Investigação de modelos de coerência local para sumários multidocumento

Márcio de Souza Dias

Márcio de Souza Dias

Investigação de modelos de coerência local para sumários multidocumento

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor 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
Março de 2016

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

Souza Dias, Márcio
S541i Investigação de modelos de coerência local para sumários multidocumento / Márcio Souza Dias; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2016.
191 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. Avaliação de Coerência Local. 2. Sumarização Multidocumento. 3. Modelos de Coerência Local. 4. Erros da Qualidade Linguística. 5. Anotação de Córpus. I. Alexandre Salgueiro Pardo, Thiago , orient. II. Título.

Márcio de Souza Dias

Investigation of local coherence models for multi-document summaries

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate 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
March 2016

‘O que prevemos raramente ocorre;
o que menos esperamos geralmente acontece.”
– Benjamin Disraeli

Agradecimentos

Gostaria de agradecer primeiramente a Deus por ter me proporcionado saúde e perseverança para a concretização deste Doutorado.

Sou muito grato ao meu orientador, Prof. Thiago Alexandre Salgueiro Pardo, por estar sempre presente com, paciência, inteligência, dedicação, e bom humor, qualidades que o torna não apenas um excelente professor e orientador, mas também um amigo.

Sou grato também a minha esposa Nádia Félix por estar sempre do meu lado com o seu amparo, a meus pais, Elizabeth e Luíz, e a toda a minha família pelo incentivo, apoio e conforto durante os anos de estudo.

Gostaria de deixar os meus agradecimentos também aos companheiros de trabalho, que de alguma forma colaboraram com este trabalho. Em especial, aos colegas Lianet, Lucía, Paula (A Diretoria), Leandro, Fernando, Pedro, Edílson, Vanessa, Roque, Marco, Alessandro, Fabrício, Jackson, Amanda, Erick Mazieiro, Erick Fonseca e Andressa.

Por fim, agradeço à Universidade de São Paulo - Campus São Carlos pela infraestrutura disponibilizada, e à Universidade Federal de Goiás - Regional Catalão pela liberação concedida a mim para cursar o doutorado.

Resumo

A sumarização multidocumento consiste na tarefa de produzir automaticamente um único sumário a partir de um conjunto de textos derivados de um mesmo assunto. É imprescindível que seja feito o tratamento de fenômenos que ocorrem neste cenário, tais como: (i) a redundância, a complementaridade e a contradição de informações; (ii) a uniformização de estilos de escrita; (iii) tratamento de expressões referenciais; (iv) a manutenção de focos e perspectivas diferentes nos textos; (v) e a ordenação temporal das informações no sumário. O tratamento de tais fenômenos contribui significativamente para que seja produzido ao final um sumário informativo e coerente, características difíceis de serem garantidas ainda que por um humano. Um tipo particular de coerência estudado nesta tese é a coerência local, a qual é definida por meio de relações entre enunciados (unidades menores) em uma sequência de sentenças, de modo a garantir que os relacionamentos contribuirão para a construção do sentido do texto em sua totalidade. Partindo do pressuposto de que o uso de conhecimento discursivo pode melhorar a avaliação da coerência local, o presente trabalho propõe-se a investigar o uso de relações discursivas para elaborar modelos de coerência local, os quais são capazes de distinguir automaticamente sumários coerentes dos incoerentes. Além disso, um estudo sobre os erros que afetam a Qualidade Linguística dos sumários foi realizado com o propósito de verificar quais são os erros que afetam a coerência local dos sumários, se os modelos de coerência podem identificar tais erros e se há alguma relação entre os modelos de coerência e a informatividade dos sumários. Para a realização desta pesquisa foi necessário fazer o uso das informações semântico-discursivas dos modelos CST (*Cross-document Structure Theory*) e RST (*Rhetorical Structure Theory*) anotadas no córpus, de ferramentas automáticas, como o *parser* Palavras e de algoritmos que extraíram informações do córpus. Os resultados mostraram que o uso de informações semântico-discursivas foi bem sucedido na distinção dos sumários coerentes dos incoerentes e que os modelos de coerência implementados nesta tese podem ser usados na identificação de erros da qualidade linguística que afetam a coerência local.

Palavras-chave: Avaliação da coerência local, Sumarização multidocumento, Erros de qualidade linguística, Anotação de córpus.

Abstract

The Multi-document summarization is the task of automatically producing a single summary from a collection of texts derived from the same subject. It is essential to treat many phenomena, such as: (i) redundancy, complementarity and contradiction of information; (ii) writing styles standardization; (iii) treatment of referential expressions; (iv) text focus and different perspectives; (v) and temporal ordering of information in the summary. The treatment of these phenomena contributes to the informativeness and coherence of the final summary. A particular type of coherence studied in this thesis is the local coherence, which is defined by the relationship between statements (smallest units) in a sequence of sentences. The local coherence contributes to the construction of textual meaning in its totality. Assuming that the use of discursive knowledge can improve the evaluation of the local coherence, this thesis proposes to investigate the use of discursive relations to develop local coherence models, which are able to distinguish automatically summaries coherent from incoherent ones. In addition, a study on the errors that affect the Linguistic Quality of the summaries was conducted in order to verify what are the errors that affect the local coherence of summaries, as well as if the coherence models can identify such errors, and whether there is any relationship between coherence models and informativeness of summaries. For this research, it was necessary the use of semantic-discursive information of CST models (*Cross-document Structure Theory*) and RST (*Rhetorical Structure Theory*) noted in the corpora, automatic tools, parser as *Palavras*, and algorithms that extract information from the corpus. The results showed that the use of semantic-discursive information was successful on the distinction between coherent and incoherent summaries, and that the information about coherence can be used in error detection of linguistic quality that affect the local coherence.

