entender a semântica da sentença é analisar o comportamento do verbo em relação aos argumentos que a envolvem (Fillmore, 1968). Chamam-se de "papéis semânticos" às relações semânticas entre o verbo e seus argumentos. A tarefa de identificar os argumentos de um determinado verbo é chamada de Anotação de Papéis Semânticos (APS) (Gildea e Jurafsky, 2001, 2002).

Alva-Manchego (2013) propôs um sistema anotador de papéis semânticos para o português brasileiro que consta de 3 fases: (1) identificação do verbo alvo, (2) identificação de argumentos e (3) classificação de argumentos. Na Fig. 10, ilustra-se um exemplo do processo de anotação de papéis semânticos. Em primeiro lugar, identifica-se o verbo alvo "venceu" (v). Em seguida, identificam-se os argumentos (A) "a equipe brasileira", "a Finlândia" e "em Tampere". Por último, os argumentos são anotados com os papéis semânticos "A_0" (sujeito agente da oração), "A_1" (sujeito paciente da oração) e "*AM-LOC*" (local da ação), respectivamente. Cabe ressaltar que a terminologia "A/Arg" refere-se ao "argumento" identificado, seguido de um numero prototípico. Já a terminologia "AM/ArgM" refere-se ao argumento modificador, seguido do tipo de modificador, como tempo, local, maneira, causa, etc.

A equipe brasileira [venceu]$_v$ a Finlândia em Tampere. (1)

[A equipe brasileira]$_A$ [venceu]$_v$ [a Finlândia]$_A$ [em Tampere A]$_A$ (2)

[A equipe brasileira]$_{A_0}$ [venceu]$_v$ [a Finlândia]$_{A_1}$ [em Tampere]$_{AM\text{-}LOC}$. (3)

Figura 10: Exemplo de anotação de papéis semânticos

O resultado final do sistema, conforme a medida F1 (ver Seção 5.1), foi de 94.5% na fase de identificação de argumentos e 81.70% na fase de classificação de papéis semânticos.

Seguindo a mesma metodologia de Alva-Manchego (2013), Fonseca (2013) desenvolveu um anotador de papéis semânticos para a língua portuguesa brasileira, evitando a dependência de um *parser* sintático. O autor seguiu a arquitetura do sistema SENNA

(Collobert e Weston, 2008; Collobert et al., 2011), que obteve bons resultados para a língua inglesa. A partir da Wikipédia e do córpus PLN-BR (Bruckschen et al., 2008), Fonseca (2013) compilou um grande córpus de domínio geral, sobre o qual foi treinado um modelo de espaço vetorial para representar as sentenças de entrada do anotador. O anotador foi implementado com base no algoritmo de classificação supervisionada denominado Redes Neurais (Haykin, 1998). O resultado final, conforme a medida F1, foi de 79,10% na fase de identificação e 68.00% na fase de classificação. Observa-se que o anotador de papéis semânticos de (Alva-Manchego, 2013) é o melhor para o português brasileiro.

Os constituintes ou argumentos relacionados ao verbo podem responder a perguntas do tipo quem?, quando?, onde? e como?. No exemplo anterior, a resposta à pergunta "quem venceu?" seria "A equipe brasileira". Da mesma forma, as perguntas "quem foi vencido?" e "onde foi vencido?" seriam respondidas por "a Finlândia" e "em Tampere", respectivamente. Tais constituintes podem definir *microaspectos* informativos, como WHO_AGENT ("quem venceu"), WHO_AFFECTED ("quem foi vencido") e WHERE ("onde"), respectivamente. Assim, os papéis semânticos são, normalmente, similares aos *microaspectos*.

Dessa maneira, neste trabalho de pesquisa, propõe-se o uso do sistema APS para o português brasileiro feito por Alva-Manchego (2013) para identificar *microaspectos*. Na Tab. 10, apresentam-se as equivalências propostas entre alguns *microaspectos* e os papéis semânticos usados por Alva-Manchego (2013) e definidos por Palmer et al. (2010).

Microaspecto	**Papel semântico**	**Nome**	**Definição**
WHO_AGENT	A0 / Arg0	Agente	O sujeito da ação
WHO_AFFECTED	A1 / Arg1	Paciente	O afetado pela ação
WHERE	AM-LOC / ArgM-LOC	Local	Onde ocorreu a ação
WHEN	AM-TMP / ArgM-TMP	Tempora	Quando ocorreu a ação
HOW	AM-MNR / ArgM-MNR	Maneira	Como a ação foi realizada
WHY	AM-CAU / ArgM-CAU	Causa	Causa ou motivo da ação

Tabela 10: Equivalências entre *microaspectos* e papéis semânticos

Cabe ressaltar que, recentemente, Hartmann (2015) aprimorou o classificador de Alva-Manchego (2013), acrescentando o córpus com mais instâncias para treino e teste,

gerando, assim, um novo modelo de aprendizado. Porém, não foi considerado tal modelo por ser feito em paralelo ao desenvolvimento deste trabalho de pesquisa.

3.5 Reconhecimento de Entidades Mencionadas

Como já foi dito, as entidades nomeadas (EN) são entidades concretas ou abstratas referenciadas no texto por um nome próprio. O termo EN foi cunhado pela *Sixth Message Understanding Conference* (MUC-6) (Grishman e Sundheim, 1996). Já as entidades mencionadas (EM) são EN dependentes de contexto. Outros elementos também são considerados como EMs, como o caso das datas, por exemplo. O Reconhecimento de Entidades Mencionadas (REM) é uma subtarefa da Extração da Informação (EI) que visa identificar e classificar entidades do texto em categorias pré-definidas, tais como pessoa, organização, local, tempo, valor e acontecimento, entre outras categorias de interesse (Nadeau e Sekine, 2007).

Nesse contexto, é importante citar o HAREM[5] (Santos e Cardoso, 2007), que é um evento de avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para coleções de documentos em português organizado pela Linguateca[6]. No âmbito do primeiro HAREM[7], vários trabalhos foram apresentados, sendo o PALAVRAS_NER (Bick, 2007) o sistema que obteve os melhores resultados nas tarefas de identificação e classificação de EM. O PALAVRAS_NER baseia-se no analisador morfossintático PALAVRAS (ver Seção 2.2) para criar um conjunto de regras manuais que identificam EM nos textos. Já no contexto do segundo HAREM[8] , um dos sistemas "open source" com os melhores resultados foi o REMBRANDT (Cardoso, 2008).

O REMBRANDT[9] (**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **AN**álise **D**etalhada do **T**exto) é um sistema de REM e de Detecção de Relações entre Entidades (DRE) para o português. Segundo Cardoso (2008), o REMBRANDT explora intensamente a Wikipédia como fonte de conhecimento, e aplica um conjunto de regras gramaticais que aproveitam os vários indícios internos e externos das EM para extrair o

[5] Refere-se a **H**AREM de **A**valiação de **R**econhecedores de **E**ntidades **M**encionadas.

[6] http://www.linguateca.pt/

[7] http://www.linguateca.pt/primeiroHAREM/harem.html

[8] http://www.linguateca.pt/harem/

[9] http://xldb.di.fc.ul.pt/Rembrandt/

seu significado. Além disso, o REMBRANDT possui uma interface própria para interagir com a Wikipédia, a SASKIA, com o objetivo de facilitar as tarefas de navegação na estrutura de categorias, ligações e redirecionamentos da Wikipédia, com vista à extração de conhecimento. No contexto do segundo HAREM, o sistema REMBRANDT teve um desempenho de 56.74% de medida F1 na fase de REM e 45.02% na fase de DRE.

Na Fig. 11, ilustra-se uma sentença anotada pelo sistema REMBRANDT. Para efeito de visualização, as EM foram anotadas com as etiquetas "<EM></EM>", indicando a categoria (C) à qual a entidade pertence. Nessa sentença, a entidade "Bernardinho" foi identificada como PESSOA, a entidade "Finlândia" como LOCAL, as entidades "3" e "0" como NÚMERO e a entidade "Jogos Pan-Americanos" como ACONTECIMENTO. Nota-se que a entidade "Finlândia" foi anotada erroneamente como LOCAL porque, no contexto, faz referência a uma equipe de vôlei.

A equipe brasileira, comandada por <EM C="PESSOA"> Bernardinho </EM>, venceu a <EM C="LOCAL"> Finlândia </EM> por <EM C="NÚMERO">3</EM> sets a <EM C="NÚMERO"> 0 </EM>, nos <EM C="ACONTECIMENTO"> Jogos Pan-Americanos </EM>.

Figura 11: Exemplo de anotação do sistema REMBRANDT

Da mesma forma que os papéis semânticos, as categorias das entidades mencionadas podem definir alguns *microaspectos* informativos: o aspecto WHERE é equivalente a "local", WHEN a "tempo" e SITUATION a "acontecimento". Dessa maneira, neste trabalho de pesquisa, propõe-se o uso do sistema REMBRANDT na identificação automática de alguns *microaspectos*. Na Tab. 11, apresentam-se as equivalências propostas entre os *microaspectos* e as categorias das EM.

Microaspecto	**Categoria EM**
WHERE	local
WHEN	tempo
SITUATION	acontecimento

Tabela 11: Equivalências entre *microaspectos* e categorias EM

3.6 Sumarizadores Multidocumento

Nesta seção, descrevem-se dois dos melhores sistemas de sumarização multidocumento feitos para o Português do Brasil: o RSumm (Ribaldo et al., 2012), da abordagem superficial e o RC4 (Cardoso, 2014), da abordagem profunda. Esses sumarizadores serão utilizados na fase de avaliação com a finalidade de comparar nosso método com os melhores métodos da literatura para o Português Brasileiro.

3.6.1 RSumm

O RSumm é um sumarizador multidocumento de abordagem superficial com base em grafos desenvolvido por Ribaldo et al. (2012). Nesse trabalho, utilizou-se um enfoque híbrido, adequando-se o sistema de mapa de relacionamentos de Salton et al. (1997) com o modelo de relações CST (Radev, 2000). Também se investigaram algumas medidas derivadas de grafos para a seleção das sentenças que formam o sumário final.

Salton et al. (1997) modelam um texto simples (monodocumento) como um grafo não direcionado em que os vértices são parágrafos e as arestas são as relações de similaridade entre os parágrafos. Os algoritmos de seleção de conteúdo em grafos são: caminho denso (em inglês, *Bushy path*), caminho profundo (em inglês, *Depth-first path*) e caminho denso segmentado (em inglês, *Segmented bushy path*). Salienta-se que esses algoritmos são voltados para sumarização monodocumento, mas foram adaptados para um cenário multidocumento.

No caminho denso, a densidade de um vértice é definida como o número de conexões que este tem com o resto do grafo; assim o caminho é construído com os vértices mais densos ordenados cronologicamente (conforme aparecem no documento) para formar o sumário. O caminho profundo é similar ao caminho anterior, só que, em vez de se selecionarem os vértices mais relacionados, começa-se pelo vértice de maior densidade e, a partir dele, escolhem-se os filhos que têm mais ligações. Porém, o problema do caminho profundo é não cobrir todos os tópicos do documento; assim o caminho denso segmentado constrói diversos caminhos densos para cada tópico e, em seguida, concatena-os em ordem textual, garantindo que pelo menos um parágrafo de cada tó-

pico será selecionado para compor o sumário.

No RSumm, os textos/documentos foram modelados como grafos com ajuda do modelo CST. Assim, por exemplo, na Fig. 12, cada vértice é uma sentença (S) pertencente a um documento (D) e as arestas podem representar tanto as relações CST quanto alguma medida de similaridade, como *Maximal Marginal Relevance* (MMR) (Carbonell e Goldstein, 1998) ou similaridade de cosseno (Salton, 1988).

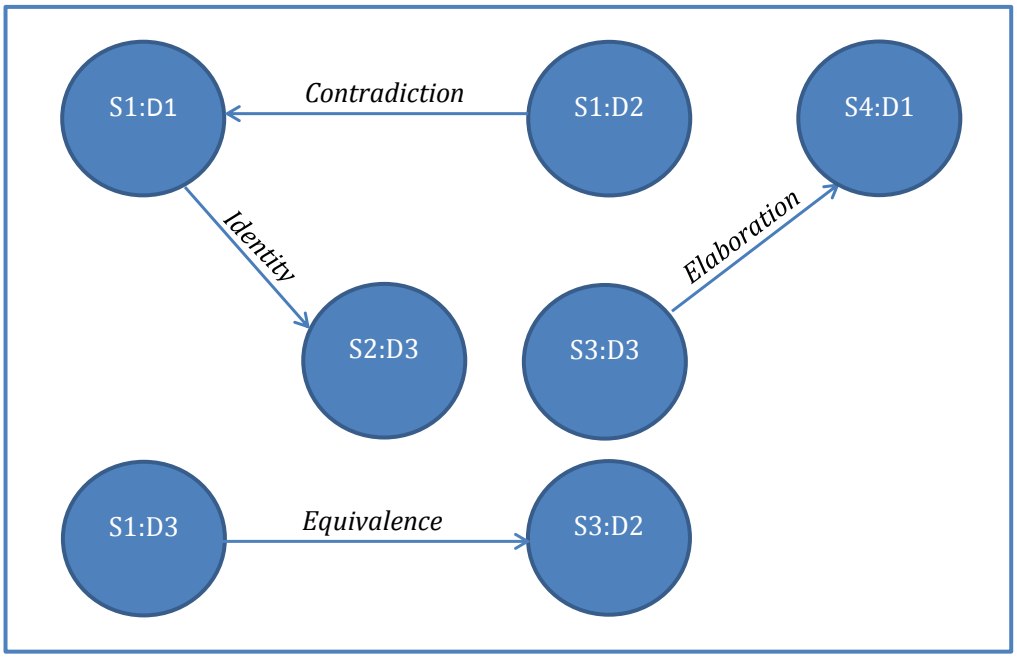

Figura 12: Exemplo de grafo com CST (Ribaldo et al., 2012)

As relações CST de alguma maneira ajudam a aprimorar a seleção das sentenças mais relevantes e desempatar sentenças com a mesma pontuação. Tal conhecimento foi usado de duas maneiras: a primeira somente considera o número total de relações por sentenças, sem considerar o tipo; e a segunda considera os tipos de relações dando um valor numérico a cada uma, conforme o seu nível de redundância. Em seguida, para selecionar o conteúdo que formará o sumário, escolhem-se somente os caminhos denso e profundo. Além disso, no começo, é feito um pré-processamento do texto, no qual se aplica lematização de palavras e eliminação de *stopwords*.

Por fim, o sumário é construído com as sentenças mais salientes (ou com as melhores pontuações) de todos os textos analisados. Para controlar a redundância entre as sentenças, utiliza-se a média dos valores de maior e menor cosseno do grafo. Incorpora-se,

também, o método de ordenação de sentenças pela posição da sentença no texto-fonte (Lima e Pardo, 2011, 2012). O critério de desempate entre as sentenças que possuem a mesma posição no texto-fonte é o seu tamanho em palavras, onde as sentenças menores devem aparecer antes no sumário. Além disso, o usuário é capaz de limitar o tamanho do sumário utilizando uma taxa de compressão de 70%, o que faz com que o sumário tenha 30% do número de palavras do maior texto-fonte.

Os resultados da avaliação dos sumários gerados pelo sistema mostram um bom nível de informatividade em comparação a outros sistemas de SA. Os resultados da ROUGE-L para o caminho denso foram: 0.4089 de precisão, 0.3704 de cobertura e 0.3871 de F1; e para o caminho profundo: 0.3977 de precisão; 0.3630 de cobertura e 0.3795 de F1. Cabe ressaltar que o RSumm é o melhor sistema de sumarização de abordagem superficial até o momento para a língua portuguesa.

Como já foi dito no inicio desta seção, utiliza-se o RSumm para comparar resultados na fase de avaliação. Além disso, será utilizado para **ranquear** as sentenças dos textos-fonte por relevância e **remover** as sentenças redundantes. Nesse caso, descartou-se o algoritmo de ordenação de sentenças de Lima e Pardo (2011, 2012), já que isso será feito conforme os padrões identificados na anotação de aspectos informativos (ver Seção 3.1.1). Destaca-se que as sentenças já terão sido previamente com aspectos mediante classificadores. Assim, selecionam-se as sentenças mais importantes anotadas com aspectos informativos. O processo de seleção de conteúdo será explicado detalhadamente na Seção 4.3.

3.6.2 RC4

O RC4 é um sumarizador multidocumento de abordagem profunda desenvolvido por Cardoso (2014). No trabalho de pesquisa dessa autora, investigou-se como modelar o processo de sumarização automática usando o conhecimento semântico-discursivo em métodos de seleção de conteúdo e o impacto disso para a produção de sumários mais informativos e representativos dos textos-fonte. Para isso, utilizaram-se as teorias semântico-discursivas RST e CST, além de subtópicos.

O córpus utilizado para avaliar os sumários foi o CSTNews, o qual foi anotado somente com relações RST e subtópicos, já que o córpus foi previamente anotado com relações CST no trabalho feito por Aleixo e Pardo (2008). No total, foram criados 13 sumarizadores com distintas configurações, sendo que o RC4 obteve os melhores resultados. Os resultados obtidos na avaliação dos métodos de sumarização automática indicaram que o uso do conhecimento semântico-discursivo como estratégias de seleção de conteúdo afeta positivamente a produção de sumários informativos. Cabe ressaltar que o RC4 é o melhor sistema de sumarização de abordagem profunda até o momento para a língua portuguesa.

3.7 Considerações Finais

Neste capítulo, apresentaram-se os recursos e ferramentas desenvolvidos para a língua Portuguesa utilizados no processo de sumarização. Tais recursos e ferramentas são utilizados tanto na fase de identificação de aspectos (*microaspectos* e *macroaspectos*) quanto na fase de seleção de conteúdo para gerar os sumários finais. O córpus CSTNews será utilizado em 3 etapas: (1) construção/avaliação do classificador de aspectos, (2) seleção de conteúdo com base nos padrões identificados nos sumários anotados do córpus, e (3) avaliação dos sumários automáticos, em comparação com os sumários humanos. É importante dizer que foi feita uma pesquisa para descobrir as melhores ferramentas existentes na literatura.

CAPÍTULO

4

Sumarização Multidocumento com base em Aspectos para o Português

Neste trabalho, adaptou-se uma abordagem profunda para produção de sumários multidocumento para o Português. Esta pesquisa contempla a análise do efeito que produzem as informações fornecidas pelos aspectos na informatividade dos sumários. Em particular, desenvolveram-se as três etapas da arquitetura genérica de um sistema SA (ver Fig. 4): identificação de aspectos informativos (na etapa de *análise*), formação de sumários (na etapa de *transformação*) e justaposição de sentenças (na etapa de *síntese*).

O objetivo principal desta pesquisa é investigar métodos de SA multidocumento usando informações significativas fornecidas pelos aspectos que ajudarão na seleção de conteúdo para a formação de sumários mais informativos. Para atingir o objetivo principal, cumpriram-se três objetivos específicos: (i) identificação de aspectos, (ii) seleção de conteúdo para formação de sumários (iii) e avaliação dos sumários gerados. Neste capítulo, explica-se a metodologia utilizada para atender os dois primeiros objetivos. Já a avaliação de sumários será explicada detalhadamente no capítulo seguinte.

Primeiramente, como parte da etapa de *análise* da arquitetura de um sistema SA, criou-se um classificador multirrótulo para identificar automaticamente aspectos infor-

mativos (*microaspectos* e *macroaspectos*). Para isso, utilizaram-se ferramentas da literatura para o Português, tais como anotador de papéis semânticos e reconhecedor de entidades nomeadas. Também se desenvolveram técnicas de aprendizado de máquina utilizando atributos da literatura. Além disso, para melhorar o desempenho do classificador, criaram-se regras manuais com base em padrões linguísticos obtidos através de uma análise detalhada das sentenças do córpus CSTNews anotado com aspectos. Os resultados da avaliação do classificador sobre o córpus CSTNews demostraram que os aspectos podem ser identificados automaticamente em textos jornalísticos com um desempenho razoável.

Em segundo lugar, como parte da etapa de *transformação*, desenvolveram-se alguns métodos de seleção de conteúdo para sumarização com base em padrões de aspectos em sumários. O conteúdo a ser selecionado são as **sentenças** provenientes dos textos-fonte. Supõe-se que essas sentenças foram previamente anotadas com aspectos informativos. Nesta etapa, utilizou-se o sumarizador RSumm (Ribaldo et al., 2012) para ranquear as sentenças por relevância e remover a redundância. Em seguida, com base nos padrões identificados na anotação manual de aspectos, criaram-se métodos para selecionar as sentenças relevantes **mais informativas**.

Por último, como parte da etapa de *síntese*, realizou-se uma justaposição das sentenças que formarão o sumário final.

Neste capítulo, explica-se detalhadamente o processo de identificação de *microaspectos* (Seção 4.1) e de *macroaspectos* (Seção 4.2). Depois, apresentam-se os métodos de seleção de conteúdo para formação de sumários (Seção 4.3). Também se apresenta um exemplo geral do processo (Seção 4.4). Por último, descrevem-se as considerações finais deste capítulo (Seção 4.5).

4.1 Identificação de Microaspectos

O processo de identificação automática de *microaspectos* foi dividido em 3 fases (ver Fig. 13). A seguir, explicam-se as fases do processo de identificação:

1. Compilar as sentenças dos 48 sumários anotados do córpus CSTNews das categorias “Cotidiano”, “Esporte”, “Mundo” e “Política”. Não foram consideradas as

categorias “Dinheiro” e “Ciência”, por terem poucos sumários anotados.

2. Anotar as sentenças com *microaspectos* usando 4 sistemas diferentes:

   (a) **Sistema APS (Anotador de Papéis Semânticos):** sistema que utiliza o APS de Alva-Manchego (2013) para anotar sentenças com *microaspectos* equivalentes aos papéis semânticos apresentados na Tab. 10. Abrangem-se os *microaspectos* WHO_AGENT, WHO_AFFECTED, WHEN, WHERE, WHY e HOW.

   (b) **Sistema APS + Regras:** sistema que usa regras desenvolvidas manualmente com base nos “falsos negativos e positivos” do sistema APS, com a finalidade de aprimorar o seu desempenho. Abrangem-se os *microaspectos* WHO_AGENT, WHO_AFFECTED, WHEN, WHERE, WHY e SCORE.

   (c) **Sistema REMBRANDT:** sistema que utiliza o reconhecedor de entidades nomeadas de Cardoso (2008) para anotar sentenças com *microaspectos* equivalentes às categorias das entidades mencionadas apresentadas na Tab. 11. Abrangem-se os *microaspectos* WHEN, WHERE e SITUATION.

   (d) **Aprendizado de Máquina (AM):** uso de técnicas de AM para criar um classificador de *microaspectos* com base em atributos léxico-semânticos. Atende todos os *microaspectos*, com exceção do GOAL.

3. Obter um conjunto de sentenças anotadas automaticamente com *microaspectos*.

Cabe ressaltar que o *microaspecto* GOAL não foi considerado por não ser identificado por nenhum dos sistemas utilizados. A seguir, descrevem-se detalhadamente os sistemas de anotação automática de *microaspectos*.

4.1.1 O Sistema APS

O sistema APS se baseia no classificador de papéis semânticos de Alva-Manchego (2013) com auxílio do *parser* PALAVRAS (Bick, 2000). Basicamente, o PALAVRAS gera as árvores sintáticas de cada sentença que servem de instâncias para que o classificador possa anotar os papéis semânticos correspondentes. Já no final, os papéis semânticos são mapeados nos *microaspectos* WHO_AGENT, WHO_AFFECTED, WHEN, WHERE,

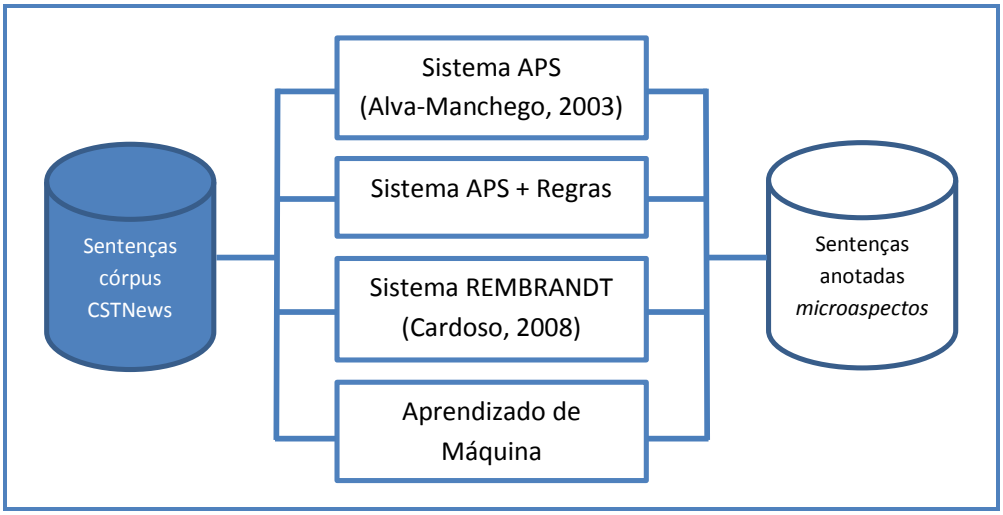

Figura 13: Metodologia do processo de identificação de *microaspectos*

WHY e HOW, conforme apresentado na Tab. 10. A seguir, relata-se o processo efetuado pelo sistema APS:

1. Dado um conjunto de sentenças a serem anotadas, utiliza-se o parser PALAVRAS a fim de gerar árvores sintáticas para cada sentença. Tais árvores são representadas em formato TigerXML[1] (ou *tree*). Neste passo, o PALAVRAS vai pré-processar as sentenças de entrada. Assim, pode se dar o caso de separação das contrações, por exemplo: “dos” em “de os”, “nesta” em “em esta”, “pelo” em “por o”, etc.

2. Executa-se um algoritmo que clona as árvores sintáticas conforme o número de verbos alvo da sentença. Os verbos auxiliares não são considerados. Desta maneira, se uma sentença possui três verbos alvo, a sua árvore sintática será clonada duas vezes, tendo-se, no total, três arvores sintáticas para a mesma sentença. Tais árvores são as instâncias de entrada do anotador de papéis semânticos de Alva-Manchego (2013).