Keywords: Evaluation of the local coherence, Multi-document summarization, Linguistic quality errors, Corpus annotation.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo

1

Introdução

1.1 Contextualização e Lacunas

Com a grande quantidade de informações textuais disponíveis atualmente, principalmente na *web*, as pessoas vem se interessando em absorver essas informações de forma mais otimizada e resumida. Para se ter uma ideia da quantidade de informação no universo *online*, um estudo realizado pela *Cisco Visual Networking Index (VNI)*[1] projetou em 2016 um tráfego de dados global anual de 1,3 zetabyte - (um zetabyte equivale a um sextilhão de bytes ou um trilhão de gigabytes). O aumento projetado do tráfego de dados global apenas entre 2015 e 2016 é de mais de 330 exabytes, valor quase igual à quantidade total do tráfego de dados global gerado em 2011 (369 exabytes). Dentre esses dados, encontram-se as informações na forma textual, objeto de estudo da Sumarização Automática Multidocumento (SAM). Desta forma, a Sumarização Automática (SA) e a SAM vêm ganhando importância na comunidade científica.

A Sumarização Automática é a tarefa de produzir sumários de maneira automática a partir de um ou mais textos fontes, sendo considerada uma subárea de pesquisa de Processamento de Língua Natural (PLN) (Mani, 2001). A SA monodocumento, já tradicional, produz um sumário a partir de um único texto fonte. Segundo Mani (2001), a SAM produz um sumário a partir de um conjunto de textos relacionados a um mesmo assunto, ou seja, ela é a extensão da sumarização monodocumento.

Tendo este cenário favorável para a SAM, um sumário multidocumento só será útil para quem o lê se ele for informativo e coerente. Assim, a informatividade de um sumário multidocumento advém das principais informações contidas no conjunto de textos fontes. Já a coerência é um fator que facilita a compreensão e a interpretação do sumário.

A obtenção de sumários multidocumento informativos e coerentes é uma tarefa complexa e

[1] http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html Acessado em 12/01/16

envolve alguns desafios, como: i) o tratamento dos fenômenos multidocumento de redundância, complementaridade e contradição de informações, ii) uniformização de estilos de escrita, iii) tratamento de expressões referenciais, iv) manutenção de focos e perspectivas diferentes nos textos e v) ordenação temporal das informações no sumário.

Para exemplificar alguns desses desafios, a Figura 1.1 apresenta um sumário criado a partir de 3 textos fonte e que possui alguns fenômenos da sumarização automática multidocumento que não foram adequadamente tratados (as numerações das sentenças - S1, S2 e S3 - não fazem parte do sumário original). Inicialmente, as sentenças apresentam informações redundantes relacionadas ao lugar onde aconteceu o terremoto; as sentenças S2 e S3 apresentam uma informação contraditória referente à magnitude do terremoto; finalmente, em cada um dos parágrafos, são apresentadas informações complementares mal organizadas e pouco coesas. Pode-se perceber que o não tratamento desses fenômenos faz com que o texto seja pouco coerente.

Os desafios da SAM devem ser tratados de forma que as redundâncias e as contradições sejam eliminadas, que haja uniformização de estilos de escrita, que todas as expressões referenciais tenham os seus antecedentes, que o foco se mantenha ao logo do texto e que os segmentos informativos complementares sejam ordenados de forma coerente e coesa.

(S1) O tremor atingiu a região às 10h13 (horário local, 22h13 de domingo, em Brasília) e seu epicentro foi localizado a 260 km da costa de Niigata, ao nordeste da capital, Tóquio, onde também foi sentido.
(S2) TÓQUIO - Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão neta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feriados.
(S3) O terremoto de 7,4 graus, que pôde ser sentido em Tóquio, foi seguido por outro tremor de menor magnitude, de 4,2 graus na escala Ritcher, às 10h34m (22h34m de domingo em Brasília).

Figura 1.1: Exemplo 1 de um sumário automático multidocumento.

Outro sumário multidocumento gerado automaticamente a partir de 3 textos fontes com problemas que afetam a sua qualidade é mostrado na Figura 1.2. Nesse sumário, expressões referenciais como “a cidade” e “dos últimos ataques de Israel” em S1 não possuem os seus respectivos antecedentes. Além disso, a sentença S2 tem um foco diferente do foco apresentado na sentença S1, e, consequentemente, isso afeta negativamente a progressão textual.

Ainda em relação ao sumário da Figura 1.2, o pronome “outros” nas sentenças S2 e S4 confunde o leitor, pois no texto os respectivos referentes não estão explícitos. O mesmo problema pode acontecer caso o leitor não saiba o que são os termos “BBC” e “Hezbollah”, uma vez que no texto não há explicação sobre eles.