3. Classifica-se cada árvore/instância da sentença pelo anotador de papéis semânticos de Alva-Manchego (2013).

4. Mapeiam-se os papéis semânticos nos *microaspectos* correspondentes.

[1] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html

5. Transforma-se o formato de saída convencional do classificador, denominado formato CoNLL[2], em um formato de rótulos "<aspect></aspect>". Assim, os *microaspectos* são anotados na sentença em um formato mais legível para o usuário.

6. Posicionam-se, por fim, os aspectos anotados no final da sentença.

Para exemplificar a saída do sistema APS, na Fig. 14, ilustra-se uma sentença anotada automaticamente com *microaspectos* (passo 5). Nota-se que o segmento "A equipe brasileira" foi anotado como WHO_AGENT por representar o sujeito gramatical agente da oração e por estar relacionado semanticamente ao verbo "vencer (venceu)". Já os segmentos "a Finlândia" e "em Tampere" foram anotados como WHO_AFFECTED (paciente) e WHERE (local), respectivamente.

<aspect APS="WHO_AGENT">A equipe brasileira</aspect> **venceu** <aspect APS="WHO_AFFECTED">a Finlândia</aspect> <aspect APS="WHERE">em Tampere</aspect>.

Figura 14: Sentença anotada com *microaspectos* pelo sistema APS (passo 5)

Por último, a anotação de *microaspectos* será feita em nível sentencial (passo 6). Isso significa que não importa se um *microaspecto* é anotado várias vezes numa mesma sentença (pela presença de vários verbos alvo), pois ele será indicado uma vez só no final da sentença (ver Fig. 15).

[A equipe brasileira venceu a Finlândia em Tampere.]**WHO_AGENT, WHO_AFFECTED, WHERE**

Figura 15: Anotação de *microaspectos* em nível sentencial do sistema APS (passo 6)

4.1.2 O Sistema APS + Regras

Tal sistema consiste de um conjunto de regras criadas manualmente com base nos padrões presentes nas sentenças identificadas como "falsos negativos" e "falsos positivos" da anotação feita pelo sistema APS. Os primeiros referem-se às sentenças cujos

[2] http://ilk.uvt.nl/conll/

aspectos o sistema APS não conseguiu classificar, mas que foram anotados manualmente. Já os segundos referem-se às sentenças que o sistema APS conseguiu classificar, mas que não foram anotados manualmente. Assim, criaram-se regras para os aspectos WHO_AGENT, WHO_AFFECTED, WHEN, WHERE e WHY. Porém, não foram criadas regras para os aspectos HOW, SITUATION e GOAL, já que não foram encontrados padrões que possam identificar esses aspectos. Já para o aspecto SCORE, também se criaram regras manuais, mesmo não tendo equivalência com algum papel semântico.

Tanto as regras do WHO_AGENT quanto as regras do WHO_AFFECTED estão baseadas somente nos "falsos positivos". Tais *microaspectos* representam a entidade "pessoa" ou "organização". Porém, o classificador de papéis semânticos não distingue se o agente/paciente da oração é representado por uma pessoa ou organização. Assim, existe uma grande possibilidade do classificador gerar muitos "falsos positivos". Para solucionar este problema, usaram-se as etiquetas semânticas para substantivos fornecidas pelo PALAVRAS para anotar PESSOA e ORGANIZAÇÃO (p.ex: "Hprof" de profissão, "Hfam" de família, "inst" de instituição, etc.). A ideia é apagar todos os segmentos identificados como *microaspectos* pelo sistema APS que não representem uma entidade (pessoa ou organização) e, assim, diminuir a quantidade de "falsos positivos" (ver Fig. 16). Já que existe a possibilidade do PALAVRAS anotar erradamente uma entidade de lugar como pessoa/organização (como é o caso da entidade "Coréia-do-Norte"), utiliza-se o léxico de local do repositório REPENTINO (Sarmento et al., 2006) para filtrar tais informações. O denominado "segmento APS", refere-se ao segmento da sentença anotado pelo sistema APS (p.ex: "<aspect APS=WHO_AGENT>O presidente</aspect>").

PESSOA = [H, HH, Hattr, Hbio, Hfam, Hideo, Hmyth, Hnat, Hprof, Hsick, Htit, hum]
ORGANIZAÇÃO = [admin, org, inst, media, party, suborg]
$\subseteq$ = "está contido em"
$\nsubseteq$ = "não está contido em"

Regra 1: Se o segmento APS contiver algum *token* associado a uma etiqueta semântica do tipo PESSOA/ORGANIZAÇÃO, e se o *token* não pertencer ao "léxico_de_local" do REPENTINO, então o segmento será corretamente anotado como WHO_AGENT/WHO_AFFECTED.

- Entrada: *"<aspect APS=WHO_AGENT>O presidente</aspect> diz que algumas de as obras já estão em andamento, <aspect APS=WHO_AGENT>outras</aspect> vão começar logo."*

  presidente_(Hprof) $\subseteq$ PESSOA $\nsubseteq$ léxico_de_local
  outras_(diff) $\nsubseteq$ PESSOA

  Saída: *"<aspect APS=WHO_AGENT>O presidente </aspect> diz que algumas de as obras já estão em andamento, outras vão começar logo."*

- Entrada: *"<aspect APS=WHO_AGENT> Ao menos 549 pessoas </aspect> morreram em consequência das enchentes que atingiram <aspect APS=WHO_AGENT> a Coréia-do-Norte </aspect> em julho"*

  pessoas_(H) $\subseteq$ PESSOA $\nsubseteq$ léxico_de_local
  Coréia-do-Norte_(hum) $\subseteq$ PESSOA $\subseteq$ léxico_de_local

  Saída: *"<aspect APS=WHO_AGENT>Ao menos pessoas</aspect> morreram em consequência das enchentes que atingiram a Coréia-do-Norte em julho."*

Figura 16: Regras do *microaspecto* WHO_AGENT/WHO_AFFECTED

Para o *microaspecto* WHEN, criaram-se regras que identificam automaticamente expressões de tempo. As regras do aspecto WHEN foram feitas com base na teoria de Baptista et al. (2008). Segundo os autores, as expressões de tempo foram organizadas em quatro grandes tipos: expressões de localização temporal (datas, horas e intervalos), de tipo TEMPO_CALENDARIO (p.ex: "no dia 7 de Julho", "ontem", "às 17:00", "entre 2013 e 2015", etc.); expressões de quantificação temporal, de tipo DURAÇÃO (p.ex: "todo o verão", "durante três meses", "por cinco anos", etc.); expressões de frequência, de tipo FREQUÊNCIA (p.ex: "diariamente", "todos os dias", "duas vezes por semana", etc.); e expressões temporais genéricas, de tipo GENÉRICO (p.ex: "Fevereiro", "inverno", etc.). Neste trabalho, desenvolveram-se **todas** as regras que identificam os quatro tipos de expressões temporais. Porém, para aprimorar o desempenho do sistema APS, utilizaram-se **somente** as regras que identificam expressões do tipo TEMPO_CALENDARIO, especificamente datas (ver Fig. 17), já que essas regras cobrem todos os casos possíveis dos "falsos negativos".

PREP = [de, em, a, por, para]

PRON = [ele(s), ela(s), este(s), esta(s), esse(s), essa(s), aquele(s), isto, isso, aquilo, aqui, aí, ali, ...]

ARTG = [a(s), o(s), um, uns, uma, umas, à(s)]

dia_da_semana = [segunda-feira, terça-feira, quarta-feira, quinta-feira, sexta-feira, sábado, domingo]

adverbio_de_tempo = [hoje, amanhã, ontem, anteontem, tarde, madrugada, noite, meia-noite, manhã]

lexico_de_tempo = [microssegundo, segundo, minuto, hora, dia, semana, mês, ano, década, milênio, semestre, bimestre, trimestre, época, tempo]

"+/-"= seguido_ou

Regra 1: Se a sentença tiver PREP + (PRON | ARTG) + averbio_de_tempo + PREP + (PRON | ARTG) + dia_da_semana +/- NÚMERO, então a sentença será anotada como WHEN.

"A chuva complicava o trânsito na manhã desta segunda-feira, 16."

na_(PREP+PRON) + manhã_(adverbio_de_tempo) + desta_(PREP+PRON) + segunda-feira_(dia_da_semana) + 16_NÚMERO

"Uma nova série de ataques criminosos foi registrada na madrugada desta terça-feira."

na_(PREP+PRON) + madrugada_(adverbio_de_tempo) + desta_(PREP+PRON) + terça-feira_(dia_da_semana)

Regra 2: Se a sentença tiver PREP + (PRON | ARTG) + dia_da_semana, então a sentença será anotada como WHEN.

"Um terremoto atingiu Japão nesta segunda-feira matando 9 pessoas."

nesta_(PREP+PRON) + segunda-feira_(dia_da_semana)

Regra 3: Se a sentença tiver PREP + (PRON | ARTG) +/- (TOKEN | NÚMERO) + lexico_de_tempo, então a sentença será anotada como WHEN.

"Aos 18 minutos, Maicon fez o primeiro gol."

Aos_(PREP+ARTG) + 18_(NÚMERO) + minutos_(lexico_de_tempo)

"No primeiro tempo houve outras jogadas ..."

No_(PREP +ARTG) + primeiro_(TOKEN) + tempo_(lexico_de_tempo)

"Os acontecimentos ocorreram nessa semana."

nessa_(PREP+PRON) + semana_(lexico_de_tempo)

Regra 4: Se a sentença tiver PREP + (PRON | ARTG) + avérbio_de_tempo, então a sentença será anotada como WHEN.

"A quarta medida foi aprovada nesta madrugada."

nessa_(PREP+PRON) madrugada_(avérbio_de_tempo)

Figura 17: Regras do *microaspecto* WHEN

Para o *microaspecto* WHERE, analisando os "falsos positivos", descobriu-se que o sistema APS costuma anotar erradamente expressões de tempo e advérbios de modo, como local. Para resolver isso, utilizaram-se todas as regras desenvolvidas para identificar expressões temporais e formou-se um léxico com algumas expressões de modo (p.ex: "em geral", "em vão", "em voz alta", etc.). A ideia é ignorar todos os segmentos erroneamente classificados como WHERE pelo sistema APS e, assim, diminuir a quantidade de "falsos positivos" (ver regra 1, Fig. 18). Além disso, com base nos "falsos negativos", criou-se uma regra que identifica expressões de local iniciadas pela preposição "em" (ver regra 2, Fig. 18).

Regra 1: Se o segmento APS contiver a PREPOSIÇÃO "em", seguida ou não de ARTIGO/PRONOME, seguido de um SUBSTANTIVO que não seja uma "expressão de tempo" ou um "advérbio de modo", então o segmento será corretamente anotado como WHERE.

- Entrada: *"Eu guardei as informações<aspect APS=WHERE>nesse computador</aspect>."*
  em_(PREPOSIÇÃO) + esse_(PRONOME) + computador_(SUBSTANTIVO)
  Saída: *"Eu guardei as informações<aspect APS=WHERE>nesse computador</aspect>."*
- Entrada: *"<aspect APS=WHERE>No domingo</aspect>, uma batalha sangrenta ocorreu."*
  em_(PREPOSIÇÃO) + o_(ARTIGO) + domingo_(expressão_de_tempo)
  Saída: *"No domingo, uma batalha sangrenta ocorreu."*
- Entrada: *"Eu pense <aspect APS=WHERE>em voz alta</aspect>."*
  em_(PREPOSIÇÃO) + voz alta_(adverbo_de_modo)
  Saída: *"Eu pense em voz alta."*

Regra 2: Se a sentença tiver a PREPOSIÇÃO "em" + expressão_capitalizada, então a sentença será anotada como WHERE.

"O senador Marcos nasceu em São Paulo"
em_(PREPOSIÇÃO) + São Paulo_(expressão_capitalizada)

Figura 18: Regras do *microaspecto* WHERE

Para o *microaspecto* WHY, criaram-se regras que identificam expressões de causa com base nos "falsos negativos e positivos" (ver Fig. 19). Cabe ressaltar que o léxico de causa pode ser ampliado. Por último, as regras do *microaspecto* SCORE foram criadas com base nos padrões identificados nas únicas 10 sentenças anotadas no córpus CSTNews (ver Fig. 20). Tais regras foram integradas no sistema APS + Regras.

léxico_de_causa = [por isso, com isso, porque, devido a, por causa de, por força de, em função de, em virtude de, em razão de, em decorrência de, em consequência de, pois, visto que, já que, ...]

Regra 1: Se a sentença tiver expressão "léxico_de_causa", então a sentença será anotada como WHY.

"O senador teve seu estado de saúde piorado, por causa de complicações gastrointestinais."

por causa de_(léxico_de_causa)

Regra 2: Se a sentença tiver PREPOSIÇÃO "por" + verbo_infinitivo, então a sentença será anotada como WHY.

"Já Poliana Okimoto ficará fora de a decisão de os 800m livre por estar com infecção intestinal."

por_(PREPOSIÇÃO) + estar_(verbo_infinitivo)

Regra 3: Se a sentença tiver a expressão "graças a" + ARTIGO, sem ser parte da expressão "dar graças a", então a sentença será anotada com aspecto WHY.

"Graças ao médico, o paciente não morreu."

graças a_(expressão) + o_(ARTIGO)

Figura 19: Regras do *microaspecto* WHY

léxico_de_score = [set(s), gol(s), jogo(s), ...]

Regra 1: Se a sentença tiver NÚMERO + léxico_de_score + "a" + NÚMERO, então a sentença será anotada como SCORE.

"A equipe brasileira venceu a Finlândia por 3 sets a 0 na Liga Mundial de Vôlei-06."

3_(NUM) + sets_(léxico_de_score) + a + 0_(NÚMERO)

Regra 2: Se a sentença tiver NÚMERO + metros + NÚMERO, então a sentença será anotada como SCORE.

"A medalha de prata ficou com a americana com 4m40."

4_(NÚMERO) + m_(metros) + 40_(NÚMERO)

Regra 3: Se a sentença tiver NÚMERO + minuto + NÚMERO + segundo + NÚMERO, então a sentença será anotada como SCORE.

"Eles fizeram história a o cravar o tempo de 7min12s27 e superar os Estados Unidos."

7_(NÚMERO) + min_(minuto) + 12_(NÚMERO) + s_(segundo) + 27_(NÚMERO)

"O Brasil conquistou a medalha de ouro na prova de natação, com o tempo de 3min15s90."

3_(NÚMERO) + min_(minuto) + 15_(NÚMERO) + s_(segundo) + 90_(NÚMERO)

Figura 20: Regras do *microaspecto* SCORE

4.1.3 O sistema REMBRANDT

O sistema REMBRANDT, feito por Cardoso (2008), visa identificar automaticamente as entidades mencionadas presentes nos textos-fonte. Neste trabalho, o sistema REMBRANDT será utilizado na identificação dos *microaspectos* WHEN, WHERE e SITUATION, por serem equivalentes às entidades "tempo", "local" e "acontecimento", respectivamente (ver Tab. 11).

Na Fig. 21, ilustra-se um exemplo de uma sentença anotada pelo sistema REMBRANDT. Diferentemente da anotação da Fig. 11, as EM já foram mapeadas com os *microaspectos* respectivos. Observa-se que a entidade "Jogos Pan-Americanos" foi reconhecida como SITUATION (acontecimento), a entidade "terça-feira" como WHEN (tempo) e as entidades "Finlândia" e "Maracanãzinho", como WHERE (local). Nota-se que a entidade "Finlândia" foi anotada erroneamente como WHERE porque, no contexto, faz referência a uma equipe de vôlei e não a um local.

No contexto dos <aspect REMBRANDT="SITUATION">Jogos Pan-Americanos </aspect>, a equipe brasileira de vôlei venceu nesta <aspect REMBRANDT="WHEN">terça-feira</aspect> a <aspect REMBRANDT="WHERE">Finlândia</aspect> por 3 sets a 0 no, <aspect REMBRANDT="WHERE">Maracanãzinho</aspect>.

Figura 21: Sentença anotada com *microaspectos* pelo sistema REMBRANDT

Da mesma maneira que os sistemas APS e APS + Regras, a anotação será feita em nível sentencial (ver Fig. 22).

[No contexto dos Jogos Pan-Americanos, a equipe brasileira de vôlei venceu nesta terça-feira a Finlândia por 3 sets a 0, no Maracanãzinho.] **SITUATION/WHEN/WHERE**

Figura 22: Anotação de *microaspectos* em nível sentencial usando REMBRANDT

4.1.4 Aprendizado de Máquina

Na atualidade, destaca-se a capacidade dos computadores de aprender tarefas automaticamente com base em alguma experiência. Essa experiência se constrói por meio de um conjunto de exemplos denominados instâncias. Cada instância contém certos atributos que, teoricamente, representam conhecimento útil à tarefa a ser automatizada. Em um sistema de Aprendizado de Máquina (AM), a experiência recebe o nome de "conjunto de treinamento". Segundo Mitchell (1997), a predição desejada em uma instância recebe o nome de rótulo, tornando-se um conjunto finito de valores, denominados classes. Em outras palavras, o AM tenta generalizar a predição de uma classe a partir de um conjunto finito de treinamento para dados de teste nunca antes vistos.

Neste trabalho, a tarefa a ser aprendida é a "identificação de *microaspectos*". Devido à disponibilidade de um córpus anotado manualmente (CSTNews), a nossa tarefa segue na linha do paradigma de AM supervisionado, em que o conjunto de treinamento está formado por pares instância-classe denominados dados rotulados. As instâncias-classes são as sentenças do córpus anotadas com aspectos informativos.

A identificação de *microaspectos* é um problema de classificação multirrótulo. Neste trabalho, aplica-se o método de transformação de problemas (Tsoumakas e Katakis, 2007), que visa transformar o problema de classificação multirrótulo em um conjunto de problemas de classificação binária. Portanto, criaram-se vários classificadores binários, sendo escolhidos os 8 melhores, para cada um dos *microaspectos* WHO_AGENT, WHO_AFFECTED, WHERE, WHEN, WHY, HOW, SITUATION e SCORE, respectivamente. O *microaspecto* GOAL não foi considerado por ter poucas instâncias anotadas.

No total, definiram-se 6 tipos de atributos léxico-semânticos (ver Tab. 12). Para extrair tais atributos, utilizou-se o formato *flat* (simples) do *parser* PALAVRAS (ver Fig. 9). Cada atributo é representado por unigramas "(1, 1)", bigramas "(2, 2)" e bigramas + trigramas "(2, 3)". Assim, para cada um dos 8 *microaspectos*, cria-se um classificador resultado da representação (unigramas, bigrama, bigrama+trigrama) de cada um dos 6 tipos de atributos. Por exemplo, o classificador denominado "*(2, 3) POS*" foi criado com base em todos os bigramas e trigramas "(2, 3)" das classes gramaticais (POS) de todas as palavras do córpus. Já o classificador "*(2, 2) lemmas+POS*" foi criado com base em

todos os bigramas "(2,2)" da união do lema e o POS (p.ex: "o+DET", "menino+N", "nadar+V") de todas as palavras do córpus. Observa-se que foi considerada a coocorrência de unigramas e bigramas de palavras denominado *bag of words*, também proposto por Makino et al. (2011). Ao final, criaram-se 144 classificadores binários (ver Tab. 71, Apêndice C). Salienta-se que esses atributos também serão utilizados na identificação de *macroaspectos*.

Tipo de atributo	**Notação**
Bag of words	bag_of_words
Lematização	lemmas
POS (*part-of-speech*)	POS
Etiquetas semânticas	semantic
Lematização + POS	lemmas+POS
POS + etiquetas semânticas	POS+semantic

Tabela 12: Atributos definidos

A técnica de aprendizado supervisionado utilizada foi SVM (*Support Vector Machine*) (Vapnik, 1995). A técnica se baseia no princípio de minimização do risco estrutural, trabalhando sobre o conceito de margem. O SVM realiza a classificação de dados por meio da construção de vários hiperplanos. O termo margem refere-se à distância mínima a partir do hiperplano de separação até as instâncias de dados mais próximas. A técnica visa criar a maior distância possível entre os hiperplanos de separação e as instâncias próximas a eles. O fato de considerar apenas instâncias próximas às margens é uma característica particular da técnica, daí o nome "vetores de suporte". Escolheu-se o SVM, entre outras técnicas da literatura (SMO, Naïve Bayes, J48), por ser atualmente a técnica mais utilizada na literatura para classificação com textos. Além disso, é a melhor técnica em tratamento de vetores especiais de grandes dimensões.

Portanto, propõe-se o uso de AM para criar classificadores que possam identificar automaticamente *microaspectos*. Objetiva-se obter o melhor classificador para cada *microaspecto* avaliando-se todos os possíveis classificadores gerados a partir dos atributos léxico-semânticos. As instâncias de treino e teste são as sentenças dos sumários anotados do córpus CSTNews.

4.2 Identificação de Macroaspectos

De mesma forma que os *microaspectos*, o processo de identificação de *macroaspectos* foi dividido em 3 fases (ver Fig. 23). A seguir, explicam-se as fases do processo de identificação:

1. Compilar as sentenças dos 48 sumários anotados do córpus CSTNews das categorias "Cotidiano", "Esporte", "Mundo" e "Política". Não foram consideradas as categorias "Dinheiro" e "Ciência", por terem poucos sumários anotados.

2. Anotar as sentenças com *macroaspectos* usando duas abordagens:

   (a) **Aprendizado de Máquina:** uso de técnicas de AM para criar classificadores de *macroaspectos*. Por um lado, serão criados classificadores com base nos atributos definidos por Teufel (1999) e utilizados em outros trabalhos da literatura (Teufel e Moens, 2002; Feltrim et al., 2006; Genoves Jr. et al., 2007; Dayrell et al., 2012). Por outro lado, serão criados classificadores com base em atributos léxico-semânticos: *bag of words*, *lemas*, *part-of-speech* (POS), etiquetas semânticas e a combinação deles (previamente utilizados na identificação de *microaspectos*). Cabe ressaltar que a maioria dos atributos utilizados nos classificadores são fornecidos pelo *parser* PALAVRAS (Bick, 2000). Esta abordagem atende os *macroaspectos* WHAT, CONSEQUENCE, COMMENT, DECLARATION e HISTORY. O restante dos *macroaspectos* não foi considerado por haver poucas sentenças anotadas.

   (b) **Regras manuais:** devido ao baixo resultado obtido pela abordagem usando AM, optou-se pela criação de regras com base na identificação de padrões linguísticos presentes nas sentenças dos sumários anotados no córpus CSTNews. Esta abordagem atende os *macroaspectos* COMPARISON, DECLARATION, GOAL, HISTORY e PREDICTION. Não foi possível identificar padrões linguísticos para o restante dos *macroaspectos*.

3. Obter um conjunto de sentenças anotadas automaticamente com *macroaspectos*.

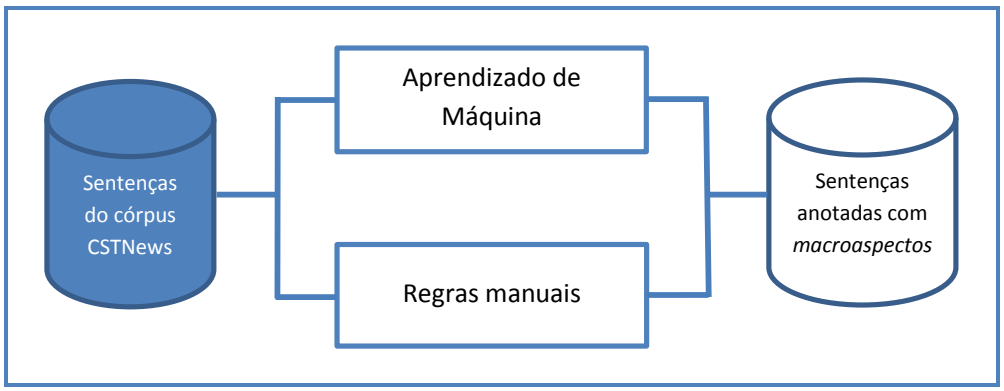

Figura 23: Metodologia do processo de identificação de *macroaspectos*

4.2.1 Aprendizado de Máquina

Da mesma maneira que para os *microaspectos*, criaram-se vários classificadores binários, sendo escolhidos os 5 melhores, para cada um dos *macroaspectos* WHAT, CONSEQUENCE, COMMENT, DECLARATION e HISTORY, respectivamente. Já os *macroaspectos* COMPARISON, PREDICTION, COUNTERMEASURES, GOAL, SITUATION e HOW não foram considerados por terem poucas instâncias anotadas.

Por um lado, criaram-se classificadores binários com base nos atributos definidos por (Teufel, 1999) e utilizados em outros autores da literatura, como Teufel e Moens (2002); Feltrim et al. (2006); Genoves Jr. et al. (2007); Dayrell et al. (2012) (ver Tab. 1). Como já foi dito, os atributos foram definidos originalmente para o gênero científico, portanto, os atributos foram adaptados para o gênero jornalístico.

Os atributos *TF-IDF*, *título*, *tamanho* e *posição* foram conservados na forma original. Já para os atributos *tempo*, *voz* e *modal*, não só foi considerado o primeiro verbo, mas sim todos os verbos da sentença. Por exemplo, costuma-se classificar como PREDICTION as sentenças que possuem algum verbo no tempo futuro. No caso da sentença "Ele melhorou e está estável, mas continuará internado", o verbo "continuará" (tempo futuro) é o terceiro verbo da sentença. Isso acontece porque o estilo de escrita jornalístico não segue o mesmo estilo dos textos científicos, em que o primeiro verbo pode estar relacionado ao papel retórico da sentença. Finalmente, para o atributo *expressão padrão*, só foram identificadas expressões para DECLARATION (por exemplo, "de acordo com", "segundo" e os verbos ilocutórios) e COMPARISON ("em relação a", "em

comparação a"). Cabe ressaltar que não foi possível identificar uma maior quantidade de expressões devido a pouca quantidade de sentenças anotadas no córpus CSTNews.

Por outro lado, criaram-se vários classificadores binários com base em 6 tipos de atributos léxico-semânticos, utilizados também para identificar *microaspectos* (ver Tab. 12). Lembra-se que para extrair tais atributos, utilizou-se o formato *flat* do *parser* PALAVRAS. No total, foram criados 90 classificadores binários, resultado da combinação dos 6 atributos representados em unigramas, bigramas e bigramas+trigramas, para os 5 *macroaspectos*. (ver Apêndice D).

A técnica de aprendizado supervisionado utilizada nas duas abordagens foi o SVM. Propõe-se, então, o uso de AM para criar classificadores que possam identificar automaticamente *macroaspectos*. Objetiva-se obter o melhor classificador para cada *macroaspecto*, avaliando-se todos os possíveis classificadores gerados a partir dos atributos definidos por Teufel (1999) e os atributos léxico-semânticos. As instâncias de treino e teste são as sentenças dos sumários anotados do córpus CSTNews.

4.2.2 Regras Manuais

A abordagem usando AM obteve resultados muito baixos, sendo que vários dos *macroaspectos* não puderam ser identificados corretamente em nenhuma sentença. Portanto, criaram-se regras manuais analisando todas as sentenças anotadas do córpus CSTNews com a finalidade de achar padrões linguísticos que pudessem representar alguns dos *macroaspectos*.

Foram criadas regras para os aspectos COMPARISON, DECLARATION, PREDICTION, HISTORY e GOAL. No entanto, não foram criadas regras para os aspectos WHAT, COMMENT, CONSEQUENCE, COUNTERMEASURES, SITUATION e HOW, por não terem sido achados padrões para criação de regras. Utilizou-se o formato *flat* do PALAVRAS que fornece informações morfossintáticas, lematização e etiquetas semânticas do tipo pessoa/organização.

A maioria das regras visam identificar expressões padrão. Assim, por exemplo, a expressão "em relação a" denota COMPARISON (ver Fig. 24). Da mesma forma, as expressões "segundo" e "de acordo com" correspondem a DECLARATION (ver Fig. 25); "previsão" corresponde a PREDICTION (ver Fig. 26); "desde" e "da história" correspon-

dem a HISTORY (ver Fig. 27); e "objetivo" corresponde a GOAL (ver Fig.28).

Outras regras se baseiam no tipo de verbo. Por exemplo, qualquer tipo de verbo ilocutório (p.ex: "declarar", "afirmar", "dizer", "informar", "anunciar", "expressar", etc.) denota DECLARATION. Já outras regras se baseiam no tempo verbal. Por exemplo, os verbos no futuro costumam expressar uma previsão (PREDICTION).

Cabe ressaltar que a maioria das regras (COMPARISON, GOAL e PREDICTION) foram criadas sobre pouca quantidade de sentenças anotadas, sendo relativamente simples de se identificar padrões linguísticos. No entanto, existe a possibilidade de acontecer *overfitting*[3] nas regras, por estas serem criadas e testadas sobre um conjunto mínimo de dados. Um modelo com *overfitting* apresenta uma alta precisão, porém tal modelo não é uma boa representação da realidade.

Regra 1: Se a sentença contiver a PREPOSIÇÃO "em", seguida de ("relação"|"comparação"), seguida do ARTIGO "a", então a sentença será anotada como COMPARISON.

"Foram autuados 208.471 contribuintes, um crescimento de 104,47% em relação a o ano passado."

em_PREPOSIÇÃO + relação + a_(ARTIGO)

Regra 2: Se a sentença tiver o VERBO "comparar", então a sentença será anotada como COMPARISON.

"As intenções de voto para Lula caíram quando se compara com os candidatos Geraldo e Heloisa."

compara_(VERBO) = comparar_(VERBO)

Figura 24: Regras do *macroaspecto* COMPARISON

[3]Termo utilizado em AM ou estatística para dizer que o modelo estatístico se ajustou demasiadamente ao conjunto de dados, não sendo capaz de generalizar adequadamente.

verbos_ilocutórios = [dizer, afirmar, anunciar, informar, destacar, expressar, referir, opinar, ...]
PESSOA/ORGANIZAÇÃO = [H, Hprof, hum, admin, org, ints, media, party, suborg]
$\subset$ = "está contido em"

Regra 1: Se a sentença tiver um VERBO contido nos "verbos_ilocutórios", então a sentença será anotada como DECLARATION.

"Marcelinho disse que logo a torcida vai se acostumar e apoiar a mudança de levantador."

disse_(VERBO) = dizer_(VERBO) $\subset$ verbos_ilocutórios

"Neste mesmo dia, o exército israelense afirmou ter matado 30 milicianos do Hezbollah."

afirmou_(VERBO) = afirmar_(VERBO) $\subset$ verbos_ilocutórios

Regra 2: Se a sentença tiver a PREPOSIÇÃO "segundo", seguida por um ARTIGO, então a sentença será anotada como DECLARATION.

"Segundo o secretário-adjunto da Receita, o Leão está mais atento, não guloso."

Segundo_(PREPOSIÇÃO) + o_(ARTIGO)

Regra 3: Se s sentença tiver a PREPOSIÇÃO "segundo", seguida por um substantivo associado a uma etiqueta semântica do tipo PESSOA/ORGANZAÇÃO, então a sentença será anotada como DECLARATION.

"Segundo Lula, o mundo precisa de uma nova matriz energética, e o etanol pode ..."

segundo_(PREPOSIÇÃO) + (Lula_(H) $\subset$ PESSOA/ORGANIZAÇÃO)

Regra 4: Se a sentença tiver a PREPOSIÇÃO "de", seguida do VERBO "acordo", seguida da PREPOSIÇÃO "com", então a sentença será anotada como DECLARATION.

"De acordo com a Infraero, será possível realizar a obra em três etapas."

de_PREPOSIÇÃO + acordo_(VERB) + com_(PREPOSIÇÃO)

Figura 25: Regras do *macroaspecto* DECLARATION

Regra 1: Se a sentença tiver um VERBO no futuro, então a sentença será anotada como PREDICTION.

"A seleção brasileira ainda enfrentará portugueses e finlandeses na fase de classificação."

enfrentará_(VERBO_no_futuro)

Regra 2: Se a sentença tiver o token "previsão", então a sentença será anotada como PREDICTION.

"A previsão de chuva na área aumenta os temores de mais devastação."

Figura 26: Regras do *macroaspecto* PREDICTION

Regra 1: Se a sentença tiver o ADVÉRBIO "já", seguido de um VERBO no tempo pretérito perfeito (PS), pretérito mais que perfeito (PS/MQP), pretérito imperfeito (IMPF) ou condicional (COND); então a sentença será anotada como HISTORY.

"As ações são atribuídas à facção criminosa PCC, que já comandou outros ataques em duas ocasiões."

já_(ADVÉRBIO) + comandou_(PS)

"ACM já tinha sofrido infarto em 1989 e já tinha recebido três pontes de safena."

já_(ADVÉRBIO) + tinha_(IMPF)

Regra 2: Se a sentença tiver o token "desde", então a sentença será anotada como HISTORY.

"O grupo criminoso desviou desde 2004 cerca de R$ 70 milhões dos cofres públicos."

"Ele está envolvido com o tráfico desde 1986 e criou sua própria rede distribuidora de drogas ..."

Regra 3: Se a sentença tiver a PREPOSIÇÃO+ARTIGO "da", seguido do token "história", então a sentença será anotada como HISTORY.

"Esse foi o pior ataque a tiros contra um campus universitário da história dos EEUU."

da_(PREPOSIÇÃO+ARTIGO) + "história"

Figura 27: Regras do *macroaspecto* HISTORY

Regra 1: Se a sentença tiver o token lematizado "objetivo", então a sentença será anotada como GOAL.

"O governo israelense objetiva uma zona de segurança cedida a uma força multinacional apoiando..."

objetiva = objetivo_(lema)

"O objetivo das buscas é garantir a apreensão dos registros de ocorrências que contêm informações ..."

objetivo = objetivo_(lema)

Figura 28: Regras do *macroaspecto* GOAL

4.3 Métodos de Seleção e Ordenação de Conteúdo

O objetivo desta etapa é gerar um sumário selecionando-se as informações mais importantes que sejam informativas (ou de interesse) para o usuário. A metodologia adotada é ilustrada na Fig. 29. A seguir, explicam-se as fases do processo de seleção de conteúdo.

1. Dado um conjunto de sentenças anotadas com aspectos (*microaspectos* e *macro-aspectos*), realiza-se um ranqueamento de sentenças por meio do sistema RSumm (Ribaldo et al., 2012). Assim, obtém-se um conjunto de sentenças ordenadas por relevância e sem presentar redundância.
2. Selecionam-se e ordenam-se as sentenças com base nos padrões identificados (*templates*) na anotação de aspectos no córpus CSTNews (Rassi et al., 2013; Felippo et al., 2014), para cada categoria em particular (ver Seção 3.1.1). Nesta fase, criaram-se dois métodos de seleção de conteúdo.
3. Obtém-se o sumário final com base em uma taxa de compressão dada pelo usuário. Nesse caso, utiliza-se 70% de compressão, que é o padrão no córpus CSTNews.

Salienta-se que o sumário será formado por um conjunto de sentenças extraídas dos textos-fonte sem uma divisão entre parágrafos. Em outras palavras, o sumário será formado por um **parágrafo** só.

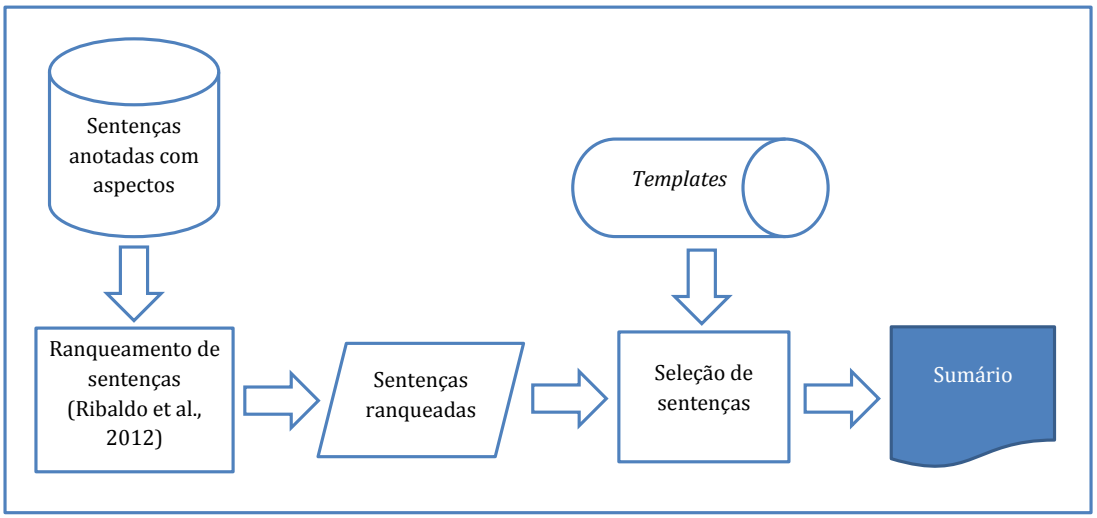

Figura 29: Metodologia do processo de seleção de conteúdo

O objetivo do sumarizador RSumm é ranquear as sentenças anotadas conforme a sua relevância nos textos-fonte. Nesta fase, **descartou-se** o algoritmo de ordenação de sentenças do Lima e Pardo (2011, 2012) (utilizado pelo RSumm), já que a ordenação de sentenças será feita pelos métodos propostos neste trabalho. Além disso, o RSumm remove todas as sentenças que apresentem redundância, fenômeno comum na sumarização multidocumento. Assim, obtém-se um conjunto menor de sentenças ordenadas por relevância em relação ao conjunto total de sentenças dos textos-fonte.

Após o ranqueamento de sentenças, desenvolveram-se dois métodos de seleção de conteúdo com base nos *templates* identificados na anotação de aspectos. O primeiro método visa dar uma pontuação para cada sentença com base no balanceamento da relevância (dada pelo RSumm) e da informatividade (dada pelos aspectos). Ressalta-se que os métodos propostos são novidade na sumarização multidocumento com base em aspectos informativos.

A relevância é dada pela posição da sentença no conjunto de sentenças. Dado um conjunto de sentenças C ranqueadas pelo RSumm, a relevância da sentença S é determinada pela fórmula a seguir:

$$RScore(S) = \frac{|C| - i}{|C|} \tag{10}$$

em que $|C|$ representa o total de sentenças no conjunto C e i representa a posição da sentença no conjunto. Por exemplo, a sentença que aparece na primeira posição ($i = 0$) em um conjunto de 15 sentenças previamente ranqueadas terá um *RScore(S)=1.000*. Já a sentença que aparece na quarta posição ($i = 3$) terá um *RScore(S)=0.800*.

A informatividade é dada pela cobertura da maioria dos aspectos definidos para uma categoria específica. Para cada categoria, com base nos *templates* identificados nas Tabs. 6, 7, 8 e 9, definiram-se os denominados "conjuntos de cobertura". Tais conjuntos estão formados pelos aspectos que acontecem com maior frequência no **1ro parágrafo** dos sumários anotados (tanto para a **maioria** quanto para **todos** os sumários). Na Tab. 13, mostra-se o conjunto de aspectos definidos para cada categoria.

Categoria	Conjunto de cobertura
Cotidiano	WHAT, WHERE, WHEN, WHO_AGENT, DECLARATION
Esportes	WHO_AGENT, WHAT, SCORE, CONSEQUENCE, SITUATION, COMMENT, WHEN, WHERE
Mundo	WHAT, WHERE, WHO_AFFECTED, WHEN, WHO_AGENT, CONSEQUENCE
Política	WHO_AGENT, WHAT, WHO_AFFECTED, WHEN, DECLARATION

Tabela 13: Cobertura de aspectos por categoria

Com base nesses conjuntos, pontua-se uma sentença conforme a cobertura de aspectos. Dada uma sentença S da categoria c anotada com aspectos, a cobertura é dada pela fórmula a seguir:

$$AScore(S) = \frac{nMatches(A(S), A(c))}{|A(c)|} \tag{11}$$

em que *nMatches* representa o total de aspectos da sentença $A(S)$ que pertencem ao conjunto de aspectos definidos $A(c)$ para a categoria em questão. Já $|A(c)|$ representa o total de aspectos definidos no conjunto de cobertura. Por exemplo, considere uma sentença da categoria "Cotidiano" com os aspectos WHAT, WHERE, WHEN e PREDICTION. Olhando-se para a Tab. 13, pode-se deduzir que a sentença contém somente 3 dos 5 aspectos definidos para essa categoria (WHAT, WHERE e WHEN), sendo *nMatches=3*. Assim, o valor do *AScore(S)=3/5=0.600*. Finalmente, a sentença S será pontuada de acordo com a seguinte fórmula:

$$SentScore(S) = (RScore(S) * (1 - \alpha)) + (AScore(S) * \alpha) \tag{12}$$

em que α represente o coeficiente de balanceamento entre as pontuações de relevância (*RScore*) e informatividade (*AScore*). Os valores de α variam entre $[0, 1]$. Assim, se $\alpha = 1$, outorga-se mais importância à informatividade (ou cobertura de aspectos), e, se $\alpha = 0$, outorga-se mais importância à relevância.

Em seguida, ordenam-se as sentenças decrescentemente conforme as novas pontuações dadas pelo *AScore(S)* e selecionam-se tantas sentenças conforme a taxa de compressão dada pelo usuário, dando como saída o sumário final. No total, criaram-se 10 sumarizadores a partir dos valores definidos entre 0 e 1 para "α". Cada sumarizador

representa uma configuração de "α" denominado *ASumm_n*, em que "n" é um identificador numérico. Por exemplo, a configuração *ASumm_8* é dada por $\alpha = 0.8$.

Já o segundo método, denominado *ASumm_OP*, está baseado na **ordenação parcial** identificada nos *templates* definidos para cada categoria do córpus. Dado um conjunto de sentenças ranqueadas pelo RSumm (relevância), o método *ASumm_OP* visa selecionar sentenças com base na ordem em que ocorrem os aspectos nos *templates*. Na Tab. 14, mostra-se o "padrão de ordem" que devem seguir as sentenças, baseado na ordenação parcial dos aspectos da categoria "Esportes" (ver Tab. 7). Assim, seleciona-se a sentença que contenha pelo menos um aspecto do "padrão de ordem" para a categoria em questão. Essa sentença será removida do conjunto inicial e será colocada na primeira posição do sumário. Em seguida, seleciona-se outra sentença que contenha o aspecto subsequente no "padrão de ordem"; remove-se a sentença do conjunto original e coloca-se na segunda posição do sumário, e assim por diante. Se a sentença selecionada já contiver o aspecto a seguir no "padrão de ordem", passa-se a selecionar outra sentença do conjunto inicial. Caso o conjunto de sentenças não contiver o aspecto no "padrão de ordem", passa-se ao próximo aspecto. Da mesma forma que o método *ASumm_n*, limita-se o sumário de acordo com a taxa de compressão dada pelo usuário.

Ordenação parcial	• WHO_AGENT $<$ WHAT • WHO_AGENT, WHAT $<$ SCORE • WHO_AGENT, WHAT $<$ CONSEQUENCE • WHO_AGENT, WHAT $<$ SITUATION • WHO_AGENT, WHAT $<$ WHERE • WHO_AGENT, WHAT, SCORE $<$ CONSEQUENCE
Padrão de ordem	WHO_AGENT $<$ WHAT $<$ (SCORE|SITUATION|WHERE) $<$ CONSEQUENCE

Tabela 14: Padrão de ordem da categoria "Esportes"

Para exemplificar o método, na Tab. 15, apresenta-se um conjunto de quatro sentenças anotadas previamente ranqueadas pelo RSumm. O método *ASumm_OP* funciona da seguinte forma:

1. Seleciona-se a sentença S2 por ser a primeira do conjunto que contém o aspecto WHO_AGENT. O conjunto de sentenças é agora formado por S1, S3 e S4. O padrão de ordem é agora "WHAT < (SCORE|SITUATION| WHERE) < CONSEQUENCE". O sumário, até o momento, está formado por S2.

2. Seleciona-se a sentença S1 por ser a primeira do conjunto que contém o aspecto WHAT. O conjunto de sentenças é agora formado por S3 e S4. O padrão de ordem é agora "(SCORE|SITUATION| WHERE) < CONSEQUENCE". O sumário, até o momento, está formado por S2 e S1.

3. Seleciona-se a sentença S4 por ser a primeira do conjunto que contém o aspecto SITUATION. O conjunto de sentenças é agora formado por S3. O padrão de ordem é agora "(SCORE|WHERE) < CONSEQUENCE". O sumário, até o momento, está formado por S2, S1 e S4.

4. Por último, seleciona-se a sentença S3 por ser a primeira e a única do conjunto que contém o aspecto SCORE. O sumário final está formado por S2, S1, S4 e S3 (respeitando essa ordem).

Sentença	**Aspectos**
S1	WHAT, WHEN, WHERE
S2	WHO_AGENT, CONSEQUENCE
S3	WHO_AGENT, WHO_AFFECTED, SCORE
S4	WHEN, SITUATION

Tabela 15: Exemplo de sentenças ranqueadas/anotadas da categoria "Esportes"

4.4 Arquitetura SA Multidocumento

Utilizando os aspectos informativos como guia para sumarizar, implementou-se uma arquitetura para SA multidocumento. Tal arquitetura descreve as três etapas da arquitetura geral de um sistema SA (*análise*, *transformação* e *síntese*), começando por um conjunto de textos-fonte/documentos que tratam um mesmo tópico como entrada e produzindo um sumário final como saída. A arquitetura é mostrada na Fig. 30.

A etapa de **análise** inicia com a análise sintática do parser PALAVRAS para todas as sentenças de entrada. Em seguida, as sentenças são anotadas automaticamente pelos classificadores de *microaspectos* e *macroaspectos*, respectivamente.

A etapa de **transformação** inicia com o conjunto de sentenças anotadas com aspectos. Em seguida, ranqueiam-se as sentenças e remove-se a redundância com o RSumm. Posteriormente, selecionam-se e ordenam-se as sentenças conforme os métodos propostos: *ASumm_n* e *Asumm_OP*.

A etapa de **síntese** inicia com um conjunto de sentenças ordenadas pelos métodos *ASumm_n* e *Asumm_OP*. Em seguida, realiza-se uma justaposição das sentenças do conjunto para mostrar o sumário final ao usuário em forma de um parágrafo.

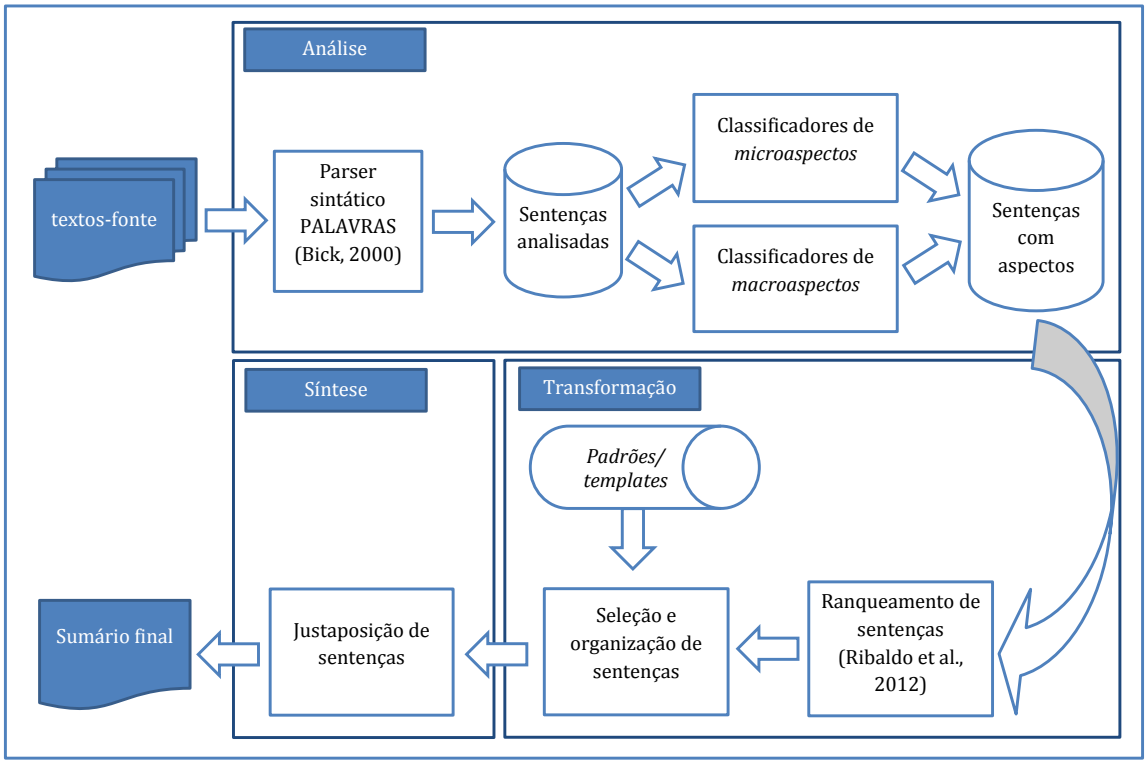

Figura 30: Arquitetura do sistema SA multidocumento

A seguir, mostra-se um exemplo completo do funcionamento da arquitetura de SA multidocumento. No início, têm-se dois documentos de entrada que falam sobre a vitória da equipe masculina de vôlei (ver Figs. 31 e 32). Cada sentença tem um identificador entre colchetes "[]". Assim, por exemplo, o identificador [S1-D2] representa a sentença 1 do documento 2. Cabe ressaltar que este exemplo foif feito manualmente.

A seleção brasileira masculina de vôlei conseguiu, nesta sexta-feira, a sétima vitória consecutiva na Liga Mundial ao derrotar a Finlândia por 3 sets a 0 - parciais de 25/17, 25/22 e 25/21 -, em jogo realizado na cidade de Tampere, na Finlândia. **[S1-D1]** Invicto na competição, o Brasil está tranquilo na liderança do Grupo B. **[S2-D1]** Os Finlandeses estão na terceira colocação, com três vitórias e quatro derrotas. **[S3-D1]** Portugal e Argentina - que duelam duas vezes neste final de semana, em Portugal - completam a chave. **[S4-D1]** Brasil e Finlândia se enfrentarão novamente neste sábado, às 12h30 (horário de Brasília), com transmissão ao vivo do canal de TV a cabo SporTV. **[S5-D1]** Nas duas ultimas rodadas da fase de classificação da Liga Mundial, a seleção brasileira receberá a portugueses e Finlandeses. **[S6-D1]** A fase final da competição deste ano acontecerá na Rússia. **[S7-D1]**

Figura 31: Documento 1 da categoria "Esportes"

A seleção brasileira masculina de vôlei, que é treinada por Bernardinho, venceu a Finlândia por 3 sets a 0, parciais de 25/17, 25/22 e 25/21, nesta sexta-feira, em Tampere (FIN), e manteve sua invencibilidade na Liga Mundial-06. **[S1-D2]** Este foi o sétimo triunfo consecutivo dos brasileiros na competição -antes, o país conquistou quatro vitórias contra a seleção argentina e duas diante de Portugal. **[S2-D2]** Os dois times voltam a se enfrentar às 12h30 deste sábado, no mesmo ginásio, que normalmente é utilizado para competições de hóquei no gelo. **[S3-D2]** A equipe brasileira masculina já conquistou cinco vezes a Liga Mundial -1993, 2001, 2003, 2004 e 2005. **[S4-D2]** Com oito títulos, a Itália é a maior vencedora da competição. **[S5-D2]**

Figura 32: Documento D2 da categoria "Esportes"

Em primeiro lugar, anotam-se os aspectos para cada sentença por meio dos classificadores de *microaspectos* e *macroaspectos* (ver Tab. 16). Por exemplo, a sentença: "*A equipe brasileira masculina já conquistou cinco vezes a Liga Mundial -1993, 2001, 2003, 2004 e 2005*", com identificador [S4-D2], contêm os *microaspectos* WHO_AGENT ("A equipe brasileira") e WHEN ("1993, 2001, 2003, 2004 e 2005"), e os *macroaspectos* HISTORY ("já conquistou") e SITUATION ("Liga Mundial").