O sumário da Figura 1.3, criado a partir de 3 textos fontes, possui sentenças que quebram a sequência lógica do texto, por exemplo, as sentenças S6, S7 e S9 possuem focos distintos. Nesse sumário, há também sentenças com informações redundantes (sentenças S7 e S4 e sentenças S8

(S10) De acordo com um correspondente da BBC em Tiro, John Simpson, a cidade, na qual ficaram apenas cerca de 3 mil pessoas, ficou completamente isolada depois dos últimos ataques de Israel.
(S2) Outros nove ficaram feridos.
(S3) A ofensiva israelense foi lançada depois de uma sequência de ataques do Hezbollah no domingo que causou as maiores baixas para Israel nas quatro semanas do conflito.
(S4) Comandos israelenses mataram outros três guerrilheiros libaneses na cidade de Tiro, onde destruíram sete plataformas de lançamento de foguetes, informaram as fontes israelenses.
(S5) Já o Exército de Israel provocou a morte de 30 militantes do Hezbollah.
(S6) Os foguetes e ataques do Hezbollah causaram a morte de 15 pessoas e deixaram mais de 200 feridas.
(S7) Durante este domingo, dia 6, foram travadas lutas sangrentas.
(S8) Enquanto isso, soldados israelenses mataram 10 integrantes da milícia do Hezbollah.
(S9) A aviação de Israel realizou durante a madrugada desta segunda-feira, dia 7, ataques a 150 alvos no Líbano.

Figura 1.2: Exemplo 2 de um sumário automático multidocumento.

e S5). Além disso, as sentenças S8 e S5 também apresentam informações contraditórias em relação ao tempo, ou seja, "Aos 26 minutos" ou "Aos 27".

Nos sumários das Figuras 1.1 a 1.3, é perceptível o não tratamento dos problemas e desafios enumerados anteriormente. Esses problemas afetam diretamente a qualidade linguística do sumário e consequentemente a sua coerência. De forma geral, os sumarizadores multidocumento automáticos estão mais preocupados com a informatividade (selecionar o conteúdo mais relevante dos textos fonte para formar o sumário) do que gerar sumários coerentes, já que tais sumarizadores não possuem um tratamento adequado dos fatores que influenciam negativamente a coerência. Uma vez que, o tratamento da coerência é uma tarefa árdua e trabalhosa.

Os maiores desafios, dentre os apresentados, estão na manutenção da coerência e da coesão dos sumários. Portanto, a coerência textual de sumários multidocumento é influenciada pela informação redundante, pela ordenação (temporal ou não) dos segmentos textuais que compõem os sumários, pela fusão de segmentos textuais com informações complementares, pelo tratamento de informações contraditórias e manutenção de focos e perspectivas diferentes nos textos. Por outro lado, a coesão observa fatores de continuidade na superfície textual, como boa pontuação e uso de itens lexicais, uso apropriado de expressões referenciais, dentre outros fenômenos. Estes últimos fatores também interferem na coerência do texto.

Mais formalmente, para Koch (1998), a coesão textual diz respeito a todos os processos de sequencialização que asseguram (ou tornam recuperável) uma ligação linguística significativa entre os elementos que ocorrem no texto, ou seja, ligação entre palavras ou frases. Em outras palavras, a coesão é uma ligação equilibrada entre os parágrafos, as sentenças e as palavras, fazendo com que fiquem afinados entre si, com o intuito de obter uma relação de significância.

(S1) O Brasil lavou a alma após o decepcionante empate com a Colômbia no último domingo e, nesta quarta-feira, aplicou uma sonora goleada por 5 a 0 sobre o Equador no Maracanã.
(S2) O Brasil tocava a bola devagar e errava muitos passes.
(S3) No primeiro tempo o Brasil foi superior.
(S4) O Equador começou a gostar do jogo e ganhar confiança para avançar e dar alguns sustos no time comandado por Dunga.
(S5) Aos 26 minutos, a torcida xingava e pedia Obina na seleção, quando Kaká chutou forte de longe e Ronaldinho Gaúcho deu uma leve desviada na bola, enganando o goleiro equatoriano.
(S6) Kaká acertou um belíssimo chute de longe no ângulo aos 31 e fez 3 a 0.
(S7) Apesar de jogar melhor e dominar a partida, o Brasil não conseguia o segundo gol e o Equador começou a acreditar que dava para empatar e estragar a festa.
(S8) Aos 27, Kaká arriscou de muito longe e Ronaldinho colocou o desviou o chute.
(S9) A 20cm da linha de fundo ele deu dois dribles humilhantes no zagueiro equatoriano e cruzou para Elano, que fez o quarto, aos 37.

Figura 1.3: Exemplo 3 de um sumário automático multidocumento

Os elementos de coesão auxiliam na transição de ideias entre as sentenças e os parágrafos. Por exemplo no trecho, "Os manifestantes fizeram um protesto em Brasília contra a política, a corrupção e a má distribuição de renda do país, porque consideram injusta a atual situação do país. Porém o ministro da Justiça considerou a manifestação um ato de rebeldia, uma vez que alguns manifestantes provocaram tumulto e destruição do bem público", as palavras "porque", "porém" e "uma vez que" têm o papel de ligar as partes do texto, assim, essas palavras são responsáveis pela coesão do texto.

A coerência está diretamente ligada à possibilidade de estabelecer um sentido para o texto, ou seja, ela faz com que o texto faça sentido para o leitor, devendo, portanto, ser entendida como um princípio de interpretabilidade, ou seja, a compreensão do texto numa situação de comunicação e a capacidade que o receptor tem para aprender o sentido deste texto (Koch & Travaglia, 2002).

Dijk & Kintsch (1983) diferem dois tipos de coerência: a local e a global. A primeira é relativa a partes do texto, como sentenças ou sequências de sentenças dentro do texto. Embora as incoerências locais possam não comprometer totalmente o sentido do texto, de qualquer forma tornam mais difícil a compreensão. A coerência global é aquela que diz respeito ao texto em sua totalidade.