Sentença	Aspectos
[S1-D1]	WHO_AGENT, WHEN, WHAT, SITUATION, WHO_AFFECTED, SCORE, WHERE, GOAL
[S2-D1]	WHO_AGENT
[S3-D1]	WHO_AGENT, GOAL
[S4-D1]	WHO_AGENT, WHEN
[S5-D1]	WHO_AGENT, WHEN, HOW
[S6-D1]	SITUATION, WHO_AGENT, PREDICTION, WHO_AFFECTED
[S7-D1]	WHEN, PREDICTION, WHERE
[S1-D2]	WHO_AGENT, WHO_AFFECTED, SCORE, WHEN, WHERE, SITUATION, WHAT
[S2-D2]	GOAL, HISTORY
[S3-D2]	WHO_AGENT, WHEN, WHERE
[S4-D2]	WHO_AGENT, HISTORY, SITUATION, WHEN
[S5-D2]	SCORE, WHO_AGENT, COMPARISON

Tabela 16: Sentenças anotadas com aspectos da categoria “Esportes”

Posteriormente, realiza-se o ranqueamento de sentenças por relevância. Na Tab. 17, mostram-se as sentenças anotadas (com aspectos) em ordem decrescente, fornecidas pelo RSumm. Observa-se que o sistema descartou três sentenças por apresentarem redundância: [S1-D2] é similiar a [S1-D1], [S4-D1] é similar a [S2-D2], e [S5-D1] é similiar a [S3-D2].

Sentença	Aspectos
[S1-D1]	WHO_AGENT, WHEN, WHAT, SITUATION, WHO_AFFECTED, SCORE, WHERE, GOAL
[S6-D1]	SITUATION, WHO_AGENT, PREDICTION, WHO_AFFECTED
[S2-D2]	GOAL, HISTORY
[S2-D1]	WHO_AGENT
[S3-D2]	WHO_AGENT, WHEN, WHERE
[S5-D2]	SCORE, WHO_AGENT, COMPARISON
[S4-D2]	WHO_AGENT, HISTORY, SITUATION, WHEN
[S3-D1]	WHO_AGENT, GOAL
[S7-D1]	WHEN, PREDICTION, WHERE

Tabela 17: Sentenças ranqueadas pelo RSumm da categoria “Esportes”

Em seguida, utiliza-se o método *ASumm_n* para ordenar as sentenças por informatividade. Para isso, definiu-se um α igual 0.75, de maneira que se priorize a cobertura de aspectos. Na Tab. 18, mostram-se as sentenças ordenadas de forma decrescente pela pontuação *AScore(S)*. Cabe ressaltar que a pontuação *AScore(S)* é calculada segundo o "conjunto de cobertura" da categoria em questão (ver Tab. 13).

Sentença	**AScore**	**Aspectos**
[S1-D1]	0.813	WHO_AGENT, WHEN, WHAT, SITUATION, WHO_AFFECTED, SCORE, WHERE, GOAL
[S3-D2]	0.420	WHO_AGENT, WHEN, WHERE
[S6-D1]	0.410	SITUATION, WHO_AGENT, PREDICTION, WHO_AFFECTED
[S4-D2]	0.365	WHO_AGENT, HISTORY, SITUATION, WHEN
[S5-D2]	0.299	SCORE, WHO_AGENT, COMPARISON
[S2-D1]	0.260	WHO_AGENT
[S7-D1]	0.215	WHEN, PREDICTION, WHERE
[S2-D2]	0.194	GOAL, HISTORY
[S3-D1]	0.149	WHO_AGENT, GOAL

Tabela 18: Sentenças ordenadas por peso *AScore* da categoria "Esportes"

Finalmente, o sumário será formado pela seleção das três primeiras sentenças da Tab. 18, conforme a taxa de compressão de 70% (aproximadamente 30% das palavras do maior documento). Na Fig. 33, apresenta-se o sumário final. A simples vista pode se observar que o sumário gerado é informativo, já que abrange a maioria dos aspectos definidos para a categoria em questão.

A seleção brasileira masculina de vôlei conseguiu, nesta sexta-feira, a sétima vitória consecutiva na Liga Mundial ao derrotar a Finlândia por 3 sets a 0 - parciais de 25/17, 25/22 e 25/21 -, em jogo realizado na cidade de Tampere, na Finlândia. **[S1-D1]** Os dois times voltam a se enfrentar às 12h30 deste sábado, no mesmo ginásio, que normalmente é utilizado para competições de hóquei no gelo. **[S3-D2]** Nas duas ultimas rodadas da fase de classificação da Liga Mundial, a seleção brasileira receberá a portugueses e Finlandeses. **[S6-D1]**

Figura 33: Sumário final da categoria "Esportes"

No capítulo seguinte, narram-se os resultados do processo de identificação de aspectos e avaliação de sumários automáticos.

4.5 Considerações Finais

Neste capítulo, descreveu-se o processo de sumarização multidocumento com base em aspectos informativos para o Português. Usando como referência a arquitetura geral de um sistema de SA, este trabalho está focado nas duas primeiras etapas: *análise* e *transformação*. A etapa de *análise* visa anotar aspectos nas sentenças dos textos-fonte/documentos de entrada por meio de um classificador de *microaspectos* e *macroaspectos*. Já a etapa de *transformação* visa selecionar e organizar as sentenças com base nos *templates* identificados na anotação de aspectos sobre os sumários do córpus CSTNews (Rassi et al., 2013; Felippo et al., 2014), dando como saída o sumário final.

Para o classificador de *microaspectos*, criaram-se sistemas com auxílio de algumas ferramentas da literatura como o Anotador de Papéis Semânticos e o Reconhecedor de Entidades Mencionadas, todas elas feitas para a língua Portuguesa. Com a finalidade de melhorar o desempenho do sistema APS, construíram-se regras manuais com base nos padrões identificados nas sentenças. Por outro lado, propôs-se uma abordagem utilizando AM em que se definiram atributos léxico-semânticos.

Para o classificador de *macroaspectos*, propôs-se uma abordagem usando AM em que se definiram atributos léxico-semânticos (também utilizados na identificação de *microaspectos*) e atributos definidos por Teufel (1999) (originalmente utilizados para identificar papéis retóricos). Como a abordagem usando AM teve resultados insatisfatórios, criaram-se regras manuais para alguns *macroaspectos* com o objetivo de melhorar esses resultados.

Para selecionar e organizar o conteúdo que formará o sumário final, criaram-se dois métodos: *ASumm_n* e *ASumm_OP*. O primeiro visa gerar um sumário por meio do balanceamento entre a relevância dada pelo RSumm e a informatividade dada pela cobertura de aspectos definidos para cada categoria do córpus. O segundo método visa gerar um sumário por meio de um "padrão de ordem" definido com base na ordenação parcial dos aspectos para cada categoria do córpus. Os sumários finais são delimitados por uma taxa de compressão dada pelo usuário. Cabe ressaltar que os métodos desenvolvidos recebem como entrada um conjunto de sentenças previamente ranqueadas por relevância pelo sistema RSumm. Tal sistema remove as sentenças que apresentem

redundância, fenômeno muito comum na SA multidocumento.

O processo de identificação de aspectos é avaliado sobre os sumários do córpus anotado com aspectos CSTNews. Já os sumários gerados são avaliados sobre os sumários humanos do mesmo córpus. No capítulo seguinte, apresentam-se os resultados da avaliação do processo de sumarização desenvolvido neste trabalho de pesquisa.

CAPÍTULO

5

Avaliação: Experimentos e Resultados

Neste capítulo, apresentam-se os experimentos realizados ao longo deste trabalho e os resultados obtidos com as suas respectivas discussões. Em primeiro lugar, descrevem-se as medidas utilizadas para avaliar os classificadores de aspectos (*microaspectos* e *macroaspectos*) (Seção 5.1). Em seguida, mostram-se os resultados obtidos da avaliação dos classificadores de *microaspectos* (Seção 5.2) e *macroaspectos* (Seção 5.3), respectivamente. Descreve-se, também, a medida utilizada para avaliar os sumários gerados (Seção 5.4). Depois, mostram-se os resultados obtidos da avaliação dos sumários gerados pelos métodos propostos (Seção 5.5), sendo que esses resultados foram comparados com os melhores sumarizadores multidocumento da literatura para o Português. Por último, apresentam-se as considerações finais deste capítulo (Seção 5.6).

5.1 Medidas de Avaliação de Classificadores

Como já foi dito, a identificação de aspectos é um problema de classificação multirrótulo. Com base na teoria de Tsoumakas e Katakis (2007), criaram-se vários classificadores binários para cada aspecto. Assim, apresentam-se as medidas de avaliação do desempenho para classificação binária conforme à matriz de confusão da Tab. 19.

	Verdadeiro (P)	**Falso (P)**
Verdadeiro (R)	VP	FN
Falso (R)	FP	VN

Tabela 19: Matriz de confusão

Observa-se que na linha superior da matriz estão as classes preditas (P) pelo sistema. Já na coluna da esquerda estão as classes anotadas manualmente chamadas de classes reais (R). Para ter uma estimativa de erro de classificação, dentro da matriz acham-se as seguintes quantidades:

- **Verdadeiros positivos (VP):** refere-se à quantidade de instâncias que o classificador conseguiu anotar automaticamente e que foram anotadas manualmente.
- **Falsos negativos (FN):** refere-se à quantidade de instâncias que o classificador NÃO conseguiu anotar automaticamente, mas que foram anotadas manualmente.
- **Falsos positivos (FP):** refere-se à quantidade de instâncias que o classificador conseguiu anotar automaticamente, mas que NÃO foram anotadas manualmente.
- **Verdadeiros negativos (VN):** refere-se à quantidade de instâncias em que o classificador NÃO conseguiu anotar automaticamente e que NÃO foram anotadas manualmente.

As estimativas de erro são calculadas por meio da quantidade de instâncias/exemplos, dando origem às métricas. As métricas são calculadas conforme as classes positiva (SIM) e negativa (NÃO) para cada aspecto. A seguir, explicam-se as métricas usadas neste trabalho:

- **Cobertura (classe SIM):** também chamada de "taxa verdadeira positiva". Refere-se à taxa de exemplos verdadeiramente positivos que foram classificados como tal.

$$C_S = \frac{VP}{VP + FN} \tag{13}$$

- **Cobertura (classe NÃO):** também chamada de "taxa verdadeira negativa" ou "especificidade". Refere-se à taxa de exemplos verdadeiramente negativos que foram classificados como tal.

$$C_N = \frac{VN}{VN + FP} \tag{14}$$

- **Precisão (classe SIM):** também chamada de "valor preditivo positivo". Refere-se à taxa de exemplos classificados como positivos que efetivamente o são.

$$P_S = \frac{VP}{VP + FP} \tag{15}$$

- **Precisão (classe NÃO):** também chamada de "valor preditivo negativo". Refere-se à taxa de exemplos classificados como negativos que efetivamente o são.

$$P_N = \frac{VN}{VN + FN} \tag{16}$$

- **Medida F1:** refere-se à "média harmônica" ponderada da precisão e da cobertura, em que as duas métricas têm o mesmo peso ($\alpha = 1$). O cálculo é feito tanto para a classe positiva quando para a classe negativa.

$$F_\alpha = \frac{(1 + \alpha) * P * C}{\alpha * (P + C)} \tag{17}$$

$$F_1 = \frac{2 * P * C}{P + C} \tag{18}$$

- **Acurácia:** refere-se à taxa do total de acertos (VP + VN) sobre o total de exemplos.

$$P = \frac{VP + VN}{VP + VN + FP + FN} \tag{19}$$

5.2 Avaliação da Identificação de Microaspectos

Nesta seção, avaliam-se os classificadores propostos para identificar *microaspectos*. Cabe ressaltar que os sistemas propostos (APS, APS+Regras e REMBRANDT) são **também** chamados de classificadores. Assim, os resultados dos sistemas são apresentados conforme as medidas de avaliação da Seção 5.1. No total, anotaram-se 322 sentenças nas quatro categorias principais no córpus CSTNews: "Cotidiano" (102), "Esportes" (60), "Mundo" (94) e "Política" (66) (ver Seção 3.1.1). Por um lado, utilizaram-se as

322 sentenças para avaliar os sistemas anotadores. Por outro lado, para avaliar os classificadores que utilizaram atributos léxico-semânticos, utilizou-se somente o 30% das sentenças do conjunto, já que o 70% restante foi utilizado para treinamento. A seguir, apresentam-se os resultados obtidos pelos sistemas para cada *microaspecto*. Já na Seção 5.2.9, apresentam-se os resultados dos classificadores que utilizaram a abordagem de AM.

5.2.1 WHO_AGENT

Para identificar automaticamente o aspecto WHO_AGENT, foram utilizados os sistemas APS e APS+Regras. No entanto, não foi utilizado o sistema REMBRANDT, por ser incapaz de identificar o sujeito agente da oração. O sistema foi testado sobre o córpus CSTNews com um total de 130 sentenças anotadas manualmente com o aspecto WHO_AGENT.

Na Tab. 20, apresentam-se os resultados dos sistemas testados. Observa-se que o sistema APS foi um pouco melhor que o sistema APS+Regras. Porém, um grande defeito do sistema APS é anotar qualquer entidade sujeito da oração, mesmo sem saber se aquela entidade é uma pessoa ou organização. Assim, existe a probabilidade do sistema APS ter acertado muito (0.815) e ter errado muito também (0.522). As regras do WHO_AGENT foram criadas com base nos "falsos positivos", com a finalidade de melhorar a precisão. Contudo, essa melhora de 0.664 fez que a cobertura diminuísse a 0.592. Mesmo assim, o sistema APS+Regras é mais confiável, porque consegue saber se o sujeito da ação é uma entidade pessoa/organização, conforme a definição do aspecto WHO_AGENT (ver Apêndice A).

WHO_AGENT	**Cobertura**	**Precisão**	**F1**	**Acurácia**
APS	0.815	0.522	0.637	0.624
APS+Regras	**0.592**	**0.664**	**0.626**	**0.714**

Tabela 20: Resultados para o *microaspecto* WHO_AGENT

Na Tab. 21, apresenta-se a matriz de confusão do sistema APS+Regras. Nota-se que das 130 sentenças anotadas, só 77 foram anotadas corretamente e 53 não foram anotadas. A seguir, analisam-se algumas sentenças identificadas como "falsos negativos"

e “falsos positivos” do sistema APS+Regras.

WHO_AGENT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	77	53	**SIM**	0.592	0.664	0.626	0.714
Falso	39	153	**NÃO**	0.797	0.743	0.769	

Tabela 21: Matriz de confusão do *microaspecto* WHO_AGENT

Falsos negativos

Na Tab. 22, mostram-se algumas sentenças cujos aspectos não foram identificados automaticamente pelo sistema APS+Regras, mas que foram anotados manualmente. Pode-se observar, em nível discursivo, que na sentença da categoria “Cotidiano”, a entidade “o presidente” atua como WHO_AGENT, porque foi a entidade que fez a afirmação introduzida pelo segmento “Segundo”. Contudo, em nível semântico, o sistema não consegue identificar o segmento “o presidente” como agente por não haver um verbo elocutivo (p. ex: “o presidente afirmou que”, “o presidente disse que”). Por outro lado, na categoria “Esportes”, o segmento “o Brasil” representa o agente relacionado ao verbo “estar (está)”, mas, mesmo assim, o classificador não conseguiu identificá-lo como WHO_AGENT. Da mesma maneira, na categoria “Mundo”, o segmento “A agência meteorológica do Japão” atua como agente do verbo “chegar (chegou)”, mas não foi anotado pelo sistema. Por último, na categoria “Política”, o segmento “pela Mesa Diretora do Senado” representa o agente da passiva, portanto, é difícil para o classificador identificalo como agente.

Categoria	Sentença
Cotidiano	Segundo o presidente, a prioridade é a realização de obras nas regiões metropolitanas de grandes centros urbanos.
Esportes	Com o resultado, o Brasil está na liderança do grupo B, perto da classificação para a próxima fase do campeonato.
Mundo	A agência meteorológica do Japão chegou a emitir alerta de Tsunami, mas o cancelou uma hora após.
Política	Amanhã será decidido, pela Mesa Diretora do Senado, se a quarta representação será encaminhada ao Conselho de Ética.

Tabela 22: Falsos negativos do *microaspecto* WHO_AGENT

Falsos positivos

Na Tab. 23, mostram-se algumas das sentenças cujos aspectos foram identificados automaticamente pelo sistema APS+Regras, mas que não foram anotadas manualmente. Observa-se que nas categorias "Cotidiano", "Esportes" e "Política" a, os segmentos "Lula", "Maradona" e "Ele", respectivamente, foram identificados corretamente como WHO_AGENT, porém não foram anotadas manualmente. Isso pode ter sido erro de anotação humana. Na sentença da categoria "Mundo", o segmento "O furação Dean" não representa uma organização/pessoa; o sistema errou ao considerar erroneamente "Dean" uma pessoa.

Categoria	**Sentença**
Cotidiano	Depois de as vaias, <aspect APS="WHO_AGENT">Lula</aspect> **desistiu** de declarar abertos os jogos, como estava planejado.
Esportes	<aspect APS="WHO_AGENT">Maradona</aspect> **voltou** a ter problemas de saúde em o fim de semana e foi internado novamente em um hospital em Buenos Aires.
Mundo	<aspect APS="WHO_AGENT">O furacão Dean</aspect> **passou** por a costa sul de a Jamaica, inundando a capital e espalhando árvores e telhados.
Política	<aspect APS="WHO_AGENT">Ele</aspect> **disse** que é ideal que se crie uma comissão de três relatores para os processos em conjunto.

Tabela 23: Falsos positivos do *microaspecto* WHO_AGENT

5.2.2 WHO_AFFECTED

Para identificar o aspecto WHO_AFFECTED, foram utilizados os sistemas APS e APS+Regras. No entanto, não foi utilizado o sistema REMBRANDT, por este ser incapaz de identificar o sujeito paciente da oração. O sistema foi testado sobre o córpus CSTNews com um total de 60 sentenças anotadas manualmente com o aspecto WHO_AFFECTED. Na Tab. 24, apresentam-se os resultados dos sistemas avaliados. Observa-se que o sistema APS+Regras superou o sistema APS. O uso de regras para identificar a existência de uma entidade pessoa/organização dentro de um segmento teve êxito. Porém, a cobertura diminuiu de 0.767 a 0.417. Isso quer dizer que, assim como para o aspecto WHO_AGENT, o sistema APS estava acertando por acaso.

WHO_AFFECTED	Cobertura	Precisão	F1	Acurácia
APS	0.767	0.203	0.321	0.394
APS+Regras	**0.417**	**0.368**	**0.391**	**0.758**

Tabela 24: Resultados para o *microaspecto* WHO_AFFECTED

Na Tab. 25, mostra-se a matriz de confusão do sistema APS+Regras. Nota-se que, das 60 sentenças anotadas, só 25 sentenças foram anotadas pelo sistema APS+Regras corretamente. A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos" do sistema APS+Regras.

WHO_AFFECTED	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	25	35	**SIM**	0.417	0.368	0.391	0.758
Falso	43	219	**NÃO**	0.836	0.862	0.849	

Tabela 25: Matriz de confusão do *microaspecto* WHO_AFFECTED

Falsos negativos

Na Tab. 26, mostram-se algumas das sentenças cujos aspectos não foram identificados automaticamente pelo sistema APS+Regras. Observa-se claramente que, na sentença da categoria "Cotidiano", o classificador não conseguiu identificar o segmento "200 pessoas" como sujeito paciente da oração, relacionado ao verbo "vitimar (vitimado)". Isso é um erro do sistema APS. Caso similar ocorre na sentença da categoria "Mundo" e "Política". Já na sentença da categoria "Esportes", acontece um problema de sujeito oculto. Assim, a entidade "Maradona", relacionada ao verbo "internar (internado)", não foi identificada.

Categoria	Sentença
Cotidiano	Em o acidente, o avião passou por a pista de Congonhas com velocidade acima de o normal, atravessou uma avenida e atingiu um prédio, vitimando 200 pessoas.
Esportes	Maradona voltou a ter problemas de saúde em o fim de semana e foi internado novamente em um hospital em Buenos Aires.
Mundo	O terremoto deixou 9 pessoas mortas (todos idosos) e mais de 700 feridos, além de casas e viadutos destruídos.
Política	Cristovam Buarque e Luciano Bivar têm, cada um, 1% dos votos.

Tabela 26: Falsos negativos do *microaspecto* WHO_AFFECTED

Falsos positivos

Na Tab. 27, apresentam-se algumas das sentenças que o sistema APS+Regras não devia ter identificado. Nota-se que na sentença da categoria "Cotidiano", a organização "A secretaria da Fazenda" foi identificado corretamente, porém, essa sentença não foi anotada manualmente. Caso similar ocorre com a sentença da categoria "Esportes". Um problema do classificador APS+Regras é que, na maioria dos casos, anotam-se segmentos posicionados à esquerda do verbo alvo. Isso pode ser observado nas sentenças das categorias "Mundo" e "Política", em que os segmentos "de a polícia" e "por a bancada de o PT", respectivamente, foram erroneamente classificados.

Categoria	**Sentença**
Cotidiano	<aspect APS="WHO_AFFECTED">A Secretaria da Fazenda</aspect> também **foi atingida** por uma bomba.
Esportes	Ronaldinho fez uma sequência de dribles em o Equador e **cruzou** <aspect APS="WHO_AFFECTED">para Elano</aspect>, que fez o quarto gol.
Mundo	Depois de isso, fez quatro cirurgias plásticas para **escapar** <aspect APs="WHO_AFFECTED">de a polícia</aspect>.
Política	A unificação **foi proposta** <aspect APS="WHO_AFFECTED">por a bancada de o PT </aspect> e tem apoio de o PSOL.

Tabela 27: Falsos positivos do *microaspecto* WHO_AFFECTED

5.2.3 WHEN

Foram testados 4 tipos de sistemas: APS, REMBRANDT, APS+REMBRANDT e APS+Regras. Os sistemas foram testados sobre o córpus CSTNews com um total de 75 sentenças anotadas manualmente com o aspecto WHEN. Os resultados dos 4 sistemas são apresentados na Tab. 28. O sistema APS+Regras obteve os melhores resultados.

WHEN	**Cobertura**	**Precisão**	**F1**	**Acurácia**
APS	0.693	0.477	0.565	0.752
REMBRANDT	0.547	0.719	0.621	0.845
APS+REMBRANDT	0.840	0.492	0.621	0.761
APS+Regras	**0.947**	**0.504**	**0.657**	**0.770**

Tabela 28: Resultados para o *microaspecto* WHEN

Na Tab. 29, apresenta-se a matriz de confusão do melhor sistema (APS+Regras) para o *microaspecto* WHEN. Mesmo assim, a precisão é relativamente baixa (0.504), por causa da grande quantidade de "falsos positivos". A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

WHEN	**Verdadeiro (P)**	**Falso (P)**	**Classe**	**Cobertura**	**Precisão**	**F1**	**Acurácia**
Verdadeiro	71	4	**SIM**	0.947	0.504	0.657	0.770
Falso	70	177	**NÃO**	0.717	0.978	0.827	

Tabela 29: Matriz de confusão do *microaspecto* WHEN

Falsos negativos

Na Tab. 30, mostram-se algumas das sentenças que o sistema APS+Regras não devia ter identificado. Observa-se que na sentença da categoria "Cotidiano", os segmentos "às 8h", "às 9h" e "meia hora depois" não foram identificados. Isso quer dizer que deve ser criada uma regra para identificar tempo em formato de horas. Na sentença da categoria "Política", não foi identificado o advérbio de tempo "Amanhã". Tentou-se solucionar esse problema criando uma regra que identifica advérbios de tempo isolados, ou seja, que não foram antecedidos por alguma preposição (ver regra nº 4 da Fig.17). Porém, a regra identificava muitas sentenças que não foram anotadas manualmente, gerando muitos "falsos positivos". Por causa disso, optou-se por não a utilizar.

Categoria	**Sentença**
Cotidiano	A Companhia de Engenharia de Tráfego (CET) anunciou que o índice de congestionamento era de 54 quilômetros às 8h, 113 km às 9h e 110 km meia hora depois, valores bem acima de as médias para os horários, que eram de 36, 82 e 76 quilômetros respectivamente.
Política	Amanhã será decidido, por a Mesa Diretora de o Senado, se a quarta representação será encaminhada a o Conselho de Ética.

Tabela 30: Falsos negativos do *microaspecto* WHEN

Na Tab. 31, mostram-se algumas das sentenças que o sistema APS+Regras não devia ter identificado. As sentenças da categoria "Cotidiano", "Mundo" e "Política" foram anotadas corretamente, mas não foram anotadas manualmente (erro de anotação humana). Já na sentença da categoria "Esportes", o classificador errou totalmente ao classificar o

segmento “a 0” como tempo.

Categoria	Sentença
Cotidiano	<aspect APS=“WHEN”>Antes de a festa</aspect>, ele visitou a Vila Olímpica e conversou com atletas de vários países de o mundo que estavam lá e ouviu de vários de eles elogios sobre a qualidade de o que o Brasil estava oferecendo em a Vila Olímpica.
Esportes	A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets <aspect APS=“WHEN”>a 0</aspect>, em Tampere (FIN), mantendo sua invencibilidade em a Liga Mundial de Vôlei-06.
Mundo	Esta proposta será debatida por a ONU <aspect APS=“WHEN”>hoje</aspect> ou amanhã.
Política	O grupo criminoso desviou <aspect APS=“WHEN”>desde 2004</aspect> cerca de R$ 70 milhões de os cofres públicos.

Tabela 31: Falsos positivos do *microaspecto* WHEN

5.2.4 WHERE

Da mesma forma que para o *microaspecto* WHEN, avaliaram-se quatro tipos de sistemas: APS, REMBRANDT, APS+REMBRANDT e APS+Regras. Os sistemas foram testados sobre o córpus CSTNews com um total de 56 sentenças anotadas manualmente com o aspecto WHERE. Na Tab. 32, apresentam-se os resultados dos sitemas anotadores. Nota-se que o sistema APS+Regras obteve os melhores resultados.

WHERE	Cobertura	Precisão	F1	Acurácia
APS	0.679	0.447	0.539	0.798
REMBRANDT	0.804	0.425	0.556	0.776
APS+REMBRANDT	0.946	0.363	0.525	0.702
APS+Regras	**0.804**	**0.474**	**0.596**	**0.811**

Tabela 32: Resultados para o *microaspecto* WHERE

Na Tab. 33, apresenta-se a matriz de confusão do melhor sistema (APS+Regras) para o *microaspecto* WHERE. Observa-se que a precisão é baixa (0.474) por causa da grande quantidade de “falsos positivos”. A seguir, analisam-se algumas sentenças identificadas como “falsos negativos” e “falsos positivos”.

WHERE	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	45	11	**SIM**	0.804	0.474	0.596	0.811
Falso	50	216	**NÃO**	0.812	0.952	0.876	

Tabela 33: Matriz de confusão do *microaspecto* WHERE

Falsos negativos

Na Tab. 34, mostram-se algumas das sentenças cujos aspectos não foram identificados pelo sistema APS+Regras. Nota-se que o segmento "para a Bahia", na categoria "Cotidiano", contém a preposição "para" em lugar da preposição "em" (ver Fig. 18). Da mesma forma, o segmento "Norris Hall" não está associado à preposição "em".

Categoria	Sentença
Cotidiano	A família pediu que o corpo fosse levado diretamente para a Bahia.
Mundo	Pouco depois, Norris Hall, edifício de a engenharia, foi alvo de outro ataque a tiros.

Tabela 34: Falsos negativos do *microaspecto* WHERE

Falsos positivos

Na Tab. 35, mostram-se algumas das sentenças cujos aspectos o sistema APS+Regras não devia ter identificado como WHERE. Tanto na categoria "Cotidiano" quanto na categoria "Política", o sistema classifica corretamente os segmentos "em o Palácio de Aclamação" e "em Rondônia" como expressões de lugar, contudo, não foram anotadas manualmente. Já na categoria "Esportes" e "Mundo", o sistema APS identificou erroneamente os segmentos "em o salto com vara" e "em um de seus transformadores".

Categoria	Sentença
Cotidiano	O velório será <aspect APS="WHERE"> em o Palácio da Aclamação. </aspect>
Esportes	A brasileira Fabiana Murer conquistou a medalha de ouro <aspect APS="WHERE">em o salto com vara</aspect> a o saltar 4m60.
Mundo	A maior usina nuclear de o mundo teve incêndio <aspect APS="WHERE">em um de seus transformadores</aspect>, mas o fogo foi controlado.
Política	Algumas pessoas pertencem a o alto escalão político <aspect APS="WHERE">em Rondônia.</aspect>.

Tabela 35: Falsos positivos do *microaspecto* WHERE

5.2.5 WHY

No total, avaliaram-se dois tipos de sistemas: APS e APS+Regras. O sistema REMBRANDT foi descartado por não possuir uma categoria equivalente ao *microaspecto* WHY. Os sistemas foram testados sobre o córpus CSTNews com um total de 32 sentenças anotadas manualmente com o aspecto WHY. Na Tab. 36, mostram-se os resultados dos sistemas avaliados. Nota-se que o sistema APS+Regras ganhou do sistema APS por uma considerável diferença em todas as métricas, com exceção da acurácia. Mesmo assim, a cobertura é baixa (0.469).

WHY	**Cobertura**	**Precisão**	**F1**	**Acurácia**
APS	0.156	0.500	0.238	0.901
APS+Regras	**0.469**	**0.789**	**0.588**	**0.935**

Tabela 36: Resultados para o *microaspecto* WHY

Na Tab. 37, apresenta-se a matriz de confusão do melhor sistema (APS+Regras) para o *microaspecto* WHY. Pode-se observar que, mesmo com o auxílio das regras, encontram-se 17 "falsos negativos" de 32 sentenças anotadas. A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

WHY	**Verdadeiro (P)**	**Falso (P)**	**Classe**	**Cobertura**	**Precisão**	**F1**	**Acurácia**
Verdadeiro	15	17	**SIM**	0.469	0.789	0.588	0.935
Falso	4	286	**NÃO**	0.986	0.944	0.966	

Tabela 37: Matriz de confusão do *microaspecto* WHY

Falsos negativos

Na Tab. 38, mostram-se algumas das sentenças cujos aspectos não foram identificados pelo sistema APS+Regras, mas que foram anotadas manualmente. Nota-se que, nas sentenças da categoria "Esportes" e "Mundo", não existe uma expressão de causa explícita, dessa maneira, as regras não conseguiram classificar as sentenças como WHY (ver Fig. 19). Por outro lado, a sentença da categoria "Política" não representa uma causa explícita ou implícita, portanto, não devia ser anotada manualmente.

Categoria	Sentença
Esporte	Maradona voltou a ter problemas de saúde em o fim de semana e foi internado novamente em um hospital em Buenos Aires.
Mundo	Duas mulheres de 80 anos morreram no desmoronamento de suas casas.
Política	Tem candidato aí com o salto 15, eu vou nas sandálias da humildade, disse Alckmin, apontando a arrogância do adversário.

Tabela 38: Falsos negativos do *microaspecto* WHY

Falsos positivos

Na Tab. 39, mostram-se algumas das sentenças que o sistema APS+Regras não devia ter identificado como WHY. Nota-se que, na sentença da categoria "Cotidiano", identificou-se a expressão "pois", porém, a expressão não denota uma causa, estritamente falando. Já na categoria "Política", a expressão "porque" representa uma causa explícita, portanto, devia ser anotada manualmente.

Categoria	Sentença
Cotidiano	Seu discurso foi focado em as questões ambientais, citando que a produção de biocombustíveis não afeta a segurança alimentar, <aspect APS="WHY">pois</aspect> a cana-de-açúcar ocupa apenas 1% de as terras agricultáveis de o Brasil.
Política	Eu não moverei uma palha contra a oposição <aspect APS="WHY">porque</aspect> vocês moverão um paiol inteiro, afirmou o presidente Luiz Inácio Lula da Silva, candidato à reeleição por o PT, sobre os ataques de seus adversários.

Tabela 39: Falsos positivos do *microaspecto* WHY

5.2.6 HOW

O único sistema testado foi o APS. O sistema REMBRANDT foi desconsiderado por não ter uma categoria equivalente ao aspecto HOW. Como já foi dito, não foi possível criar regras manuais porque existem muito poucas sentenças anotadas e não foi possível identificar padrões. O sistema foi testado sobre o córpus CSTNews com um total de 9 sentenças anotadas manualmente com o aspecto HOW. Na Tab. 40, mostra-se a matriz

de confusão do sistema APS para o *microaspecto* HOW. Observa-se que a medida F1 foi baixa (0.040) em função da grande quantidade de "falsos positivos". A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

HOW	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	8	**SIM**	0.111	0.024	0.040	0.851
Falso	40	273	**NÃO**	0.872	0.972	0.919	

Tabela 40: Matriz de confusão do *microaspecto* HOW

Falsos negativos

Na Tab. 41, apresentam-se algumas das sentenças cujos aspectos o sistema APS não identificou automaticamente como HOW. Nota-se que, nas sentenças da categoria "Cotidiano" e "Política", os segmentos "com bombas e tiros" e "em dois turnos" não foram identificados pelo sistema.

Categoria	Sentença
Cotidiano	Os bandidos atacaram agências bancárias, bases policiais e prédios públicos com bombas e tiros.
Política	Para ser aprovada, a PEC precisa ser votada em dois turnos da Câmara.

Tabela 41: Falsos negativos do *microaspecto* HOW

Falsos positivos

Na Tab. 42, mostram-se algumas das sentenças que o sistema APS não devia ter identificado como HOW. Nota-se que os segmentos marcados nas sentenças das categorias "Cotidiano" e "Política" foram erradamente identificados como HOW, por seguir o padrão "com" e "como" no começo dos segmentos.

Categoria	Sentença
Cotidiano	A falha em o reversor – mecanismo que ajuda o avião a frear – foi detectada por o sistema de a aeronave, que continuou voando em os dias seguintes <aspect APS="HOW">com o reversor desligado</aspect>.
Política	O assunto surgiu depois que seu advogado questionou a legitimidade de o colegiado e de a própria Polícia Federal em investigar o caso, uma vez que, <aspect APS="HOW">como senador, Renan gozaria de foro especial</aspect>.

Tabela 42: Falsos positivos do *microaspecto* HOW

5.2.7 SITUATION

O sistema escolhido foi o REMBRANDT. O sistema APS foi descartado por não ter um papel semântico equivalente ao aspecto SITUATION. O sistema foi testado sobre o córpus CSTNews com um total de 13 sentenças anotadas manualmente. Na Tab. 43, mostra-se a matriz de confusão do sistema REMBRANDT para o *microaspecto* SITUATION. Nota-se que a cobertura foi baixa (0.231), enquanto a precisão foi alta (0.750). A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

SITUATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	3	10	**SIM**	0.231	0.750	0.353	0.966
Falso	1	308	**NÃO**	0.997	0.969	0.983	

Tabela 43: Matriz de confusão do *microaspecto* SITUATION

Falsos negativos

Na Tab. 44, mostram-se algumas das sentenças que o sistema REMBRANDT não conseguiu identificar como SITUATION. Nota-se que o segmento "Liga Mundial de Vôlei-06", na categoria "Esportes", é uma entidade mencionada que indica um evento competitivo, portanto, devia ser identificada pelo sistema. Por outro lado, o segmento "em esta batalha", na categoria "Mundo", não é uma entidade mencionada (ou nome próprio), portanto, o sistema não conseguiu identifica-la.

Categoria	Sentença
Esportes	A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.
Mundo	Nesta batalha, 15 soldados israelenses morreram ao serem atingidos por um míssil.

Tabela 44: Falsos negativos do *microaspecto* SITUATION

Falsos positivos

Na Tab. 45, apresenta-se a única sentença que o sistema REMBRANDT identificou automaticamente como SITUATION, mas que não foi anotada manualmente. Nota-se claramente que o segmento "Jogos Olímpicos de Pequim" indica uma competição, mas

não foi anotado manualmente (erro de anotação).

Categoria	Sentença
Esportes	A ginasta Jade Barbosa foi escolhida, em votação na Internet, para ser a representante do Brasil no revezamento da tocha dos <aspect EM= "ACONTECIMENTO|ORGANIZADO">Jogos Olímpicos de Pequim </aspect>

Tabela 45: Falsos positivos do *microaspecto* SITUATION

5.2.8 SCORE

O único sistema testado foi o APS+Regras. É importante lembrar que o aspecto SCORE não tem equivalência com algum papel semântico, portanto, só foram criadas regras manuais e integradas ao sistema. Já o sistema REMBRANDT foi desconsiderado por não ter uma categoria equivalente ao aspecto SCORE. O sistema foi testado sobre o córpus CSTNews com um total de 10 sentenças anotadas manualmente, Na Tab. 46, mostra-se a matriz de confusão do sistema APS+Regras para o aspecto SCORE. Observa-se que as regras manuais tiveram um ótimo desempenho, identificando todas as sentenças anotadas manualmente. Cabe ressaltar que as regras foram criadas sobre uma quantidade mínima de sentenças anotadas, por isso o resultado é excelente.

SCORE	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	10	0	**SIM**	1.000	1.000	1.000	1.000
Falso	0	312	**NÃO**	1.000	1.000	1.000	

Tabela 46: Matriz de confusão do *microaspecto* SCORE

5.2.9 Resultados dos Classificadores usando Atributos Léxico-Semânticos

Diferentemente dos sistemas testados anteriormente, na abordagem usando AM os classificadores foram treinados e testados com as 322 sentenças (ou instâncias) do córpus CSTNews, anotadas com aspectos. A estratégia de treinamento e teste foi de repetidas divisões em subconjuntos (em várias iterações) com *estratificação*, já que se garante que haja as mesmas proporções de classes dentro de cada subconjunto. A ideia de se usar córpus estratificado é de amenizar o problema de "desbalanceamento de classes", que pode influenciar no desempenho do classificador (Newman e Merz, 1998).

Portanto, o córpus foi 10 vezes estratificado aleatoriamente, sendo que, para cada iteração, a divisão do córpus foi de 70% para o conjunto de treinamento (225 instâncias) e 30% para o conjunto de teste (97 instâncias). Não foi utilizado o tradicional *10-fold cross-validation* porque, devido ao córpus ser muito pequeno, o *fold* de teste teria poucas instâncias. Com a técnica de estratificação, garante-se uma melhor distribuição de classes e, por conseguinte, resultados mais justos.

Como já foi dito, a técnica de AM supervisionada usada foi SVM. No trabalho futuro, serão utilizadas outras técnicas, como Árvores de Decisão (Breiman et al., 1984), Redes Neurais (Haykin, 1998) ou Redes Bayesianas (Mitchell, 1997).

A avaliação de cada classificador foi feita conforme as métricas estatísticas obtidas da matriz de confusão: "Precisão", "Cobertura", "F1" e "Acurácia". O resultado final é a "**média dos valores obtidos em cada uma das 10 iterações do córpus estratificado**". Esta abordagem atende os *microaspectos* WHO_AGENT, WHO_AFFECTED, WHERE, WHEN, WHY, HOW, SITUATION e SCORE.

Na Tab. 47, mostram-se os resultados dos melhores classificadores que utilizaram atributos léxico-semânticos para cada *microaspecto*. Observa-se que o desempenho dos classificadores, em termos de F1, é baixo na maioria dos casos, a exceção dos aspectos WHO_AGENT, WHEN e WHERE. Isso se deve às poucas instâncias anotadas. O melhor resultado foi obtido pelo classificador "*(1, 1) semantic*" para o aspecto WHEN. O pior resultado foi obtido por todas as combinações de classificadores para o aspecto SCORE. A causa do aspecto SCORE ter ido mal é a pouca quantidade de sentenças anotadas (10) no córpus. Nota-se que a maioria dos classificadores é representada por unigramas "(1, 1)". Por último, o atributo mais representativo é "*POS+semantic*", alcançando os melhores resultados nos aspectos WHO_AGENT, WHERE e WHY. Na Tab. 71 do Apêndice C, mostram-se os resultados de todos os classificadores utilizando atributos léxico-semânticos.

Microaspecto	Classificador	Cobertura	Precisão	F1	Acurácia
WHO_AGENT	(2, 3) POS+semantic	0.538	0.636	0.583	0.691
WHO_AFFECTED	(1, 1) lemmas	0.222	1.000	0.364	0.854
WHEN	(1, 1) semantic	0.522	0.750	0.615	0.845
WHERE	(2, 3) POS+semantic	0.471	0.615	0.533	0.856
WHY	(2, 3) POS+semantic	0.200	0.500	0.286	0.897
HOW	(1, 1) bag_of_words	0.250	1.000	0.400	0.938
SITUATION	(1, 1) lemmas+POS	0.333	1.000	0.500	0.959
SCORE	Todos	0.000	0.000	0.000	0.000

Tabela 47: Melhores classificadores de *microaspectos* usando atributos léxico-semânticos

5.3 Avaliação dos Classificadores de Macroaspectos

Nesta seção, avaliam-se os classificadores propostos para identificar *macroaspectos*. De igual maneira que na Seção 5.2.9, na abordagem usando AM, o conjunto das 322 sentenças anotadas do córpus CSTNews foi 10 vezes estratificado, sendo que, para cada iteração, o conjunto foi dividido em 70% para treino (225 instâncias) e 30% para teste (97 instâncias). Assim, o resultado final é a **media dos valores obtivos em cada uma das 10 iterações**. Esta abordagem atende os *macroaspectos* WHAT, CONSEQUENCE, COMMENT, DECLARATION e HISTORY.

Diferentemente da abordagem anterior, na abordagem usando regras compilaram-se as 322 sentenças do córpus para teste. Para avaliar esta abordagem, usaram-se as mesmas métricas de avaliação que na abordagem com AM. Como já foi dito, somente foram criadas regras manuais para os *macroaspectos* COMPARISON, DECLARATION, GOAL, HISTORY e PREDICTION. A seguir, apresentam-se os resultados obtidos pelos classificadores para cada *macroaspecto*.

5.3.1 WHAT

Os classificadores foram testados sobre um total de 60 sentenças anotadas manualmente com o aspecto WHAT. Na Tab. 48, apresentam-se os resultados do classificador

usando os atributos definidos por Teufel (1999). Observa-se que o classificador para a classe "sim" teve melhores resultados do que para classe "não". Para a classe "sim", a cobertura (0.660) foi melhor que a precisão (0.550). Já na Tab. 49, apresentam-se os resultados do melhor classificador usando atributos léxico-semânticos (ver Tab. 72 do Apêndice C), denominado "*(2, 2) bag_of_words*", criado com base em todos os bigramas "(2,2)" de todas as palavras do córpus. Observa-se que a classe "sim" foi melhor do que a classe "não" com uma alta cobertura (0.800) e uma precisão relativamente baixa (0.519). Cabe ressaltar que os resultados são bons por causa da grande quantidade de sentenças anotadas com WHAT.

WHAT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	33	17	**SIM**	0.660	0.550	0.600	0.546
Falso	27	20	**NÃO**	0.426	0.541	0.476	

Tabela 48: Resultados do *macroaspecto* WHAT usando atributos de Teufel (1999)

WHAT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	40	10	**SIM**	0.800	0.519	0.630	0.515
Falso	37	10	**NÃO**	0.213	0.500	0.299	

Tabela 49: Resultados do *macroaspecto* WHAT usando atributos léxico-semânticos

5.3.2 CONSEQUENCE

Os classificadores foram testados sobre um total de 14 sentenças anotadas manualmente com o aspecto CONSEQUENCE. Na Tab. 51, apresentam-se os resultados do classificador usando os atributos definidos por Teufel (1999). Observa-se que o classificador para a classe "não" teve melhores resultados do que para classe "sim". Para a classe "sim", tanto a cobertura quanto a precisão são nulos. Os resultados claramente mostram que não é possível identificar CONSEQUENCE usando os atributos de Teufel (1999). Por outro lado, na Tab. 51, mostra-se o melhor classificador usando atributos léxico-semânticos: "*(1, 1) lemmas*", criado com base em todos os unigramas "(1, 1)" de todas os lemas das palavras do córpus. Para a classe "sim", o classificador obteve uma cobertura quase nula (0.071) e uma precisão perfeita (1.000). Embora o classificador

usando atributos léxico-semânticos seja o melhor, não deve ser considerado como um classificador apto para identificar CONSEQUENCE, devido aos resultados quase nulos. Em conclusão, não foi possível identificar o aspecto CONSEQUENCE.

CONSEQUENCE	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	14	**SIM**	0.000	0.000	0.000	0.856
Falso	0	83	**NÃO**	1.000	0.856	0.922	

Tabela 50: Resultados do *macroaspecto* CONSEQUENCE usando atributos de Teufel (1999)

CONSEQUENCE	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	13	**SIM**	0.071	1.000	0.133	0.866
Falso	0	83	**NÃO**	1.000	0.865	0.927	

Tabela 51: Resultados do *macroaspecto* CONSEQUENCE usando atributos léxico-semânticos

5.3.3 COMPARISON

Na Tab. 52, apresentam-se os resultados usando regras manuais sobre um conjunto de 6 sentenças anotadas manualmente com os aspecto COMPARISON. Observa-se que a classe "não" teve melhores resultados do que a classe "sim". No entanto, a classe "sim" obteve uma medida F1 relativamente alta (0.667). A acurácia também foi bastante alta (0.991). Os resultados mostram que é possível identificar COMPARISON usando regras manuais. Cabe ressaltar que os resultados são bons por causa da pouca quantidade de sentenças anotadas, o que indica um possível *overfitting*.

4.3. COMPARISON	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	3	3	**SIM**	0.500	1.000	0.667	0.991
Falso	0	316	**NÃO**	1.000	0.991	0.995	

Tabela 52: Resultados do *macroaspecto* COMPARISON usando atributos de Teufel (1999)

5.3.4 COMMENT

Os classificadores foram testados sobre um total de 7 sentenças anotadas manualmente com o aspecto COMMENT. Na Tab. 53, apresentam-se os resultados do classificador usando os atributos de Teufel (1999). Observa-se que o classificador para a classe "não" teve melhores resultados do que para classe "sim". Para a classe "sim", tanto a cobertura quanto a precisão são nulos. Os resultados afirmam que não é possível identificar COMMENT usando os atributos de Teufel (1999). Já na Tab. 54, apresentam-se os resultados do melhor classificador usando atributos léxico-semânticos: "*(2, 2) semantic*", criado com base em todos os bigramas "(2, 2)" das etiquetas semânticas de todas as palavras do córpus. Para a classe "sim", o classificador obteve uma cobertura baixa (0.143) e uma precisão perfeita (1.000). Mesmo que o classificador com base nos atributos léxico-semânticos seja o melhor (0.025 de F1), não é um classificador competente para identificar COMMENT. Em conclusão, não é possível identificar COMMENT. Cabe ressaltar que os resultados são baixos devido a pouca quantidade de sentenças anotadas.

COMMENT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	7	**SIM**	0.000	0.000	0.000	0.928
Falso	0	90	**NÃO**	1.000	0.928	0.963	

Tabela 53: Resultados do *macroaspecto* COMMENT usando atributos de Teufel (1999)

COMMENT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	6	**SIM**	0.143	1.000	0.025	0.938
Falso	0	90	**NÃO**	1.000	0.938	0.968	

Tabela 54: Resultados do *macroaspecto* COMMENT usando atributos léxico-semânticos

5.3.5 DECLARATION

Os classificadores foram testados sobre um total de 17 sentenças anotadas manualmente com o aspecto COMMENT. Na Tab. 55, apresentam-se os resultados do classificador usando os atributos definidos por Teufel (1999). Nota-se que o classificador para a classe "não" teve melhores resultados do que para a classe "sim". Para a classe

"sim", tanto a cobertura quanto a precisão são nulas. Os resultados afirmam que não é possível identificar DECLARATION usando os atributos de Teufel (1999). Na Tab. 55, mostram-se os resultados do melhor classificador usando atributos léxico-semânticos: "*(1, 1) lemmas+POS*", criado com base em todos os unigramas "(1, 1)" do lema junto com a classe gramatical de todas as palavras do córpus. Para a classe "sim", o classificador obteve uma cobertura média (0.529) e uma precisão bastante alta (0.900). Assim, o classificador usando atributos léxico-semânticos obteve os melhores resultados (0.667 de F1).

DECLARATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	17	**SIM**	0.000	0.000	0.000	0.825
Falso	0	80	**NÃO**	1.000	0.825	0.904	

Tabela 55: Resultados do *macroaspecto* DECLARATION usando atributos de Teufel (1999)

DECLARATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	9	8	**SIM**	0.529	0.900	0.667	0.907
Falso	1	79	**NÃO**	0.988	0.908	0.946	

Tabela 56: Resultados do *macroaspecto* DECLARATION usando atributos léxico-semânticos

Na Tab. 57, apresentam-se os resultados usando regras manuais sobre um conjunto de 58 sentenças anotadas manualmente com os aspecto DECLARATION. Observa-se que a classe "não" teve melhores resultados do que a classe "sim" por uma diferença mínima. Tanto a cobertura (0.879) quanto a precisão (0.944) para a classe "sim" foram altas, obtendo-se, por consequência, uma medida F1 bastante alta (0.911). Cabe ressaltar que a acurácia também foi bastante alta (0.969). Os resultados claramente mostram que é possível identificar DECLARATION usando regras manuais (ver Fig. 25).

DECLARATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	51	7	**SIM**	0.879	0.944	0.911	0.969
Falso	3	261	**NÃO**	0.989	0.974	0.981	

Tabela 57: Resultados do *macroaspecto* DECLARATION usando regras manuais

5.3.6 GOAL

Na Tab. 58, apresentam-se os resultados usando regras manuais sobre um conjunto de 10 sentenças anotadas manualmente com os aspecto GOAL. Observa-se que a classe "não" teve melhores resultados do que a classe "sim". Para a classe "sim", a cobertura foi baixa (0.400), enquanto a precisão foi alta (0.800). Ressalta-se, também, o bom desempenho em termos de acurácia (0.978). Os resultados mostram que é possível identificar GOAL usando regras manuais (ver Fig. 28). Também é preciso dizer que foi fácil de se identificar regras por causa da pouca quantidade de instâncias anotadas, e isso pode gerar *overfitting*.

GOAL	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	4	6	**SIM**	0.400	0.800	0.533	0.978
Falso	1	311	**NÃO**	0.997	0.981	0.989	

Tabela 58: Resultados do *macroaspecto* GOAL usando regras manuais

5.3.7 HISTORY

Os classificadores foram testados sobre um total de 9 sentenças anotadas manualmente com o aspecto HISTORY. Na Tab. 59, apresentam-se os resultados do classificador usando os atributos definidos por Teufel (1999). Nota-se que o classificador para a classe "não" teve melhores resultados do que para a classe "sim". Para a classe "sim", tanto a cobertura quanto a precisão são nulas, portanto, a medida F1 também é nula (0.000). Os resultados mostram claramente que não é possível identificar HISTORY usando os atributos de Teufel (1999). Já na Tab. 60, mostram-se os resultados do classificador usando atributos léxico-semânticos: "*(2, 3) semantic*", criado com base em todos os bigramas e trigramas "(2, 3)" das etiquetas semânticas de todas as palavras do córpus. O classificador obteve uma cobertura bastante baixa (0.111) e uma precisão média (0.500). Embora o classificador baseado em atributos léxico-semânticos tenha obtido os melhores resultados, não é apto para identificar HISTORY, por causa do baixo desempenho. É importante dizer que os resultados são bastante baixos por causa da pouca quantidade de sentenças anotadas com HISTORY.

HISTORY	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	9	**SIM**	0.000	0.000	0.000	0.907
Falso	0	88	**NÃO**	1.000	0.907	0.951	

Tabela 59: Resultados do *macroaspecto* HISTORY usando atributos de Teufel (1999)

HISTORY	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	8	**SIM**	0.111	0.500	0.182	0.907
Falso	1	87	**NÃO**	0.989	0.916	0.951	

Tabela 60: Resultados do *macroaspecto* HISTORY usando atributos léxico-semânticos

Na Tab. 61, apresentam-se os resultados usando regras manuais sobre um conjunto de 29 sentenças anotadas manualmente com os aspecto HISTORY. Observa-se que a classe "não" teve melhores resultados do que a classe "sim" por uma grande diferença. Para a classe "sim", a cobertura foi relativamente baixa (0.414), enquanto a precisão foi alta (0.750). Cabe ressaltar que a acurácia também foi bastante alta (0.935). Os resultados mostram que é possível identificar HISTORY usando regras manuais (ver Fig. 27).