Segundo Koch & Travaglia (2002), a coerência local ocorre devido ao bom uso dos elementos da língua em sequências menores, para expressar sentidos que possibilitem a comunicação. Já as incoerências locais surgem pelo mal uso desses mesmos elementos linguísticos para o mesmo fim. Exemplos desse mal uso podem ser visto nas sentenças abaixo:

1. Maria tinha limpado a casa quando chegamos, mas ainda estava limpando a casa.

2. Marcelo não foi trabalhar, entretanto estava doente.

3. O boi estava grávido.

Em (1), a incoerência esta presente pelo fato de se ter o mesmo processo verbal em duas etapas distintas de sua realização, como "terminado" e "não terminado" ao mesmo tempo, sendo isso impraticável. Já em (2), o problema está na conexão entre as duas orações da sequência, "Marcelo não foi trabalhar" e "estava doente", que possuem uma relação de oposição que contraria a relação de causa que parece ser mais aceitável ou esperada entre as ideias expressas pelas duas orações. Em (3), a incoerência é percebida por contrariar o conhecimento geral (boi não fica "grávido" ou "prenho"), contudo, isto só é verdade se o mundo representado pelo texto for o mundo real e não um mundo de fantasia ou mágico.

A referência também é um dos aspectos importantes tanto na coesão quanto na coerência. Os elementos de referência são os itens da língua que não podem ser interpretados semanticamente por si só, mas remetem a outros itens do discurso necessários à sua interpretação (Koch, 1998). Os elementos de referência vêm sendo estudados em trabalhos voltados para sumarização monodocumento e na sumarização multidocumento, no intuito de obter sumários coesos e coerentes.

Um exemplo da importância das expressões referenciais é dado por Gonçalves (2008). Observe o texto fonte da Figura 1.4, cujos termos "o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)", "Guerra" e "o agrônomo" possuem uma relação de correferência entre si, formando a chamada cadeia de correferência. Suponha que um sumarizador automático gere um sumário (Figura 1.5) do texto fonte presente na Figura 1.4. Veja que, no sumário obtido, o termo "o agrônomo" ficou sem um termo antecedente que o especifique, ou seja, a pessoa que o mesmo representa. Desta forma, um processamento no sumário (pós-edição) poderia ser feito por meio da substituição do termo "o agrônomo" por um elemento da cadeia de correferência que melhor facilite a interpretação deste termo, neste caso, o elemento da cadeia de correferência "o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)" define de forma mais clara e específica o termo geral "o agrônomo", proporcionando, assim, uma melhor compreensão do sumário, como mostra o sumário editado da Figura 1.6. Assim, um texto coerente deve preservar todos os seus termos de correferência, não deixando nenhum termo sem o seu antecedente que o explique.

Além do possível problema da não preservação dos termos referenciais por parte dos sumarizadores multidocumento (também comum na SA monodocumento), outros problemas podem ser notados. Veja o sumário multidocumento oriundo de 3 textos fonte, mostrado na Figura 1.7 (as numerações de parágrafos - §1, §2 e §3 - não fazem parte do sumário original), gerado por um sumarizador automático multidocumento. Tal sumário possui pronomes pessoais ("eu", "mim", "me" e "Eu") e pronomes possessivos ("meu" e "minha") sublinhados no primeiro parágrafo, sem uma entidade como referência, ou seja, não se sabe a quem estes pronomes estão se referindo, deixando este parágrafo do sumário incompreensível. Mas, lendo todo o sumário, subentende-se que se trata da fala da entidade "O presidente Luiz Inácio Lula da Silva". Desta forma, essa entidade deveria ter vindo antes dos pronomes, de forma que o antecedente

*"A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia. A opinião é d***o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)***. Guerra** *citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo. Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare. Para* **o agrônomo**, *o Brasil deve buscar o desenvolvimento de transgenias que tentam melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra".*

Figura 1.4: Texto fonte retirado de Gonçalves (2008, p. 17)

"Para o agrônomo, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra".

Figura 1.5: Sumário do texto mostrado na Figura 1.4 (Gonçalves, 2008, p. 17)

"Para o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina), o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra".

Figura 1.6: Sumário pós-editado

da referência fosse facilmente localizado.

Outro problema que aparece no sumário da Figura 1.7 é a redundância de informações. Como a mesma informação pode estar presente em vários textos fonte, essa não deveria aparecer no sumário mais de uma vez, como acontece nos trechos em negrito no sumário da Figura 1.7.

A não ordenação correta de sentenças é outro fator que pode prejudicar a coerência do sumário. A Figura 1.8 mostra um sumário multidocumento com problemas na ordem das sentenças. Uma ordenação mal realizada pode gerar um sumário confuso e incoerente para o leitor, pois várias informações estão desconexas sobre o assunto que possivelmente o sumário trataria.

Assim, se a estrutura de um sumarizador multidocumento possuísse um módulo que pudesse

§1.“A vaia e o aplauso são dois momentos de reação do ser humano. A única coisa que eu, particularmente, fico triste é que eu fui preparado para uma festa. . . . como se eu fosse convidado para o aniversário de um amigo meu, chegasse lá e encontrasse um grupo de pessoas que não queria a minha presença lá. Eu tenho certeza de que não é esse o pensamento do Rio de Janeiro. ***Depois que terminou o evento, várias pessoas vieram dizer que tinha sido organizado, que gente tinha recebido o convite. A mim, não me interessa o que aconteceu, já aconteceu. O importante é que foi uma abertura extraordinária dos Jogos Pan-Americanos”****, afirmou.*
§2.O presidente Luiz Inácio Lula da Silva afirmou nesta segunda-feira, durante o programa de rádio “Café com o Presidente”, que ficou triste com as vaias que recebeu durante a abertura oficial da décima quinta edição dos Jogos PanAmericanos, realizada no estádio do Maracanã, no Rio de Janeiro. ***“Depois que terminou o evento, várias pessoas vieram dizer que tinha sido organizado, que gente tinha recebido o convite. A mim, não me interessa o que aconteceu, já aconteceu. O importante é que foi uma abertura extraordinária dos Jogos Pan-Americanos.”***
§3. O presidente Luiz Inácio Lula da Silva classificou de “reação do ser humano” as vaias que recebeu, na última sexta-feira, durante a abertura dos Jogos Pan-Americanos do Rio, no Maracanã.