HISTORY	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	12	17	**SIM**	0.414	0.750	0.533	0.935
Falso	4	289	**NÃO**	0.986	0.944	0.965	

Tabela 61: Resultados do *macroaspecto* HISTORY usando regras manuais

5.3.8 PREDICTION

Na Tab. 62, apresentam-se os resultados usando regras manuais sobre um conjunto de 17 sentenças anotadas manualmente com os aspecto PREDICTION. Observa-se que a classe "não" teve melhores resultados do que a classe "sim". Para a classe "sim", a cobertura foi alta (0.765), enquanto a precisão foi baixa (0.333). Cabe ressaltar que existe uma grande quantidade de sentenças que deveriam ter sido anotadas manualmente, assim, muitos "falsos positivos" surgiram, ocasionando uma baixa precisão. Ressalta-se, também, o bom desempenho da acurácia (0.907). Os resultados mostram que pode ser

factível utilizar regras para identificar PREDICTION (ver Fig. 26). Também é preciso dizer que foi fácil de se criar as regras por causa da pouca quantidade de instâncias anotadas, podendo, novamente, gerar *overfitting*.

PREDICTION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	13	4	**SIM**	0.765	0.333	0.464	0.907
Falso	26	279	**NÃO**	0.915	0.986	0.949	

Tabela 62: Resultados do *macroaspecto* PREDICTION usando regras manuais

5.4 Medida de Avaliação de Sumários

Neste trabalho de pesquisa avalia-se a informatividade dos sumários automáticos. A informatividade dos sumários é avaliada em relação a um sumário de referência feito por humanos. Aquele sumário de referência normalmente é composto por sentenças dos textos-fonte que, conforme o critério linguístico humano, são consideradas essenciais para compor o sumário final.

Como já se mencionou anteriormente, a ferramenta que mede a informatividade dos sumários automáticos é a ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), proposta por Lin (2004). Esta ferramenta faz uma avaliação próxima ao julgamento humano, por isso, é a mais comumente usada para avaliar sumários automáticos. Basicamente, a medida ROUGE computa a coocorrência de n-gramas entre o sumário automático e um ou mais sumários de referência humanos. Esses n-gramas são considerados como sequências de 1 a 4 palavras. Por exemplo, a ROUGE-2 calcula a coocorrência de 2-gramas ou bigramas. Neste trabalho, as medidas a serem utilizadas são ROUGE-1, ROUGE-2 e ROUGE-L[1].

Os resultados da ROUGE são dados em termos de precisão, cobertura e média harmônica (também chamada de medida F) em relação ao sumário de referência. A seguir, apresentam-se as fórmulas:

[1]Baseado no problema *Longest Common Subsequence* (LCS), em que se faz uma comparação da maior subsequência de palavras em comum entre os sumários humanos e automáticos.

$$Precisão = \frac{Número\ de\ n\text{-}gramas\ em\ comum\ com\ o\ sumário\ de\ referência}{Número\ de\ n\text{-}gramas\ do\ sumário\ automático} \quad (20)$$

$$Cobertura = \frac{Número\ de\ n\text{-}gramas\ em\ comum\ com\ o\ sumário\ de\ referência}{Número\ de\ n\text{-}gramas\ do\ sumário\ de\ referência} \quad (21)$$

$$F = \frac{2 * Presição * Cobertura}{Presição + Cobertura} \quad (22)$$

É importante dizer que existem fatores que não podem ser avaliados pela ROUGE como a coerência e a coesão. Para poder avaliar esses fatores, deve se elaborar um cenário de avaliação humana conformado por juízes competentes da língua. Neste trabalho, avaliou-se somente a informatividade dos sumários, portanto, utilizou-se a ROUGE.

5.5 Avaliação dos Métodos de Sumarização

Nesta seção, relata-se a avaliação dos métodos de SA multidocumento propostos por meio das medidas da ROUGE (Lin, 2004). Além disso, os métodos propostos (*ASumm_n* e *ASumm_OP*) são comparados com dois dos melhores sumarizadores para o Português: o RSumm (Ribaldo et al., 2012) da abordagem superficial e o RC4 (Cardoso, 2014) da abordagem profunda.

O córpus utilizado para avaliar os sumários gerados foi o CSTNews (Cardoso et al., 2011). Como já foi explicado na Seção 3.1, o córpus contêm 50 coleções de textos jornalísticos sobre um mesmo assunto. Originalmente o córpus disponibilizava somente de 1 sumário extrativo e 1 *abstract* para cada coleção de documentos. Dias et al. (2014), enriqueceram o córpus criando mais 5 novos sumários extrativos e mais 5 novos *abstracts* para cada coleção, formando um total de 6 sumários extrativos e 6 *abstracts*.

Já que os métodos propostos foram criados com base nos padrões identificados nas quatro categorias principais do córpus, a avaliação foi feita sobre 48 sumários extrativos automáticos: "Cotidiano" (14), "Esportes" (10), "Mundo" (14) e "Política" (10). Não foram considerados os sumários das categorias "Dinheiro" (1) e "Ciência" (1).

Assim, avaliou-se a medida ROUGE dos 48 sumários extrativos automáticos com todos os sumários extrativos de referência. Além disso, avaliou-se a medida ROUGE dos 48 sumários extrativos com todos os *abstracts* de referência. Ressalta-se que cada sumário está limitado a uma taxa de compressão de 70%, ou seja, o sumário deve conter 30% (aproximadamente) do total de palavras do maior texto-fonte/documento.

Na Tab. 63, apresentam-se os valores da medida ROUGE-1 (unigramas), ROUGE-2 (bigramas) e ROUGE-L (maior subsequência de palavras em comum) para todos os sumários extrativos de referência do córpus CSTNews. As letras "C", "P" e "F" representam as métricas "Cobertura", "Precisão" e "medida F", respectivamente (ver Seção 5.4). O valor de cada métrica é calculado pela **média** da soma das medidas ROUGE entre o sumário gerado pelo método e os 6 sumários extrativos de referência, para cada coleção de documentos. Nas linhas 1 a 4, apresentam-se os resultados de quatro configurações do método *ASumm_n*: *ASumm_10* ($\alpha = 1.0$), *ASumm_7* ($\alpha = 0.7$), *ASumm_5* ($\alpha = 0.5$), e *ASumm_1* ($\alpha = 0.1$). Na linha 5, mostram-se os resultados do método *ASumm_OP*. Já na linha 6 e 7, mostram-se os resultados dos sumarizadores RSumm e RC4, respectivamente. Os melhores resultados estão ressaltados em **negrita**.

	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
Método	**C**	**P**	**F**	**C**	**P**	**F**	**C**	**P**	**F**
ASumm_10	0,4358	0,5153	0,4697	0,2843	0,3076	0,3029	0,4097	0,4833	0,4411
ASumm_7	0,4344	0,5195	0,4703	0,2802	0,3390	0,3048	0,4067	0,4865	0,4403
ASumm_5	0,4623	0,5376	0,4949	0,3064	0,3604	0,3295	0,4347	0,5055	0,4653
ASumm_1	**0,4768**	**0,5502**	**0,5083**	**0,3231**	**0,3760**	**0,3455**	**0,4508**	**0,5201**	**0,4805**
ASumm_OP	0,4326	0,5439	0,4759	0,2844	0,3607	0,3140	0,4048	0,5095	0,4456
RSumm	0,4777	0,5363	0,5030	0,3216	0,3627	0,3392	0,4504	0,5054	0,4741
RC4	**0,5147**	**0,5663**	**0,5370**	**0,3755**	**0,4166**	**0,3931**	**0,4923**	**0,5419**	**0,5137**

Tabela 63: Avaliação ROUGE dos sumários extrativos

Na configuração *ASumm_10*, tentou-se dar total importância à cobertura de aspectos (ou informatividade). Porém, o resultado foi o pior entre todos os métodos. Pode-se observar que os resultados melhoram conforme se dê mais importância à relevância do que à cobertura, caso das configurações *ASumm_7* e *ASumm_5*. A configuração *ASumm_1* obteve o melhor resultado para o método *ASumm_n* e foi o **segundo me-**

lhor método de sumarização, mostrando que a cobertura dos aspectos pode ajudar na formação de sumários informativos. Já o método *ASumm_OP* ficou na quarta posição, afirmando que os "padrões de ordem" gerados a partir da ordem em que ocorrem os aspectos no córpus não foram úteis na formação de sumários mais informativos. Por último, o método RC4 obteve os melhores resultados, superando o RSumm. Nota-se que os resultados da ROUGE-1 são os mais altos. Em geral, os resultados da ROUGE-1, ROUGE-2 e ROUGE-L são muito próximos.

Na Tab. 64 apresentam-se os resultados obtidos para todos os *abstracts* de referência do córpus CSTNews. De igual maneira que na avaliação anterior, observa-se que o pior método foi *ASumm_10* e o melhor método continua sendo o RC4. O método *ASumm_1* obteve o **segundo melhor resultado**. Diferentemente dos resultados da Tab. 63, os resultados da avaliação sobre os *abstracts* são um pouco mais baixos. Isso acontece por causa de que os *abstracts* são partes reescritas dos textos-fonte, gerando uma menor quantidade de ocorrências de n-gramas entre o sumário automático e o sumário de referência.

	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
Método	**C**	**P**	**F**	**C**	**P**	**F**	**C**	**P**	**F**
ASumm_10	0,3927	0,4379	0,4122	0,1872	0,2115	0,1977	0,3521	0,3916	0,3692
ASumm_7	0,3933	0,4442	0,4149	0,1883	0,2142	0,1993	0,3518	0,3960	0,3707
ASumm_5	0,4205	0,4587	0,4375	0,2062	0,2260	0,2149	0,3790	0,4136	0,3944
ASumm_1	**0,4296**	**0,4653**	**0,4452**	**0,2149**	**0,2334**	**0,2229**	**0,3871**	**0,4194**	**0,4012**
ASumm_OP	0,3971	0,4708	0,4255	0,1954	0,2333	0,2100	0,3556	0,4211	0,3810
RSumm	0,4296	0,4538	0,4398	0,2137	0,2254	0,2185	0,3886	0,4107	0,3979
RC4	**0,4511**	**0,4649**	**0,4563**	**0,2372**	**0,2442**	**0,2397**	**0,4102**	**0,4224**	**0,4148**

Tabela 64: Avaliação ROUGE dos *abstracts*

Salienta-se que o RC4 é um método de sumarização profundo que utiliza conhecimento semântico-discursivo fornecido por um córpus *gold standard* anotado manualmente com relações RST e CST. Já o método *ASumm_1* simula um ambiente real de sumarização desde a identificação dos aspectos (conhecimento semântico-discursivo) nas sentenças dos textos-fonte até a seleção das sentenças que formarão o sumário final, podendo, obviamente, existir ruído no processo (p.ex: sentenças não analisadas

pelo *parser* ou sentenças erroneamente anotadas com aspectos). É por esse motivo que os resultados do método RC4 foram superiores aos resultados do método *ASumm_1*.

Para verificar se há significância estatística entre os métodos, realizou-se o **Teste de Wilcoxon** (Søgaard et al., 2014) sobre os resultados da medida F1 para a ROUGE-1, ROUGE-2 e ROUGE-L. As hipóteses de pesquisa são:

- **H0:** A hipótese nula é de que não há diferença significativa entre o desempenho do método RC4 e o desempenho do método *ASumm_1*, tanto para os sumários extrativos quanto para os *abstracts*.
- **H1:** A hipótese alternativa é de que há diferença significativa entre os desempenhos.

É possível refutar a hipótese H0 se o *p-value* do teste for menor que 0,05. Isso nos dá 95% de significância estatística. Na Tab. 65, mostram-se os resultados obtidos tanto para os sumários extrativos quanto para os *abstracts*. Todos os resultados mostram que não há evidências para rejeitar a hipótese nula, a exceção do resultado da ROUGE-L para os sumários extrativos, em que se têm evidências de que há diferença estatisticamente significativa entre os dois métodos. Pode-se concluir, então, que os dois métodos obtiveram resultados bem similares mesmo o RC4 seja melhor que o *ASumm_1*.

	Extrativos	**Abstracts**
ROUGE-1	0.2036	0.2897
ROUGE-2	0.0923	0.1321
ROUGE-L	**0.0458**	0.0970

Tabela 65: Resultados do Teste de Wilcoxon

5.6 Considerações Finais

Neste capítulo, mostraram-se os resultados da avaliação do processo de identificação de aspectos informativos (*microaspectos* e *macraspectos*) e do processo de formação de sumários.

Na identificação de *microaspectos*, avaliaram-se dois tipos de abordagens: usando

sistemas (APS, APS+Regras e REMBRANDT) e usando técnicas de AM com atributos léxico-semânticos. As duas abordagens foram testadas sobre o mesmo córpus (CST-News). Os resultados da abordagem utilizando sistemas mostraram que o sistema APS+Regras foi o melhor para a maioria dos *microaspectos* (WHO_AGENT, WHEN, WHO_AFFECTED, WHERE, WHY e SCORE). Isso quer dizer claramente que as regras melhoraram o desempenho do sistema APS. Já o sistema APS só conseguiu o melhor resultado para o *microaspecto* HOW. Da maneira igual, o sistema REMBRANDT só obteve um resultado bom para o *microaspecto* SITUATION. É importante ressaltar os problemas identificados pelos sistemas no processo de identificação de *microaspectos*:

- Em algumas ocasiões, o sistema APS teve problemas ao não conseguir classificar alguns papéis semânticos ou ao classificar papéis de maneira errada, afetando o desempenho do sistema APS+Regras.

- Algumas sentenças não foram analisadas sintaticamente pelo *parser* PALAVRAS, consequentemente não foram anotadas com papéis semânticos, aumentando, assim, a quantidade de "falsos negativos".

- O sistema REMBRANDT só identificava entidades nomeadas escritas em caixa alta (a exceção das expressões temporais), causando um baixo desempenho do sistema. É por esse motivo que também não foi considerado para identificar WHO_AGENT e WHO_AFFECTED como pessoa/organização. Por exemplo, na segmento "o presidente disse que", a entidade "presidente" não seria identificada como WHO_AGENT pelo REMBDRANT.

Diferentemente da abordagem usando sistemas, a abordagem utilizando técnicas de AM foi testada com apenas 30% do córpus CSTNews. Assim, pode-se dizer que o baixo desempenho dos classificadores usando AM se deve à pouca quantidade de instâncias de treino e teste. Acredita-se que a existência de mais instâncias/sentenças no córpus possa melhorar os resultados dos classificadores de *microaspectos*.

Na identificação de *macroaspectos*, avaliaram-se duas abordagens: usando AM e usando regras manuais. A abordagem usando AM visa criar classificadores binários com base nos atributos definidos por Teufel (1999) e atributos léxico-semânticos. Já a

abordagem usando regras está baseada nos padrões linguísticos identificados sobre todas sentenças anotadas no córpus. As duas abordagens foram avaliadas sobre o córpus CSTNews. A avaliação da abordagem usando AM foi feita com apenas 30% do córpus CSTNews. O melhor resultado foi obtido pelo classificador do *macroaspecto* WHAT utilizando atributos léxico-semânticos, por ter um maior número de instâncias anotadas. Demoustrou-se que os atributos definidos por Teufel (1999) são mais apropriados para textos científicos do que para textos jornalísticos. Pode-se acrescentar que o baixo desempenho dos classificadores se deve a pouca quantidade de instâncias anotadas de treino e teste. De igual maneira que para os *microaspectos*, acredita-se que a existência de mais instâncias no córpus possa melhorar o desempenho dos classificadores.

Diferentemente da abordagem usando AM, a abordagem utilizando regras foi avaliada com o córpus anotado completo, ou seja, com um total de 322 sentenças. Os resultados obtidos para alguns *macroaspectos* (COMPARISON, DECLRATION, GOAL, HISTORY e PREDICTION) são razoáveis, provando que é possível identificar *macroaspectos* usando regras manuais.

Um dos grandes fatores pelo qual o desempenho das regras (tanto para *microaspectos* quanto para *macroaspectos*) não foram melhores é a anotação de aspectos do córpus CSTNews (Rassi et al., 2013). Pode-se perceber, em várias ocasiões, que as regras identificaram automaticamente sentenças que não foram anotadas manualmente (mas que deveriam ter sido anotadas), como aconteceu com WHERE (que não ocorre na categoria "Política") e PREDICTION (que não ocorre em todos os verbos no futuro), afetando o desempenho das regras gerando "falsos positivos".

Em suma, os resultados confirmam a primeira hipótese de que é possível identificar automaticamente aspectos informativos, assim como afirmar que existe um conjunto recorrente de aspectos para cada categoria textual específica (ver Apêndice B).

Na formação de sumários, avaliaram-se os dois métodos propostos (*ASumm_n* e *ASumm_OP*) e os dois métodos da literatura (RSumm e RC4). Os melhores resutlados foram obtidos pelo método RC4. Os segundos melhores resultados formam obtidos pelo método proposto *ASumm_1*. Salienta-se que os resultados entre os dois métodos são próximos, com a grande diferença de que o método *ASumm_1* representa um sistema completo de sumarização (*análise*, *transformação* e *sínteses*). Comprova-se, assim,

a segunda hipótese de que existe uma ou mais estruturas típicas de aspectos (ou *templates*) em sumários e existem métodos eficazes com base nessas estruturas para selecionar o conteúdo que produzirão o sumário final.

No capítulo seguinte, relatam-se as considerções finais gerais deste trabalho de pesquisa.

CAPÍTULO

6

Considerações Finais

Neste trabalho de mestrado, desenvolveu-se a primeira investigação de sumarização multidocumento com base em aspectos informativos. Implementaram-se as três etapas da arquitetura geral de um sistema de SA: *análise*, *transformação* e *síntese*.

Na etapa de *análise*, criaram-se vários classificadores de aspectos (*microaspectos* e *macroaspectos*, respectivamente) utilizando anotador de papéis semânticos, reconhecedor de entidades mencionadas, regras manuais e técnicas de AM com atributos variados. Avaliaram-se os classificadores sobre as sentenças anotadas do córpus CSTNews (Rassi et al., 2013; Felippo et al., 2014). Os resultados foram satisfatórios, demostrando que alguns aspectos podem ser identificados automaticamente em textos jornalísticos com um desempenho razoável. Comprova-se, assim, a primeira hipótese deste trabalho.

Na etapa de *transformação*, em primeiro lugar, utilizou-se o sistema RSumm (Ribaldo et al., 2012) para ranquear as sentenças por relevância e remover a redundância. Salienta-se que as sentenças só foram ordenadas por relevância, descartando-se o método de Lima e Pardo (2011, 2012). Em segundo lugar, elaboraram-se dois métodos inovadores para selecionar e reordenar as sentenças por informatividade: *ASumm_n* e *ASumm_OP*. Desta forma, realiza-se um novo ranqueamento com base tanto na relevância quanto na informatividade das sentenças.

Na etapa de *síntese*, forma-se o sumário final em formato de parágrafo por meio de uma justaposição de sentenças. Avaliou-se a informatividade dos sumários gerados em relação aos sumários humanos do córpus CSTNews através da medida ROUGE (Lin, 2004). Assim, compararam-se os métodos propostos com os melhores métodos de sumarização multidocumento para o Português: RSumm (Ribaldo et al., 2012) (da abordagem superficial) e RC4 (Cardoso, 2014) (da abordagem profunda). O método RC4 obteve o melhor resultado. Contudo, o método proposto *ASumm_1* obteve o segundo melhor resultado, superando ao método RSumm. Mediante o Teste de Wilcoxon (Søgaard et al., 2014), comprovou-se que entre o método *ASumm_1* e o método RC4 não existe diferença estatisticamente significativa. Comprova-se, assim, a segunda hipótese deste trabalho.

Neste capítulo, também se apresentam algumas contribuições e discutem-se as limitações encontradas. Por último, propõem-se alguns trabalhos futuros.

6.1 Contribuções

Além de ser a primeira investigação de sumarização multidocumento com base em aspectos informativos, este trabalho contribuiu com o desenvolvimento teórico e prático desta linha de pesquisa. São várias as contribuições desta pesquisa, entre as principais, citam-se:

- Desenvolveu-se um sistema (ou classificador) anotador de aspectos informativos, tanto para *microaspectos* quanto para *macroaspectos*. Dentro do processo de identificação, detalham-se os seguintes subprocessos:

  - Automatizou-se o processo de anotação de papéis semânticos, tendo-se como entrada as sentenças dos textos-fonte e dando como saída, as sentenças anotadas com papéis semânticos. Para isso utilizou-se o *parser* PALAVRAS (Bick, 2000), que fornece as árvores sintáticas das sentenças de entrada para que o classificador de Alva-Manchego (2013) possa anotar os papéis semânticos correspondentes de maneira automática. Por último, implementou-se um algoritmo para mapear os papéis com os *microaspectos* respectivos. A todo este

processo chamou-se de sistema APS.

  - A fim de melhorar o desempenho do sistema APS, criaram-se regras manuais que identificam quatro tipos de expressões: tempo, causa, local e placar (ou *score*). Cabe ressaltar que foi criado um algoritmo que identifica expressões temporais com base na teoria de (Baptista et al., 2008). Por outro lado, a fim de melhorar o desempenho dos classificadores de *macroaspectos*, criaram-se regras manuais que identificam expressões padrão, tipos de verbo e tempos verbais nas sentenças. Neste passo, utilizaram-se recursos como o repositório REPENTINO (Sarmento et al., 2006) e as etiquetas semânticas do PALAVRAS.

  - Utilizou-se e avaliou-se o sistema reconhecedor de entidades mencionadas REMBRANDT (Cardoso, 2008) para identificar alguns *microaspectos*.

  - Utilizaram-se técnicas de AM com base em vários atributos da literatura, com a finalidade de se criar classificadores de *microaspectos* e *macroaspectos*, respectivamente. Mesmo o resultado dos classificadores não seja satisfatório, realizou-se uma pesquisa exaustiva sobre esta abordagem que deve ser considerada no futuro.

- Elaboraram-se dois métodos inovadores com base em aspectos para selecionar e ordenar as sentenças que formarão o sumário final: *ASumm_n* e *ASumm_OP*. Ressalta-se que o método *ASumm_n* obteve resultados promissórios, superando, inclusive, ao melhor método da abordagem superficial para o Português do Brasil (RSumm).

- Por último, como resultado da pesquisa realizada neste trabalho de mestrado, em termos de publicações, tem-se, até o momento, um artigo publicado em um evento internacional como primeiro autor. Tal artigo foi selecionado para futura publicação na edição especial da revista *Research in Computing Science* (RCS)[1].

[1] http://rcs.cic.ipn.mx

6.2 Limitações

As principais limitações foram identificadas no processo de identificação de aspectos. Por exemplo, em algumas ocasiões, o sistema anotador de papéis semânticos de Alva-Manchego (2013) apresentava problemas ao classificar de maneira errada papéis semãnticos. O mesmo acontece com o *parser* PALAVRAS, ao não conseguir analisar sintaticamente as sentenças ou ao segmentar erroneamente um texto em sentenças. Já o sistema REMBRANDT só identificava entidades nomeadas escritas em caixa alta (a exceção das expressões temporais).

Afirma-se que o baixo desempenho dos classificadores (*microaspectos* e *macroaspectos*) se deve a pouca quantidade de instâncias anotadas de treino e teste no córpus CSTNews. Acredita-se que a existência de mais instâncias anotadas no córpus possa melhorar o desempenho dos classificadores de maneira considerável.

Como já foi comentado, um dos grandes fatores pelo qual o desempenho dos classificadores não foi melhor é a anotação de aspectos do córpus CSTNews (Rassi et al., 2013). Entende-se que a anotação de aspectos requere de um alto nível de conhecimento semântico, discursivo e do mundo por parte dos anotadores que para os computadores resulta difícil de modelar. Assim, por exemplo, os anotadores não identificaram algumas expressões de tempo e de local que o computador automaticamente identificou. Ressalta-se a ausência da anotação do aspecto WHERE na categoria "Política". Isso é um dos motivos principais que afeta o desempenho dos classificadores e gera "falsos positivos". Recomenda-se fazer uma anotação manual mais focada no que o computador possa identificar.

6.3 Trabalhos Futuros

Entre os principais trabalhos futuros que se desprendem desta pesquisa de mestrado, encontram-se:

- Anotação manual de aspectos sobre os textos-fonte/documentos do córpus CSTNews a fim de acrescentar as instâncias de treino e teste para os classificadores de aspectos.

- Análise dos aspectos anotados sobre os textos-fonte/documentos do córpus CST-News, para que, ao igual que Rassi et al. (2013), possam se identificar novos padrões aspectuais (ou *templates*) para formação de sumários.

- Testar os métodos propostos (*ASumm_n* e *ASumm_OP*) sobre o novo conjunto de sentenças anotadas com aspectos.

- Aprimorar os métodos desenvolvidos tanto no processo de identificação (APS, REM, regras, AM) quanto no processo de formação de sumários (*ASumm_n* e *ASumm_OP*).

- Desenvolver novos métodos de sumarização com base na análise dos *templates* identificados na nova anotação sobre os textos-fontes.

Referências Bibliográficas

Afantenos, S., V. Karkaletsis, P. Stamatopoulos, and C. Halatsis (2008). Using Synchronic and Diachronic Relations for Summarizing Multiple Documents Describing Evolving Events. *Journal of Intelligent Information Systems 30*(3), 183–226.

Aleixo, P. and T. Pardo (2008). CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discuriva Multidocumento CST. Technical Report NILC-TR-08-05, Série de Relatórios do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP).

Aluísio, S. M. and O. N. Oliveira Jr. (1996). A Detailed Schematic Structure of Research Papers Introductions: An Application in Support-Writing Tools. *Revista de la Sociedad Espanhola para el Procesamiento del Lenguaje Natural 19*, 141–147.

Alva-Manchego, F. (2013). Anotação Automática Semissupervisionada de Papéis Semânticos para o Português do Brasil. Dissertação, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Baptista, J., C. Hagège, and N. Mamede (2008). Identificação, Classificação e Normalização de Expressões Temporais do Português: a Experiência do Segundo HAREM e o Futuro. In M. Cristina and D. Santos (Eds.), *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*, pp. 35–54. Linguateca.

Barrera, A. and R. Verma (2011). Automated Extractive Single-document Summariza-

tion: Beating the Baselines with a New Approach. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, New York, NY, USA, pp. 268–269. ACM.

Barrera, A., R. Verma, and R. Vicent (2011). SemQuest: University of Houston's Semantics-based Question Answering System. In *Proceedings of the 4th Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, pp. 1–8. National Institute of Standards and Technology.

Baum, L. E. and T. Petrie (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics 37*, 1554–1563.

Bick, E. (2000). *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Arhus, Denmark: University of Arhus.

Bick, E. (2007). Functional Aspects on Portuguese NER. In D. Santos and N. Cardoso (Eds.), *Reconhecimento de Entidades Mencionadas em Português: Documentação e Actas do HAREM, a Primeira Avaliação Conjunta na Área*, pp. 145–155. Springer.

Bing, L., P. Li, Y. Liao, W. Lam, W. Guo, and R. J. Passonneau (2015). Abstractive Multi-Document Summarization via Phrase Selection and Merging. *Computing Research Repository (CoRR)*, 1–11.

Boutell, M. R., J. Luo, X. Shen, and C. M. Brown (2004). Learning Multi-label Scene Classification. *Pattern Recognition 37*(9), 1757–1771.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Monterey, CA, USA: Wadsworth and Brooks.

Bruckschen, M., F. Muniz, J. Souza, J. Fuchs, K. Infante, M. Muniz, P. Gonçalves, R. Vieira, and S. Aluísio (2008). Anotação Linguística em XML do Corpus PLN-BR. Technical Report NILC–TR–09–08, University of São Paulo, Brazil.

Carbonell, J. and J. Goldstein (1998). The use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, New York, NY, USA, pp. 335–336. ACM.

Cardoso, N. (2008). REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In M. Cristina and D. Santos (Eds.), *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*, pp. 195–211. Linguateca.

Cardoso, P. (2014). *Exploração de Métodos de Sumarização Automática Multidocumento com base em Conhecimento semântico-discursivo*. Ph. D. thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP), São Carlos, SP, Brasil.

Cardoso, P., E. Maziero, M. Castro Jorge, E. Seno, A. Di Felippo, L. Rino, M. Nunes, and T. Pardo (2011). A Discourse Annotated Corpus for Single and Multi-document Summarization of News Texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, Cuiabá/MT, Brazil, pp. 88–105.

Cardoso, P., T. Pardo, and M. Nunes (2011). Métodos para Sumarização Automática Multidocumento usando Modelos Semântico-Discursivos. In *Proceedings of the 3rd RST Brazilian Meeting*, Cuiabá, MT, Brazil, pp. 59–74.

Castro Jorge, M. (2010). Sumarização Automática Multidocumento: Seleção de Conteúdo com base no Modelo CST (Cross-Document Structure Theory). Disertação de mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Castro Jorge, M. (2015). *Modelagem Gerativa para Sumarização Automática Multidocumento*. Ph. D. thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP, São Carlos, SP, Brazil.

Castro Jorge, M. and T. Pardo (2010). Experiments with CST-based Multidocument Summarization. In *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, Uppsala, Sweden, pp. 74–82.

Castro Jorge, M. and T. Pardo (2011). A Generative Approach for Multi-document Summarization using the Noisy Channel Model. In *Proceedings of the 3rd RST Brazilian Meeting*, Cuiabá, MT, Brazil, pp. 75–87.

Celikyilmaz, A. and D. Hakkani-Tür (2011). Discovery of Topically Coherent Sentences for Extractive Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Stroudsburg, PA, USA, pp. 491–499. Association for Computational Linguistics.

Clarke, J. and M. Lapata (2010). Discourse Constraints for Document Compression. *Computational Linguistics 36*(3), 411–441.

Collobert, R. and J. Weston (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML-2008)*, pp. 160–167. ACM.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research 12*, 2493–2537.

Conroy, J. and D. O'leary (2001). Text Summarization via Hidden Markov Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, New York, NY, USA, pp. 406–407. ACM.

Contractor, D., Y. Guo, and A. Korhonen (2012). Using Argumentative Zones for Extractive Summarization of Scientific Articles. In *Proceedings of COLING*, Mumbai, India, pp. 663–678.

Dayrell, C., A. C. Jr., G. Lima, D. M. Jr., A. Copestake, V. Feltrim, S. Tagnin, and S. Aluisio (2012). Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Dias, M., A. Bokan, C. Chuman, C. Barros, E. Maziero, F. Nobrega, J. Souza, M. Sobrevilla, M. Delege, C. M., N. Silva, P. Cardoso, P. Balage, L. R., V. Marcasso, A. Di Felippo, M. Nunes, and T. Pardo (2014). Enriquecendo o Corpus CSTNews - a Criacao de Novos Sumarios Multidocumento. In *Proceedings of the I Workshop on Tools and*

Resources for Automatically Processing Portuguese and Spanish - ToRPorEsp, São Carlos, SP, Brazil, pp. 1–8.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the ACM (JACM) 16*(2), 264–285.

Felippo, A., L. Rino, T. Pardo, P. Cardoso, E. Seno, P. Balage Filho, A. Rassi, M. Dias, M. Jorge, E. Maziero, A. Zacarias, J. Souza, R. Camargo, and V. Agostini (2014). Corpus Annotation of Textual Aspects in Multi-Document Summaries. In S. M. Aluísio and S. E. O. Tagnin (Eds.), *New Language Technologies and Linguistic Research: A Two-Way Road*, Chapter 10, pp. 171–192. Cambridge Scholars Publishing.

Fellbaum, C. (1998). *WordNet: an Electronic Lexical Database*. MIT Press.

Feltrim, V. (2004). *Uma Abordagem Baseada em Córpus e em Sistemas de Crítica para a Construção de Ambientes Web de Auxílio à Escrita Acadêmica em Português*. Tese de doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP).

Feltrim, V. D., S. Teufel, M. G. V. Nunes, and S. Aluísio (2006). Argumentative Zoning Applied to Critiquing Novices Scientific Abstracts. In *Computing Attitude and Affect in Text: Theory and Applications*, Number 20 in The Information Retrieval Series, pp. 233–246. Springer.

Fillmore, C. J. (1968). The Case for Csase. In E. Bach and R. T. Harms (Eds.), *Universals in Linguistic Theory*, pp. 0–88. New York: Holt, Rinehart and Winston.

Fonseca, E. R. (2013). Uma Abordagem Conexionista para Anotação de Papéis Semânticos. Dissertação, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Gantz, J. and D. Reinsel (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. *International Data Corporation iView*.

Genest, P., G. Lapalme, and M. Yousfi-Monod (2009). HexTac: the Creation of a Manual Extractive Run. In *Proceedings of the 4th Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, pp. 14–15.

Genest, P.-E. and G. Lapalme (2012). Fully Abstractive Approach to Guided Summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, Stroudsburg, PA, USA, pp. 354–358. Association for Computational Linguistics.

Genoves Jr., L., V. Feltrim, C. Dayrell, and S. Aluisio (2007). Automatically Detecting Schematic Structure Components of English Abstracts. In *Proceedings of the RANLP 2007, Workshop on Natural Language Processing for Educational Resources*, Borovets, Bulgaria, pp. 23–29.

Gildea, D. and D. Jurafsky (2001). Identifying Semantic Roles in Text. In *17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, Seattle, Washington.

Gildea, D. and D. Jurafsky (2002). Automatic labeling of semantic roles. *Comput. Linguist. 28*(3), 245–288.

Grishman, R. and B. Sundheim (1996). Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, Stroudsburg, PA, USA, pp. 466–471. Association for Computational Linguistics.

Haghighi, A. and L. Vanderwende (2009). Exploring Content Models for Multi-document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, Stroudsburg, PA, USA, pp. 362–370. Association for Computational Linguistics.

Hartmann, N. (2015). Anotação Automática de Papéis Semânticos de Textos Jornalísticos e de Opinião sobre Árvores Sintáticas não Revisadas. Dissertação, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Hochbaum, D. S. (1997). *Approximation Algorithms for NP-hard Problems*. Boston, MA, USA: PWS Publishing Co.

Karlsson, F. (1990). Constraint Grammar as a Framework for Farsing Running Text. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING '90, Stroudsburg, PA, USA, pp. 168–173. Association for Computational Linguistics.

Landauer, T., P. Foltz, and D. Laham (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes 25*, 259–284.

Li, P., Y. Wang, W. Gao, and J. Jiang (2011). Generating aspect-oriented multi-document summarization with event-aspect model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Stroudsburg, PA, USA, pp. 1137–1146. Association for Computational Linguistics.

Lima, J. and T. Pardo (2011). Ordenação de Sentenças em Sumários Multidocumento: Uma Abordagem Utilizando Relações CST. In *Proceedings of the 2nd STIL Student Workshop on Information and Human Language Technology*, Cuiabá, MT, Brazil, pp. 1–3.

Lima, J. and T. Pardo (2012). Ordenação de Sentenças em Sumários Multidocumento. Technical Report NILC-TR-12-02, Série de Relatórios do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP).

Lin, C.-Y. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain.

Louis, A., A. Joshi, and A. Nenkova (2010). Discourse Indicators for Content Selection in Summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, Stroudsburg, PA, USA, pp. 147–156. Association for Computational Linguistics.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

Makino, T., H. Takamura, and M. Okumura (2011). Balanced Coverage of Aspects for Text Summarization. In *Proceedings of the 4th Text Analysis Conference (TAC 2011)*,

Gaithersburg, Maryland, USA, pp. 1–8. National Institute of Standards and Technology.

Makino, T., H. Takamura, and M. Okumura (2012). Balanced Coverage of Aspects for Text Summarization. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, New York, NY, USA, pp. 1742–1746. ACM.

Mani, I. (1999). *Advances in Automatic Text Summarization*. Cambridge, MA, USA: MIT Press.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company.

Mann, W. and S. Thompson (1987). Rhetorical Structure Theory: A Theory of Text Organization. *Reprinted from the Structure of Discourse, ISI Reprint Series*, 87–190.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press.

Maziero, E. and T. Pardo (2011). Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning. Cuiabá, MT, Brazil, pp. 1–10. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology.

McKeown, K., R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg (2005). Do Summaries Help? In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, New York, NY, USA, pp. 210–217. ACM.

McKeown, K. and D. Radev (1995). Generating Summaries of Multiple News Articles. In *Proceedings of 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Volume 1, Seattle, WA, pp. 74–82.

Mitchell, T. M. (1997). *Machine Learning* (1 ed.). New York, NY, USA: McGraw-Hill, Inc.

Nadeau, D. and S. Sekine (2007). A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes 30*(1), 3–26. Publisher: John Benjamins Publishing Company.

Nenkova, A. (2005a). Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, pp. 1436–1441. AAAI Press.

Nenkova, A. (2005b). Discourse Factors in Multi-document Summarization. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 4*, AAAI'05, pp. 1654–1655. AAAI Press.

Nenkova, A. and K. McKeown (2011). *Automatic Summarization*, Volume 5. Foundations and Trends in Information Retrieval.

Newman, C. B. D. and C. Merz (1998). *UCI Repository of Machine Learning Databases*. Irvine, CA, USA: Dept. of Information and Computer Sciences, University of California.

O'Donnell, M. (1997). Variable Length On-line Document Generation. In *Proceedings of the 6th European Workshop on Natural Language Generation, Gerhard-Mercator University*, Duisburg, Germany, pp. 1–5.

Otterbacher, J. C., D. R. Radev, and A. Luo (2002). Revisions that Improve Cohesion in Multi-document Summaries: a Preliminary Study. In *Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4*, AS '02, Stroudsburg, PA, USA, pp. 27–36. Association for Computational Linguistics.

Owczarzak, K. and H. Dang (2011). Who Wrote What Where: Analyzing the Content of Human and Automatic Summaries. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, Portland, Oregon, pp. 25–32. Association for Computational Linguistics.

Palmer, M., D. Gildea, and N. Xue (2010). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

Pardo, T. (2005, Fevereiro). GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades. Série de Relatórios do NILC NILC-TR-05-05, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Pardo, T. and L. Rino (2002). DMSumm: Review and Assessment. *E. Ranchhod and N. J. Mamede (Eds.), Advances in Natural Language Processing*, 263–273.

Platt, J. C. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft Research.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Radev, D. R. (2000). A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue - Volume 10*, SIGDIAL '00, Stroudsburg, PA, USA, pp. 74–83. Association for Computational Linguistics.

Radev, D. R. and K. R. McKeown (1998). Generating Natural Language Summaries from Multiple on-line Sources. *Computational Linguistics 24*(3), 470–500.

Rassi, A., A. Zacarias, E. Maziero, J. Souza, L. Castro, P. Balage, P. Cardoso, R. Camargo, V. Agostini, A. Filippo, E. Seno, L. Rino, and T. Pardo (2013). Anotação de Aspectos Textuais em Sumários do Córpus CSTNews. Technical Report NILC-TR-13-01, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP).

Read, J., B. Pfahringer, G. Holmes, and E. Frank (2011). Classifier Chains for Multi-label Classification. *Machine Learning 85*(3), 333–359.

Ribaldo, R., A. T. Akabane, L. H. M. Rino, and T. A. S. Pardo (2012). Graph-Based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In H. de Medeiros Caseli, A. Villavicencio, A. J. S. Teixeira, and F. Perdigão (Eds.), *PROPOR*, Volume 7243 of *Lecture Notes in Computer Science*, pp. 260–271. Springer.

Rino, L., T. Pardo, C. Silla Jr., C. Kaestner, and M. Pombo (2004). Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence – SBIA*, São Luis, Brazil, pp. 235–244.

Russell, S. J. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach* (2 ed.). Pearson Education.

Salton, G. (Ed.) (1988). *Automatic Text Processing*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Salton, G. and M. J. McGill (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Salton, G., A. Singhal, M. Mitra, and C. Buckley (1997). Automatic Text Structuring and Summarization. *Information Processing and Management: an International Journal - Special issue: methods and tools for the automatic construction of hypertext 33*(2), 193–207.

Santos, D. and N. Cardoso (2007). *Reconhecimento de Entidades Mencionadas em Português*. Documentação e Actas do HAREM, a Primeira Avaliação Conjunta na Área. Oslo/Lisboa, Portugual: Linguateca.

Sarmento, L., A. S. Pinto, and L. Cabral (2006). REPENTINO - A Wide-Scope Gazetteer for Entity Recognition in Portuguese. In R. Vieira, P. Quaresma, M. da Graça Volpes Nunes, N. J. Mamede, C. Oliveira, and M. C. Dias (Eds.), *Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, Volume LNAI 3960, pp. 31–40. Springer.

Shannon, C. E. (2001). A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review 5*(1), 3–55.

Søgaard, A., A. Johannsen, B. Plank, D. Hovy, and H. M. Alonso (2014). What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL*, Baltimore, Maryland, USA, pp. 1–10.

Souza, V. and V. Feltrim (2013). A Coherence Analysis Module for SciPo: Providing Suggestions for Scientific Abstracts Written in Portuguese. *Journal of the Brazilian Computer Society 19*(1), 59–73.

Stasko, J., C. Görg, and Z. Liu (2008). Jigsaw: Supporting Investigative Analysis Through Interactive Visualization. *Information Visualization 7*(2), 118–132.

Steinberger, J., H. Tanev, M. Kabadjov, and R. Steinberger (2010). JRC's Participation in the Guided Summarization Task at TAC 2010. In *Proceedings of the Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA, pp. 1–12. National Institute of Standards and Technology.

Svore, K. M. (2007). Enhancing Single-document Summarization by Combining RankNet and Third-party Sources. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 448–457.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge, UK: Cambridge University Press.

Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text*. Ph. D. thesis, University of Edinburgh, Edinburgh, Scotland.

Teufel, S. and M. Moens (1999). Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In *Advances in automatic Text Summarization*, pp. 155–171. MIT Press.

Teufel, S. and M. Moens (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics 28*(4), 409–445.

Tsoumakas, G. and I. Katakis (2007). Multi-label Classification: An Overview. *International Journal on Data Warehousing and Mining 3*(3), 1–13.

Uzêda, V., T. A. S. Pardo, and M. V. Nunes (2010). A comprehensive comparative evaluation of rst-based summarization methods. *ACM Trans. Speech Lang. Process. 6*(4), 4:1–4:20.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Wan, X. (2008). An Exploration of Document Impact on Graph-based Multi-document Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, Stroudsburg, PA, USA, pp. 755–762. Association for Computational Linguistics.

Wan, X. and J. Yang (2006). Improved Affinity Graph based Multi-document Summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, Stroudsburg, PA, USA, pp. 181–184. Association for Computational Linguistics.

White, M., T. Korelsky, C. Cardie, V. Ng, D. Pierce, and K. Wagstaff (2001). Multidocument Summarization via Information Extraction. In *Proceedings of the 1st International Conference on Human Language Technology Research*, HLT '01, Stroudsburg, PA, USA, pp. 1–7. Association for Computational Linguistics.

Zhang, R., Y. Ouyang, and W. Li (2011). Guided Summarization with Aspect Recognition. In *Proceedings of the 4th Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, pp. 1–6. National Institute of Standards and Technology.

Zhang, Z., S. Blair-Goldensohn, and D. R. Radev (2002). Towards CST-enhanced summarization. In *Proceedings of 18th National Conference on Artificial Intelligence*, Menlo Park, CA, USA, pp. 439–445. American Association for Artificial Intelligence.

Zhou, L., M. Ticrea, and E. Hovy (2005). Multi-document Biography Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1–8.

APÊNDICE

A

Definição de Aspectos

Na Tab. 66, listam-se as definições dos aspectos identificados na anotação feita sobre o córpus CSTNews (Rassi et al., 2013), acompanhadas de exemplos prototípicos para cada categoria do córpus. Ilustra-se um exemplo para cada categoria principal do córpus CSTNews: "Cotidiano", "Esportes", "Mundo" e "Política". As categorias não ilustradas, que apresentam a marca "—", indicam a não ocorrência de casos no córpus para o aspecto correspondente.

Aspectos	**Definição e exemplo**
WHO_AGENT	*A entidade (pessoa ou organização) responsável por causar/provocar a ocorrência de um fato/evento.*
	Cotidiano: O Ministério Público Federal apreendeu nesta terça-feira, 7, os registros dos últimos cinco anos do livro de ocorrências da torre de controle do Aeroporto de Congonhas, zona sul de São Paulo, durante um mandado de busca e apreensão.
	Esporte: A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.

	Mundo: Um atirador matou ao menos 30 pessoas em dois diferentes locais da Universidade Técnica da Virgínia, em Blacksburg (Virgínia), nesta segunda-feira, no pior ataque a tiros contra um campus universitário da história dos Estados Unidos.
	Política: O ministro da Fazenda, Guido Mantega, apresentou nesta terça-feira a proposta do governo em troca do apoio do PSDB na votação da PEC (Proposta de Emenda Constitucional) que prorroga a cobrança da CPMF até 2011.
WHO_AFFECTED	*A entidade (pessoa ou organização) que sofre os efeitos de um fato/evento.*
	Cotidiano: Depois que os presos entregaram o revólver usado para dar início ao motim, a Tropa de Choque da Polícia Militar entrou no presídio e liberou os 30 reféns - sendo 16 crianças.
	Esporte: A ginasta Jade Barbosa foi escolhida em votação na Internet, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim.
	Mundo: 17 pessoas morreram após a queda de um avião na República Democrática do Congo.
	Política: Na segunda representação, Renan é acusado de trabalhar para reverter dívida de R$100 milhões da Schincariol junto ao INSS.
WHEN	*A data/período de tempo (estritamente temporal) de ocorrência de um fato/evento.*
	Cotidiano: Um homem suspeito de ter roubado o relógio Rolex do apresentador de televisão Luciano Huck foi detido na quarta-feira, 16, em Taboão da Serra, na Grande São Paulo.
	Esporte: A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos
	Mundo: Antes de chegar à Jamaica, Dean matou ao menos nove pessoas nas ilhas de Santa Lúcia, Dominica, República Dominicana e Haiti, no Caribe.
	Política: O senador João Pedro (PT-AM), relator da segunda representação contra Renan Calheiros (PMDB-AL) no Senado, confirmou que vai apresentar nesta quarta-feira, 26, na reunião do Conselho de Ética, pedido de sobrestamento das investigações sobre o caso.
WHERE	*A localização geográfica ou física de um fato/evento.*

	Cotidiano: Uma nova série de ataques criminosos foi registrada na madrugada desta segunda-feira, dia 7, em São Paulo e municípios do interior paulista.
	Esporte: A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.
	Mundo: Um acidente envolvendo dois trens, ao norte do Cairo, deixou por volta de 80 mortos e 165 feridos, segundo fontes policiais e médicas.
	Política: Na sexta-feira, em encontro com sindicalistas em São Paulo, Lula disse que venceria a eleição no primeiro turno - tendência apontada pelas pesquisas.
WHY	*Uma explicação do porquê um fato/evento acontece (ou aconteceu).*
	Cotidiano: O crescimento nas autuações de contribuintes que caíram na malha fina se deu porque os auditores passaram a contar com programas mais modernos de computadores que analisam todas as irregularidades fiscais dos contribuintes, inclusive de anos anteriores, e não mais por grupos de infrações.
	Esporte: Maradona voltou a ter problemas de saúde no fim de semana e foi internado novamente em um hospital em Buenos Aires.
	Mundo: O avião saiu de Lugushwa a Bukavu e caiu sobre uma floresta após se chocar com uma montanha, prejudicado pelo mau tempo.
	Política: Renan é alvo de um processo por quebra de decoro acusado de receber recursos da construtora Mendes Junior para pagamento de despesas pessoais, como aluguel e pensão para a jornalista Mônica Veloso, com quem tem uma filha.
HOW	*O modo como um fato/evento ocorre.*
	Cotidiano: Depois que os presos entregaram o revólver usado para dar início ao motim, a Tropa de Choque da Polícia Militar entrou no presídio e liberou os 30 reféns - sendo 16 crianças.
	Esporte: Fabiana conseguiu o ouro em três tentativas.
	Mundo: —
	Política: Em cada um dos turnos, precisa de 308 votos favoráveis.
SCORE	*O resultado numérico de um fato/evento (score, tempo, distância, etc., sobretudo relativo a esportes).*
	Cotidiano: —

	Esporte: A seleção brasileira, sob direção de Dunga, conquistou o oitavo título da Copa América, goleando a Argentina por 3 a 0.
	Mundo: —
	Política: —
COMMENT	*Um comentário do autor sobre um fato/evento.*
	Cotidiano: O presidente deu grande ênfase ao fim do protecionismo agrícola, que enriquece os ricos e empobrece os pobres.
	Esporte: Neste domingo, o esporte brasileiro alegrou a torcida verde-amarelo.
	Mundo: —
	Política: —
COMPARISON	*Dados ou estatísticas diferentes comparando duas ou mais entidades.*
	Cotidiano: Foram autuados 208.471 contribuintes, em crescimento de 104,47% em relação ao mesmo período do ano passado.
	Esporte: —
	Mundo: —
	Política: Quando se compara com uma pesquisa sem a lista oficial dos candidatos, Lula sobe de 27% para 31%, Geraldo de 4% a 14% e Heloisa de 1% a 6%.
CONSEQUENCE	*Um fato/evento causado por outro fato/evento.*
	Cotidiano: A Secretaria da Fazenda também foi atingida por uma bomba.
	Esporte: A brasileira Fabiana Murer conquistou a medalha de ouro no salto com vara ao saltar 4m60, um novo recorde pan-americano, 20cm a mais que sua antiga marca.
	Mundo: Mundo: O furacão Dean passou pela costa sul da Jamaica, inundando a capital e espalhando árvores e telhados.
	Política: Em alguns casos, os parlamentares estão sendo abandonados pelos partidos, especialmente por ser ano eleitoral
COUNTERMEASURES	*Medidas que visam solucionar/antecipar/impedir problemas relacionados a um fato/evento.*
	Cotidiano: Segundo informações do jornalista Ricardo Noblat, o presidente Luiz Inácio Lula da Silva mandou a FAB colocar dois aviões à disposição da família do senador.