Figura 1.7: Exemplo 4 de um sumário automático multidocumento

avaliar eficientemente a coerência de seus possíveis sumários, grande parte dos problemas citados não ocorreriam nos sumários produzidos. De modo que a construção de tal módulo seja possível é necessário entender a coerência e a possibilidade da mesma ser reconhecida em modelos computacionais. Para isso, a coerência vem sendo objeto de estudo de vários trabalho e teorias, como a Teoria de *Centering* de Grosz et al. (1995), Grade de Entidades de Barzilay & Lapata (2008), e *Rethorical Structure Theory* (RST) de Mann & Thompson (1987).

A Teoria de *Centering* faz uso de restrições e regras que governam as relações entre o foco de atenção dos enunciados (sentenças) do texto (discurso) e nas escolhas de expressões de referência para modelar a coerência.

A abordagem de Grade de Entidades é baseada em entidades (substantivos e pronomes) e inspirada nos conceitos da Teoria *Centering*. Essa abordagem considera que a coerência é obtida a partir do modo como as entidades são introduzidas e discutidas ao longo do texto, ou seja, o modo como as entidades são distribuídas em textos coerentes.

Para *Rhetorical Structure Theory* (RST), um texto coerente necessita possuir uma boa organização textual (estrutura retórica), ou seja, caso um texto seja coerente é sempre possível obter sua estrutura retórica.

A Teoria de *Centering* e a *Rhetorical Structure Theory* são teorias linguísticas caras para serem implementadas computacionalmente de forma integral (já que elas não foram desenvolvidas para este fim computacional) e normalmente são utilizadas como teorias complementares para modelos automáticos voltados para a coerência.

(S1) "Tudo foi resolvido", afirmou Raymond Boucher, advogado de 242 vítimas.
(S2) Este seria o maior pagamento já feito pela Igreja desde que surgiu o escândalo de abuso sexual envolvendo religiosos em 2002 e elevaria o total de indenizações pago pela Igreja desde 1950, nos Estados Unidos, a US$ 2 bilhões (R$ 3,7 bilhões).
(S3) Desde 2002, mais de mil pessoas deram entrada em processos contra a Igreja Católica por abusos sexuais na Califórnia e, nos últimos anos, a arquidiocese de Los Angeles já pagou US$ 114 milhões a 86 vítimas.
(S4) Os advogados de mais de 500 pessoas que se dizem vítimas de abusos sexuais cometidos por padres e religiosos católicos no Estado da Califórnia anunciaram, na noite de sábado, ter feito um acordo de US$ 660 milhões (R$ 1,227 bilhão) com a Arquidiocese de Los Angeles para encerrar os processos movidos contra ela.

Figura 1.8: Sumário multidocumento com problema de ordenação sentencial.

1.2 Objetivos do Trabalho

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é a exploração e desenvolvimento de modelos voltados a avaliar de forma automática a coerência local em sumários multidocumento gerados automaticamente.

Nesta tese em particular, investigações, incrementos e produções de modelos que fazem uso de informações discursivas capazes de auxiliar na identificação da coerência local em sumários multidocumento foram realizados. Entende-se por discurso o texto ou fala, compostos de várias unidades menores, que seriam as sentenças (Vieira & Lima, 2001). Por meio da análise do discurso há algoritmos para a resolução de referência, compreensão de diálogos e modelos de interpretação de textos e de distinção da coerência entre textos coerentes e incoerentes. Desta forma, teorias e métodos oriundos de discurso vem sendo utilizados em várias frentes de pesquisas em PLN, inclusive na coerência textual. Devido a isso, esta tese investigou e explorou trabalhos e teorias discursivas que pudessem ser úteis no desenvolvimento de modelos automáticos que possam distinguir sumários coerentes dos incoerentes. Uma das informações discursivas investigada e utilizada é a *Rhetorical Structure Theory* (RST) (Mann & Thompson, 1987) (ver seção 2.4.1), devido a característica de ser uma teoria voltada para a coerência de um texto ou discurso. Além disso, o córpus utilizado nesta pesquisa já está anotado com relações RST. Desta forma foi possível verificar que os textos coerentes, anotados com as relações discursivas da RST, possuem um padrão de relações discursivas da RST, ou seja, uma distribuição de relações RST que se repetem nos sumários coerentes. Assim, um modelo que utiliza esse padrão pode ser usado na avaliação da coerência local de sumários.

Outra informação discursiva apurada e útil na construção de um modelo de avaliação da coerência local e no incremento de modelos da literatura é a *Cross-Document Structure Theory* (CST) (Radev, 2000). Essa teoria é uma das principais voltadas para a sumarização multido-

cumento e propõe um conjunto de relações que permitem identificar similaridades, diferenças, contradições e informações complementares entre partes de textos sobre um mesmo assunto (ver Subseção 2.4.2). De forma similar a RST, mas agora tratando de relacionamento entre partes de diferentes textos sobre o mesmo assunto, verificou-se que o uso das relações CST em textos de referência tem um padrão que foi utilizado na distinção de sumários multidocumento de acordo com a coerência local. Além disso, o córpus adotado nesta tese já possui a anotação de relações discursivas da CST.

Com os modelos literatura adaptados e incrementados com informação discursiva, além de novos modelos discursivos criados neste trabalho, objetiva-se também a verificação do possível relacionamento de tais modelos com os possíveis erros da qualidade linguística dos sumários multidocumento. Esse possível relacionamento cria a possibilidade de obtenção de modelos que avaliam erros específicos da qualidade linguística.

É importante notar que este trabalho não teve a pretenção de fazer nenhum tipo de pós-edição dos sumários e sim de criar modelos automáticos que fossem capazes de avaliar sumários gerados automaticamente de acordo com a sua coerência local.

1.2.2 Objetivos Específicos

Os objetivos específicos são compostos por:

- Aumentar o poder de discriminação dos modelos de coerência da literatura com o acréscimo de informações discursivas;
- Desenvolver modelos independentes de língua;
- Desenvolver uma anotação de erros de qualidade linguística para o córpus de sumários automáticos;
- Verificar o desempenho dos modelos de coerência na possível identificação de erros da Qualidade Linguística do sumários;
- Investigar o possível relacionamento dos modelos de coerência com a informatividade dos sumários multidocumento;
- Contribuir com o estado da arte fornecendo um estudo em relação a utilização de conhecimentos até então não implementados para a avaliação da coerência local para a Sumarização Automática Multidocumento.

1.3 Tese e Hipóteses

Baseado nos objetivos deste trabalho, nos trabalhos de Lin et al. (2011) e Feng et al. (2014) que utilizam conhecimento discursivo no desenvolvimento de modelos de coerência para textos fonte e na falta de um módulo automático nos sumarizadores automáticos multidocumento

que avalie a coerência local nos sumários gerados de forma mais robusta que utilize conhecimentos linguístico-computacionais mais profundos, a tese deste trabalho é que conhecimento discursivo pode ser usado de forma satisfatória na avaliação da coerência local em sumários multidocumento, tanto no enriquecimento de modelos já existentes quanto na criação de modelos puramente discursivos.

De forma específica, seguem as seguintes hipóteses que direcionaram este trabalho:

- As informações das teorias discursivas escolhidas são úteis para a avaliação de coerência local.
- Os sumários coerentes possuem uma organização textual padrão baseado em relações discursivas que os distinguem dos sumários incoerentes.
- A utilização de técnicas de Aprendizado de Máquina proporcionará maior eficiência se comparada a métodos heurísticos.
- Os modelos de coerência local tem poder variado de discriminação de certos tipos de erros linguísticos.

1.4 Ineditismo e Contribuições

Este trabalho é o primeiro que focou em um estudo aprofundado na avaliação da coerência local para a sumarização multidocumento, pois até então havia um experimento em sumários multidocumento realizado por Barzilay & Lapata (2008) com os seus modelos de Grade de Entidades desenvolvidos especificamente para textos fontes. Neste estudo proposto nesta tese, destaca-se também o possível relacionamento dos modelos de coerência com os erros da Qualidade Linguística dos sumários automáticos multidocumento, algo que nenhum outro trabalho realizou. Além disso, uma análise de uma possível ligação entre os modelos de coerência com a informatividade dos sumários é algo novo também realizado neste trabalho.

Outro ponto a salientar sobre a originalidade deste trabalho é o uso de relações CST como elemento discriminador da coerência nos sumários multidocumento. O uso dessa teoria discursiva se deu pela própria natureza dos textos avaliados neste trabalho.

Algumas contribuições obtidas com este trabalho podem ser enumeradas:

- Modelos da literatura adaptados e incrementados com informação discursiva;
- Modelo formado somente com informação discursiva;
- Definição, anotação e análise de erros relacionados a qualidade linguística encontrados nos sumários automáticos;
- Análise do possível relacionamento entre os modelos de coerência e os erros da Qualidade Linguística dos sumários automáticos;
- Investigação do possível relacionamento entre os modelos de coerência e a informatividade dos sumários automáticos;

1.5 Organização da Tese

Este trabalho está organizado da seguinte forma:

- **Capítulo 2 - Fundamentação Teórica**: Conceitos e definições importantes sobre os elementos essenciais desta pesquisa serão descritos.
- **Capítulo 3 - Trabalhos Relacionados**: Os principais trabalhos da literatura relacionados ao tema desta pesquisa serão expostos.
- **Capítulo 4 - Adaptação de Métodos da Literatura**: Descrição da reimplementação dos modelos da literatura utilizados nesta tese serão realizados.
- **Capítulo 5 - Enriquecimento de Métodos com Informação Discursiva**: Detalhamento da incorporação de conhecimento discusivo nos modelos da literatura que originalmente não utilizam desse conhecimento será praticado. Além disso, modelos discursivos desenvolvidos nesta tese também serão detalhados.
- **Capítulo 6 - Experimentos e Resultados com Sumários Automáticos**: Avaliação e análise da aplicação dos modelos trabalhados nesta tese em sumários automáticos multidocumento serão detalhados. Além da anotação de erros de qualidade linguística no córpus de sumários automáticos que será relatada.
- **Capítulo 7 - Considerações Finais**: Pontos finais sobre o trabalho serão expostos, considerando as suas limitações e possíveis trabalhos futuros.