	Esporte: Para evitar o segundo cartão amarelo, o treinador fez a substituição do jogador que estava muito nervoso. (exemplo que não consta do CSTNews)
	Mundo: Foi decretado estado de emergência preventiva no local.
	Política: Entretanto, a oposição passará o dia tentando obstruir os trabalhos em plenário com o único objetivo de retardar a votação e dificultar a tarefa governista.
DECLARATION	*Um discurso ou fala de alguém ou de uma fonte por citação direta ou indireta.*
	Cotidiano: Segundo um informante da delegacia, os dois teriam vendido o acessório de luxo avaliado em cerca de R$10 mil.
	Esporte: O Barcelona jogou o Mundial para valer no ano passado. Nós faremos o mesmo, disse o meia, após a partida em Stamford Bridge. (exemplo que não consta do CSTNews)
	Mundo: Segundo o jornal Choson Sinbo, mais de 7 mil casas foram destruídas ou danificadas, e quase 16 mil hectares de terra cultivada foram inundados.
	Política: Eu não moverei uma palha contra eles [oposição] porque vocês moverão um paiol inteiro, afirmou o presidente Luiz Inácio Lula da Silva, candidato à reeleição pelo PT, sobre os ataques de seus adversários.
GOAL	*Finalidade/razão para um fato/evento que irá acontecer.*
	Cotidiano: O objetivo das buscas é garantir a apreensão dos registros de ocorrências que contêm informações sobre as falhas no controle de tráfego aéreo.
	Esporte: Boca entra em campo para ganhar após 5 partidas. (exemplo que não consta do CSTNews)
	Mundo: A Operação Farrapos, da Polícia Federal, com o objetivo de desarticular uma quadrilha internacional de tráfico de drogas, prendeu 14 dos 17 suspeitos, após 2 anos de investigações.
	Política: Entretanto, a oposição passará o dia tentando obstruir os trabalhos em plenário com o único objetivo de retardar a votação e dificultar a tarefa governista.
HISTORY	*Informação de contexto sobre uma história/um passado relacionado ao fato/evento.*

	Cotidiano: ACM já tinha sofrido infarto em 1989 e já tinha recebido três pontes de safena.
	Esporte: A equipe brasileira já conquistou cinco vezes a Liga Mundial.
	Mundo: Este foi o maior acidente ferroviário egípcio desde 2002, após o incêndio de um trem que deixou 376 mortos.
	Política: Esta pesquisa foi a primeira da série CNI/Ibope com a lista oficial dos candidatos à Presidência, fornecido pelo TSE.
PREDICTION	*Informação sobre a factibilidade de fatos/eventos futuros (podendo, inclusive, ser um evento com ocorrência certa).*
	Cotidiano: Esse trabalho permitirá avaliar os riscos aos quais estão expostos os passageiros e tripulantes de aeronaves e tomar medidas necessárias para aumentar a segurança no setor aéreo.
	Esporte: O próximo confronto será contra os rivais mais perigosos, a seleção de Cuba.
	Mundo: Na sexta-feira, choveu muito acima do esperado e há previsão de mais tempestades hoje.
	Política: Com 2 pontos percentuais para mais e para menos, os resultados assegurariam vitória de Lula no primeiro turno.
SITUATION	*Uma ocasião em que ocorreu um fato/evento. Envolve uma transação, um campeonato, um compromisso ou outros tipos de situação em uma data ou local inespecíficos.*
	Cotidiano: O presidente Luiz Inácio Lula da Silva afirmou nesta segunda-feira, durante o programa de rádio "Café com o Presidente", que vai anunciar obras de infra-estrutura e saneamento que transformarão o Brasil em um "verdadeiro canteiro de obras".
	Esporte: A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos.
	Mundo: Nesta batalha, 15 soldados israelenses morreram ao serem atingidos por um míssel.
	Política: Na sexta-feira, em encontro com sindicalistas em São Paulo, Lula disse que venceria a eleição no primeiro turno - tendência apontada pelas pesquisas.
WHAT	*Um fato/evento descrito no texto.*

	Cotidiano: Após quase 24 horas de tensão, terminou no fim da manhã desta quarta-feira a rebelião na Central de Custódia de Presos de Justiça (CCJP) no Maranhão.
	Esporte: A brasileira Fabiana Murer conquistou a medalha de ouro no salto com vara ao saltar 4m60, um novo recorde pan-americano, 20 cm a mais que sua antiga marca.
	Mundo: 17 pessoas morreram após a queda de um avião na República Democrática do Congo.
	Política: Ocorre hoje a votação da PEC (Proposta de Emenda Constitucional) que prorroga a cobrança da CPMF (Contribuição Provisória sobre Movimentação Financeira) até 2011 e mantém a alíquota de 0,38%.

Tabela 66: Aspectos do córpus CSTNews (Rassi et al., 2013)

APÊNDICE

B

Aspectos por Categoria

Nas tabelas deste apêndice, são listados os aspectos (*microaspectos* e *macroaspectos*) definidos por Rassi et al. (2013) para as quatro categorias principais do córpus CSTNews: “Cotidiano”, “Esportes”, “Mundo” e “Política”.

Macroaspectos	**Microaspectos**
COMMENT	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
COUNTERMEASURES	WHERE
DECLARATION	WHY
GOAL	HOW
HISTORY	
PREDICTION	
SITUATION	
WHAT	

Tabela 67: Aspectos definidos para a categoria “Cotidiano”

Macroaspectos	**Microaspectos**
COMMENT	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
DECLARATION	WHERE
GOAL	WHY
HISTORY	SCORE
PREDICTION	SITUATION
WHAT	
HOW	

Tabela 68: Aspectos definidos para a categoria “Esportes”

Macroaspectos	**Microaspectos**
CONSEQUENCE	WHO_AGENT
DECLARATION	WHO_AFFECTED
COUNTERMEASURES	WHEN
HISTORY	WHERE
PREDICTION	WHY
WHAT	GOAL
	SITUATION

Tabela 69: Aspectos definidos para a categoria “Mundo”

Macroaspectos	**Microaspectos**
COUNTERMEASURES	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
DECLARATION	WHERE
GOAL	WHY
HISTORY	HOW
PREDICTION	
WHAT	
SITUATION	

Tabela 70: Aspectos definidos para a categoria “Política”

APÊNDICE

C

Resultados dos Classificadores usando Atributos Léxico-Semânticos

Nas Tabs. 71 e 72, apresentam-se os resultados da classe “sim” dos classificadores de *microaspectos* e *macroaspectos* usando atributos léxico-semânticos. O desvio padrão é posicionado do lado do direito de cada valor com a marca “+/-”.

Microaspecto	**Classificador**	**Cobertura**	**Precisão**	**F1**	**Acurácia**
WHO_AGENT	(1, 1) bag_of_words	0.436+/-0.12	0.81+/-0.12	0.567+/-0.12	0.732+/-0.06
	(2, 2) bag_of_words	0.051+/-0.06	1.0+/-0.76	0.098+/-0.1	0.619+/-0.02
	(2, 3) bag_of_words	0.051+/-0.06	1.0+/-0.84	0.098+/-0.11	0.619+/-0.02
	(1, 1) lemmas	0.462+/-0.15	0.818+/-0.14	0.59+/-0.12	0.742+/-0.06
	(2, 2) lemmas	0.051+/-0.06	1.0+/-0.76	0.098+/-0.1	0.619+/-0.02
	(2, 3) lemmas	0.051+/-0.06	1.0+/-0.84	0.098+/-0.11	0.619+/-0.02
	(1, 1) POS	0.487+/-0.10	0.655+/-0.1	0.559+/-0.09	0.691+/-0.05
	(2, 2) POS	0.487+/-0.13	0.613+/-0.11	0.543+/-0.09	0.67+/-0.07
	(2, 3) POS	0.462+/-0.14	0.581+/-0.14	0.514+/-0.13	0.649+/-0.09
	(1, 1) semantic	0.513+/-0.19	0.606+/-0.1	0.556+/-0.13	0.67+/-0.07
	(2, 2) semantic	0.436+/-0.12	0.68+/-0.16	0.531+/-0.07	0.691+/-0.05
	(2, 3) semantic	0.436+/-0.16	0.654+/-0.12	0.523+/-0.13	0.68+/-0.06
	(1, 1) lemmas+POS	0.410+/-0.18	0.8+/-0.17	0.542+/-0.14	0.722+/-0.06
	(2, 2) lemmas+POS	0.051+/-0.06	1.0+/-0.76	0.098+/-0.1	0.619+/-0.02

	(2, 3) lemmas+POS	0.051+/-0.06	1.0+/-0.84	0.098+/-0.11	0.619+/-0.02
	(1, 1) POS+semantic	0.462+/-0.13	0.621+/-0.09	0.529+/-0.11	0.67+/-0.06
	(2, 2) POS+semantic	0.487+/-0.14	0.633+/-0.1	0.551+/-0.1	0.68+/-0.06
	(2, 3) POS+semantic	**0.538+/-0.15**	**0.636+/-0.08**	**0.583+/-0.11**	**0.691+/-0.06**
WHO_AFFECTED	(1, 1) bag_of_words	0.167+/-0.1	0.75+/-0.4	0.273+/-0.16	0.835+/-0.03
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(1, 1) lemmas	**0.222+/-0.16**	**1.0+/-0.3**	**0.364+/-0.22**	**0.854+/-0.03**
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(1, 1) POS	0.056+/-0.07	0.5+/-0.48	0.1+/-0.12	0.814+/-0.01
	(2, 2) POS	0.111+/-0.11	0.333+/-0.35	0.167+/-0.16	0.794+/-0.04
	(2, 3) POS	0.222+/-0.17	0.4+/-0.23	0.286+/-0.17	0.794+/-0.05
	(1, 1) semantic	0.278+/-0.22	0.5+/-0.15	0.357+/-0.18	0.814+/-0.04
	(2, 2) semantic	0.111+/-0.12	0.5+/-0.4	0.182+/-0.17	0.814+/-0.03
	(2, 3) semantic	0.111+/-0.18	0.4+/-0.51	0.174+/-0.26	0.804+/-0.05
	(1, 1) lemmas+POS	0.222+/-0.13	0.8+/-0.34	0.348+/-0.18	0.845+/-0.03
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(1, 1) POS+semantic	0.222+/-0.14	0.5+/-0.27	0.308+/-0.18	0.814+/-0.05
	(2, 2) POS+semantic	0.222+/-0.17	0.571+/-0.25	0.32+/-0.2	0.825+/-0.04
	(2, 3) POS+semantic	0.222+/-0.16	0.5+/-0.29	0.308+/-0.19	0.812+/-0.05
WHEN	(1, 1) bag_of_words	0.091+/-0.11	0.667+/-0.6	0.16+/-0.17	0.781+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.763+/-0.01
	(2, 3) bag_of_words	0.0+/-0.03	0.0+/-0.4	0.0+/-0.06	0.763+/-0.01
	(1, 1) lemmas	0.174+/-0.14	0.667+/-0.39	0.276+/-0.21	0.784+/-0.04
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.763+/-0.01
	(2, 3) lemmas	0.0+/-0.03	0.0+/-0.4	0.0+/-0.06	0.763+/-0.01
	(1, 1) POS	0.043+/-0.09	0.5+/-0.75	0.08+/-0.15	0.763+/-0.04
	(2, 2) POS	0.13+/-0.15	0.375+/-0.27	0.194+/-0.17	0.742+/-0.05
	(2, 3) POS	0.261+/-0.21	0.375+/-0.17	0.308+/-0.18	0.722+/-0.06
	(1, 1) semantic	**0.522+/-0.13**	**0.75+/-0.21**	**0.615+/-0.14**	**0.845+/-0.06**
	(2, 2) semantic	0.391+/-0.11	0.75+/-0.18	0.514+/-0.12	0.825+/-0.04
	(2, 3) semantic	0.435+/-0.1	0.769+/-0.14	0.556+/-0.1	0.835+/-0.03
	(1, 1) lemmas+POS	0.174+/-0.14	0.667+/-0.45	0.276+/-0.21	0.784+/-0.06
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.763+/-0.01
	(2, 3) lemmas+POS	0.0+/-0.03	0.0+/-0.4	0.0+/-0.06	0.763+/-0.01
	(1, 1) POS+semantic	0.348+/-0.25	0.727+/-0.2	0.471+/-0.24	0.814+/-0.06
	(2, 2) POS+semantic	0.478+/-0.06	0.688+/-0.18	0.564+/-0.08	0.825+/-0.04
	(2, 3) POS+semantic	0.478+/-0.16	0.611+/-0.13	0.537+/-0.14	0.804+/-0.05
WHERE	(1, 1) bag_of_words	0.118+/-0.13	1.0+/-0.0	0.211+/-0.19	0.845+/-0.02

	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(1, 1) lemmas	0.176+/-0.17	0.75+/-0.6	0.286+/-0.26	0.845+/-0.03
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(1, 1) POS	0.059+/-0.17	0.25+/-0.34	0.095+/-0.22	0.804+/-0.02
	(2, 2) POS	0.294+/-0.15	0.455+/-0.29	0.357+/-0.14	0.814+/-0.06
	(2, 3) POS	0.353+/-0.16	0.5+/-0.27	0.414+/-0.17	0.825+/-0.07
	(1, 1) semantic	0.412+/-0.16	0.636+/-0.16	0.5+/-0.15	0.856+/-0.03
	(2, 2) semantic	0.235+/-0.21	0.667+/-0.42	0.348+/-0.27	0.845+/-0.05
	(2, 3) semantic	0.235+/-0.17	0.667+/-0.21	0.348+/-0.2	0.845+/-0.03
	(1, 1) lemmas+POS	0.118+/-0.17	0.667+/-0.6	0.2+/-0.26	0.835+/-0.03
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(1, 1) POS+semantic	0.353+/-0.23	0.6+/-0.32	0.444+/-0.23	0.845+/-0.05
	(2, 2) POS+semantic	0.412+/-0.34	0.636+/-0.21	0.5+/-0.31	0.856+/-0.06
	(2, 3) POS+semantic	**0.471+/-0.22**	**0.615+/-0.22**	**0.533+/-0.19**	**0.856+/-0.05**
WHY	(1, 1) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.01
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(1, 1) lemmas	0.0+/-0.06	0.0+/-0.6	0.0+/-0.11	0.897+/-0.01
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 2) POS	0.1+/-0.13	0.333+/-0.62	0.154+/-0.19	0.887+/-0.03
	(2, 3) POS	0.2+/-0.23	0.333+/-0.33	0.25+/-0.24	0.876+/-0.03
	(1, 1) semantic	0.1+/-0.13	0.333+/-0.47	0.154+/-0.2	0.887+/-0.03
	(2, 2) semantic	0.0+/-0.13	0.0+/-0.83	0.0+/-0.22	0.897+/-0.01
	(2, 3) semantic	0.0+/-0.1	0.0+/-0.79	0.0+/-0.17	0.897+/-0.01
	(1, 1) lemmas+POS	0.0+/-0.09	0.0+/-0.92	0.0+/-0.17	0.897+/-0.02
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(1, 1) POS+semantic	0.1+/-0.13	0.5+/-0.74	0.167+/-0.21	0.897+/-0.03
	(2, 2) POS+semantic	0.1+/-0.2	0.333+/-0.62	0.154+/-0.3	0.887+/-0.04
	(2, 3) POS+semantic	**0.2+/-0.2**	**0.5+/-0.56**	**0.286+/-0.28**	**0.897+/-0.04**
HOW	**(1, 1) bag_of_words**	**0.25+/-0.23**	**1.0+/-0.8**	**0.4+/-0.35**	**0.938+/-0.02**
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(1, 1) lemmas	0.25+/-0.31	1.0+/-0.6	0.4+/-0.4	0.938+/-0.03
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0

	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 2) POS	0.0+/-0.12	0.0+/-0.6	0.0+/-0.18	0.907+/-0.02
	(2, 3) POS	0.125+/-0.18	0.25+/-0.54	0.167+/-0.2	0.897+/-0.04
	(1, 1) semantic	0.25+/-0.35	1.0+/-0.79	0.4+/-0.44	0.938+/-0.02
	(2, 2) semantic	0.0+/-0.12	0.0+/-0.98	0.0+/-0.22	0.918+/-0.01
	(2, 3) semantic	0.125+/-0.16	1.0+/-0.81	0.222+/-0.25	0.928+/-0.01
	(1, 1) lemmas+POS	0.25+/-0.35	1.0+/-0.8	0.4+/-0.47	0.938+/-0.03
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(1, 1) POS+semantic	0.0+/-0.12	0.0+/-1.0	0.0+/-0.22	0.918+/-0.02
	(2, 2) POS+semantic	0.0+/-0.11	0.0+/-0.34	0.0+/-0.17	0.907+/-0.02
	(2, 3) POS+semantic	0.125+/-0.17	0.5+/-0.77	0.2+/-0.23	0.918+/-0.02
SITUATION	(1, 1) bag_of_words	0.333+/-0.28	1.0+/-0.63	0.5+/-0.36	0.959+/-0.02
	(2, 2) bag_of_words	0.167+/-0.21	1.0+/-0.8	0.286+/-0.32	0.948+/-0.01
	(2, 3) bag_of_words	0.167+/-0.21	1.0+/-0.8	0.286+/-0.32	0.948+/-0.01
	(1, 1) lemmas	0.333+/-0.31	0.667+/-0.64	0.444+/-0.37	0.948+/-0.02
	(2, 2) lemmas	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(2, 3) lemmas	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.938+/-0.01
	(2, 2) POS	0.0+/-0.13	0.0+/-0.8	0.0+/-0.23	0.938+/-0.02
	(2, 3) POS	0.333+/-0.27	0.5+/-0.6	0.4+/-0.27	0.938+/-0.02
	(1, 1) semantic	0.333+/-0.26	0.5+/-0.49	0.4+/-0.3	0.938+/-0.03
	(2, 2) semantic	0.167+/-0.3	1.0+/-0.79	0.286+/-0.39	0.948+/-0.02
	(2, 3) semantic	0.167+/-0.31	1.0+/-0.76	0.286+/-0.39	0.948+/-0.02
	(1, 1) lemmas+POS	**0.333+/-0.34**	**1.0+/-0.74**	**0.5+/-0.44**	**0.958+/-0.02**
	(2, 2) lemmas+POS	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(2, 3) lemmas+POS	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(1, 1) POS+semantic	0.333+/-0.31	0.667+/-0.54	0.444+/-0.35	0.948+/-0.03
	(2, 2) POS+semantic	0.333+/-0.33	0.667+/-0.58	0.444+/-0.35	0.948+/-0.02
	(2, 3) POS+semantic	0.333+/-0.29	0.667+/-0.6	0.444+/-0.34	0.948+/-0.02
SCORE	(1, 1) bag_of_words	0.0+/-0.2	0.0+/-0.6	0.0+/-0.3	0.969+/-0.01
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(1, 1) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.01
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 2) POS	0.0+/-0.2	0.0+/-0.3	0.0+/-0.24	0.969+/-0.01
	(2, 3) POS	0.0+/-0.33	0.0+/-0.78	0.0+/-0.45	0.969+/-0.02
	(1, 1) semantic	0.0+/-0.45	0.0+/-0.67	0.0+/-0.48	0.969+/-0.02
	(2, 2) semantic	0.0+/-0.27	0.0+/-0.8	0.0+/-0.4	0.969+/-0.01

	(2, 3) semantic	0.0+/-0.27	0.0+/-0.8	0.0+/-0.4	0.969+/-0.01
	(1, 1) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.01
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(1, 1) POS+semantic	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.01
	(2, 2) POS+semantic	0.0+/-0.31	0.0+/-0.81	0.0+/-0.43	0.969+/-0.02
	(2, 3) POS+semantic	0.0+/-0.27	0.0+/-0.4	0.0+/-0.32	0.969+/-0.01

Tabela 71: Resultados dos classificadores *microaspectos* usando atributos léxico-semânticos

Macroaspecto	Classificador	Cobertura	Precisão	F1	Acurácia
WHAT	(1, 1) bag_of_words	0.48+/-0.11	0.632+/-0.1	0.545+/-0.1	0.583+/-0.08
	(2, 2) bag_of_words	**0.8+/-0.69**	**0.519+/-0.18**	**0.63+/-0.39**	**0.515+/-0.03**
	(2, 3) bag_of_words	0.54+/-0.83	0.529+/-0.35	0.535+/-0.45	0.52+/-0.05
	(1, 1) lemmas	0.48+/-0.13	0.6+/-0.06	0.533+/-0.08	0.567+/-0.04
	(2, 2) lemmas	0.72+/-0.75	0.529+/-0.24	0.61+/-0.41	0.526+/-0.05
	(2, 3) lemmas	0.4+/-0.73	0.571+/-0.32	0.471+/-0.39	0.536+/-0.06
	(1, 1) POS	0.46+/-0.09	0.59+/-0.12	0.517+/-0.07	0.561+/-0.08
	(2, 2) POS	0.56+/-0.09	0.609+/-0.1	0.583+/-0.09	0.588+/-0.09
	(2, 3) POS	0.56+/-0.12	0.596+/-0.09	0.577+/-0.09	0.577+/-0.08
	(1, 1) semantic	0.54+/-0.11	0.574+/-0.08	0.557+/-0.07	0.557+/-0.08
	(2, 2) semantic	0.5+/-0.13	0.641+/-0.13	0.562+/-0.11	0.602+/-0.1
	(2, 3) semantic	0.52+/-0.16	0.703+/-0.1	0.598+/-0.12	0.639+/-0.08
	(1, 1) lemmas+POS	0.44+/-0.13	0.579+/-0.06	0.5+/-0.09	0.542+/-0.05
	(2, 2) lemmas+POS	0.62+/-0.76	0.525+/-0.24	0.569+/-0.4	0.515+/-0.07
	(2, 3) lemmas+POS	0.4+/-0.7	0.571+/-0.3	0.471+/-0.38	0.536+/-0.08
	(1, 1) POS+semantic	0.54+/-0.12	0.6+/-0.12	0.568+/-0.09	0.573+/-0.1
	(2, 2) POS+semantic	0.52+/-0.1	0.591+/-0.11	0.553+/-0.09	0.567+/-0.1
	(2, 3) POS+semantic	0.56+/-0.1	0.596+/-0.11	0.577+/-0.08	0.577+/-0.09
CONSEQUENCE	(1, 1) bag_of_words	0.071+/-0.09	1.0+/-0.87	0.133+/-0.15	0.866+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(1, 1) lemmas	**0.071+/-0.09**	**1.0+/-0.81**	**0.133+/-0.16**	**0.866+/-0.02**
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 2) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.01
	(2, 3) POS	0.071+/-0.15	0.143+/-0.28	0.095+/-0.19	0.804+/-0.05
	(1, 1) semantic	0.071+/-0.17	0.333+/-0.53	0.118+/-0.25	0.845+/-0.02
	(2, 2) semantic	0.071+/-0.09	0.5+/-0.8	0.125+/-0.16	0.856+/-0.02
	(2, 3) semantic	0.071+/-0.15	0.333+/-0.75	0.118+/-0.24	0.845+/-0.04
	(1, 1) lemmas+POS	0.071+/-0.11	1.0+/-0.77	0.133+/-0.18	0.866+/-0.01
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(1, 1) POS+semantic	0.0+/-0.07	0.0+/-0.46	0.0+/-0.12	0.845+/-0.01
	(2, 2) POS+semantic	0.071+/-0.09	0.333+/-0.37	0.118+/-0.15	0.844+/-0.04
	(2, 3) POS+semantic	0.071+/-0.13	0.25+/-0.35	0.111+/-0.18	0.835+/-0.05
COMMENT	(1, 1) bag_of_words	0.143+/-0.25	0.5+/-0.87	0.222+/-0.34	0.928+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0

	(1, 1) lemmas	0.143+/-0.24	0.5+/-0.83	0.222+/-0.35	0.928+/-0.02
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 2) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.01
	(2, 3) POS	0.143+/-0.13	0.25+/-0.3	0.182+/-0.17	0.907+/-0.03
	(1, 1) semantic	0.143+/-0.17	0.5+/-0.72	0.222+/-0.25	0.928+/-0.02
	(2, 2) semantic	**0.143+/-0.18**	**1.0+/-0.84**	**0.25+/-0.28**	**0.938+/-0.01**
	(2, 3) semantic	0.143+/-0.22	1.0+/-0.72	0.25+/-0.32	0.938+/-0.01
	(1, 1) lemmas+POS	0.143+/-0.24	0.5+/-0.78	0.222+/-0.32	0.928+/-0.02
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(1, 1) POS+semantic	0.0+/-0.11	0.0+/-0.61	0.0+/-0.18	0.928+/-0.01
	(2, 2) POS+semantic	0.143+/-0.19	0.333+/-0.47	0.2+/-0.27	0.918+/-0.02
	(2, 3) POS+semantic	0.143+/-0.22	0.333+/-0.67	0.2+/-0.3	0.918+/-0.03
DECLARATION	(1, 1) bag_of_words	0.176+/-0.15	0.6+/-0.29	0.273+/-0.2	0.835+/-0.04
	(2, 2) bag_of_words	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(2, 3) bag_of_words	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(1, 1) lemmas	0.412+/-0.24	0.875+/-0.25	0.56+/-0.23	0.887+/-0.05
	(2, 2) lemmas	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(2, 3) lemmas	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(1, 1) POS	0.0+/-0.04	0.0+/-0.6	0.0+/-0.07	0.825+/-0.02
	(2, 2) POS	0.294+/-0.18	0.5+/-0.22	0.37+/-0.19	0.825+/-0.05
	(2, 3) POS	0.353+/-0.19	0.462+/-0.21	0.4+/-0.17	0.814+/-0.06
	(1, 1) semantic	0.353+/-0.22	0.6+/-0.31	0.444+/-0.25	0.845+/-0.05
	(2, 2) semantic	0.235+/-0.24	0.571+/-0.51	0.333+/-0.3	0.835+/-0.05
	(2, 3) semantic	0.176+/-0.24	0.5+/-0.51	0.261+/-0.27	0.825+/-0.05
	(1, 1) lemmas+POS	**0.529+/-0.24**	**0.9+/-0.13**	**0.667+/-0.21**	**0.907+/-0.04**
	(2, 2) lemmas+POS	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(2, 3) lemmas+POS	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(1, 1) POS+semantic	0.125+/-0.18	0.333+/-0.35	0.182+/-0.22	0.812+/-0.04
	(2, 2) POS+semantic	0.294+/-0.19	0.417+/-0.19	0.345+/-0.17	0.804+/-0.04
	(2, 3) POS+semantic	0.294+/-0.2	0.385+/-0.14	0.333+/-0.15	0.794+/-0.04
HISTORY	(1, 1) bag_of_words	0.0+/-0.07	0.0+/-0.6	0.0+/-0.12	0.907+/-0.01
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(1, 1) lemmas	0.0+/-0.11	0.0+/-0.98	0.0+/-0.2	0.907+/-0.01
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 2) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.01

	(2, 3) POS	0.0+/-0.09	0.0+/-0.61	0.0+/-0.15	0.866+/-0.05
	(1, 1) semantic	0.111+/-0.15	0.5+/-0.69	0.182+/-0.23	0.907+/-0.02
	(2, 2) semantic	0.0+/-0.07	0.0+/-0.6	0.0+/-0.12	0.907+/-0.01
	(2, 3) semantic	**0.111+/-0.11**	**0.5+/-0.91**	**0.182+/-0.19**	**0.907+/-0.02**
	(1, 1) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(1, 1) POS+semantic	0.0+/-0.1	0.0+/-0.81	0.0+/-0.18	0.907+/-0.01
	(2, 2) POS+semantic	0.111+/-0.14	0.5+/-0.91	0.182+/-0.23	0.907+/-0.02
	(2, 3) POS+semantic	0.111+/-0.2	0.333+/-0.6	0.167+/-0.26	0.897+/-0.04

Tabela 72: Resultados dos classificadores de *macroaspectos* usando atributos léxico-semânticos