Capítulo

2

Fundamentação Teórica

Os principais conceitos, recursos e materiais utilizados nesta tese serão elucidados neste capítulo. Os conceitos como a Coerência, a Coesão e a Sumarização Multidocumento serão definidos e fundamentados de forma a facilitar a compreensão dos mesmos que são os principais elementos de estudo desta tese. Além disso, os recursos e materiais como as teorias discursivas, *parser* sintático e córpus também serão definidos e descritos, pois tais recursos darão suporte no desenvolvimento deste trabalho e, portanto, terão uma atenção especial neste capítulo.

2.1 Coesão e Coerência

Todo texto, seja ele de qual gênero for, necessita ser bem estruturado de maneira que o mesmo não seja apenas uma soma ou sequência de frases isoladas, para que o leitor possa entender a mensagem que o texto quer transmitir. E, dentro dessa estruturação textual, dois elementos são importantes: a coesão e a coerência, detalhadas abaixo.

2.1.1 Coesão

Para Koch (1998), a coesão é apresentada por meio de elementos linguísticos, indicações na estrutura superficial do texto, sendo de caráter claro e direto, expressando-se na organização sucessiva do texto, isto é, por meio de ligações linguísticas sucessivas entre os elementos que ocorrem na superfície textual. A coesão textual é um fator importante do texto relacionado à conexão de palavras, expressões ou frases dentro de uma sequência. O texto coeso é construído com elementos de ligação que podem ser pronomes, verbos, advérbios, conectores coesivos (termos e expressões), além do uso de sinais de pontuação (vírgula, ponto final, dois-pontos, ponto-e-vírgula).

De acordo com Koch & Travaglia (1989), um texto coeso pode parecer incoerente, por dificuldades particulares do leitor, como o desconhecimento do assunto ou a não inserção na situação. Isso salienta que a coesão ajuda a estabelecer a coerência, mas não a garante, pois ela depende muito dos leitores do texto (seu conhecimento de mundo) e da situação. Assim, a coesão ajuda a perceber a coerência na compreensão dos textos, porque é resultado da coerência no processo de produção desses mesmos textos.

A coesão é dividida em dois grupos (Koch, 1998): a coesão referencial e a coesão sequencial.

A coesão referencial utiliza mecanismo de reiteração, ou seja, um componente da superfície do texto faz remissão a outro(s) elemento(s) do universo textual. O primeiro é denominado de "forma referencial ou remissiva" e o segundo elemento de "referência, referente textual ou antecedente".

Em geral, a coesão por referência se dá quando um termo (pronomes, numerais, advérbios de lugar) remete a outro termo já mencionado no texto. Por exemplo, em "Ana Luiza está viajando. Ela está de férias", o pronome "Ela" está referenciando a Ana Luiza.

O chamado elemento de referência pode ser representado por um substantivo, um sintagma nominal (SN), um fragmento de oração, uma oração ou todo enunciado, por exemplo: "O carro estragou durante a viagem. Isso fez com que Roberta chegasse atrasada ao evento", a oração "O carro estragou durante a viagem" é o elemento de referência e o pronome demonstrativo "isso" é o elemento remissivo.

A coesão sequencial está relacionada à progressão textual, em que existem elementos que se unem para dar ideia de sequencialidade e continuidade da informação principal do texto, ou seja, as relações semânticas de causa, condição, oposição, tempo, conformidade, finalidade, são chamadas de encadeadoras do discurso. Em outras palavras, a coesão sequencial é feita por encadeamento de segmentos textuais e tem por função assinalar que a informação se desenvolve, ou seja, leva à frente o discurso. Os conectores contribuem para estabelecer relações lógicas entre as ideias do texto. Tais conectores são elementos de natureza gramatical (pronomes, conjunções, preposições, categorias verbais), léxica (sinônimos, antônimos, repetições) e mecanismos sintáticos (subordinação, coordenação, ordem dos vocábulos e orações). Em seguida, alguns exemplos do uso de conectores:

1. Luíz teve sucesso na prova porque se dedicou ao estudo.

2. Pesquisar exige disciplina e dedicação, mas o esforço vale a pena.

3. Um menino furou a bola. O garoto ficou triste.

4. Todos aqui estão contentes e felizes pelo seu sucesso.

Os conectores "porque" e "mas" estão fazendo ligações entre as orações. O conector "porque", no exemplo 1, liga a oração "Luíz teve sucesso na prova" à oração subsequente "se dedicou ao estudo". O conector "mas", no exemplo 2, estabelece ligação entre a oração "Pesquisar exige disciplina e dedicação" e a oração "o esforço vale a pena". No exemplo 3, os termos "menino"

e "garoto" são conectores de natureza léxica dada pela relação de sinônimo. Já no exemplo 4, o termo "e" é um conector de coordenação que faz a ligação entre as ideias da sentença. Assim, os recursos de coesão devem ser usados para expressar no texto a direção discursiva-argumentativa que o locutor quer imprimir no texto ou a direção que ele pretende dar ao discurso oral ou escrito.

2.1.2 Coerência

Segundo Koch & Travaglia (2002), a coerência está ligada à possibilidade de instituir um significado para o texto. Ou seja, a coerência é compreendida como um princípio da interpretação do texto.

Para de Beaugrande & Dressler (1981), a continuidade de sentidos é o que sustenta a coerência, ou seja, a configuração de conceitos e relações. O que está por trás de um texto é o mundo textual que contém mais do que o sentido das expressões na superfície do texto. Desta forma, a coerência é um produto da combinação de conceitos e relações dentro de uma rede composta por tópicos. A manutenção de um assunto do texto, ou tópico sobre o qual versa a narração, é um elemento importante para garantir, entre outros aspectos, a existência de coerência em um texto.

Para Koch & Travaglia (2002), a coerência decorre de alguns fatores das mais diversas ordens:

- Elementos Linguísticos: servem como indicações para estimular os conhecimentos armazenados na memória humana, inicializam a elaboração de inferências e ajudam na obtenção da orientação argumentativa dos enunciados que compõem o texto.

- Conhecimento de Mundo: caso o texto trate de um assunto desconhecido para o leitor, a obtenção de sentido será complicada e o texto parecerá destituído de coerência.

  O conhecimento de mundo só é adquirido a partir do que vivenciamos, por meio do contato com o mundo que nos cerca e tendo experiências em uma série de fatos.

  Assim, para estabelecer a coerência de um texto também é preciso que haja correspondência, ao menos parcial, entre o conhecimento nele ativado e o conhecimento de mundo, pois, caso contrário, não haverá condições de construir a mensagem que o texto quer transmitir, dentro do qual as palavras e expressões do texto ganham sentido.

- Conhecimento Compartilhado: é o conhecimento comum entre o produtor (escritor) e o receptor (leitor/ouvinte) da mensagem do texto. Quanto maior for essa parcela, menor será a necessidade de explicitar o conteúdo do texto, pois o receptor será capaz de suprir as lacunas, por meio de inferências, por exemplo.

  Os elementos textuais que transmitem o conhecimento partilhado entre os interlocutores constituem a chamada informação "velha" ou dada, ao passo que tudo aquilo que for introduzido a partir dela constituirá informação nova trazida pelo texto.

Para que um texto seja coerente, é preciso haver um equilíbrio entre a informação dada e a informação nova. Caso um texto contenha apenas informação nova, o mesmo poderia ser incompreensivo, pois faltaria ao receptor o conhecimento prévio necessário para a compreensão do texto. Por outro lado, um texto só com informação dada seria altamente redundante, isto é, "caminharia em círculos", sem preencher seu propósito comunicativo.

- Inferências: utilizando seu conhecimento de mundo, o receptor da mensagem do texto estabelece uma relação não explícita dele com o texto de forma que ele busca compreender e interpretar; ou, então, entre segmentos de texto e os conhecimentos necessários para a sua compreensão.

  Compete ao receptor ser capaz de atingir os diversos níveis implícitos, se quiser alcançar uma compreensão mais profunda do texto que ouve ou lê. Por exemplo, a sentença "Guilherme comprou uma Mercedes novinha em folha" pode possuir as seguintes inferências:

  1. Guilherme tem um carro.
  2. Guilherme tinha dinheiro para comprar um carro.
  3. Guilherme é rico.
  4. Guilherme é melhor companhia que João.

  Pode-se observar que nem todas essas inferências são necessárias: 3 é menos necessária do que 1 e 2; 4 é a menos necessária e só será feita dependendo do contexto em que a sentença aparece.

- Informatividade: diz respeito ao grau de previsibilidade da informação contida no texto.

  Em um texto que contiver apenas informação previsível ou redundante, o grau de informatividade será baixo. Por exemplo, a sentença "O oceano é água" é previsível. Outro texto que possuir informação não previsível terá um grau maior de informatividade, por exemplo, "O oceano é água. Entretanto, ele se compõe, na verdade, de uma solução de gases e sais". Por fim, se toda a informação de um texto for inesperada ou imprevisível, ele pode ter um grau máximo de informatividade, podendo, à primeira vista, parecer incoerente por exigir do receptor um grande esforço de decodificação. Por exemplo, em "O oceano não é água. Na verdade, ele é composto de uma solução de gases e sais", não é uma informação trivial e exige uma certa reflexão por parte do leitor ou ouvinte para compreendê-la.

- Focalização: relacionada com a concentração dos usuários (produtores e receptores) em apenas uma parte do seu conhecimento, bem como a perspectiva da qual são vistos os componentes do mundo textual. O produtor fornece ao receptor determinados indícios sobre o que está focalizando, ao passo que o receptor terá de recorrer a crenças e conhecimentos partilhados sobre o que está sendo focalizado, para poder entender o texto de modo adequado.

Diferenças de focalização podem causar problemas sérios de compreensão, impedindo, por vezes, o estabelecimento da coerência.

A mesma palavra poderá ter sentido diferente, dependendo da focalização. No caso de palavras homônimas, a focalização comum dos interlocutores permitirá inferir o sentido do termo naquela situação específica. Por exemplo, o termo "vela" em "Traga-me uma vela nova", pode ter vários sentidos de acordo com cada situação dada abaixo:

1. O marido para a mulher no momento em que acaba a luz.
2. O mecânico que está consertando um carro.
3. O armador que está construindo um barco.

- Consistência e Relevância: exige que cada enunciado de um texto seja consistente com os enunciados anteriores, isto é, que todos os enunciados do texto possam ser verdadeiros (ou não contraditórios) dentro de um mesmo contexto ou dentro dos contextos representados no texto.

  A relevância exige que o conjunto de enunciados que compõem o texto seja relevante para um mesmo tópico discursivo subjacente, isto é, que os enunciados sejam interpretados como "falando" sobre um mesmo tema.

Assim, o texto é mais que uma sequência de palavras e expressões. O mesmo surge por meio de uma competência específica do falante, que é a competência textual. Verificar o que faz com que um texto possa ser definido como coerente envolve a determinação de seus princípios de coesão e de constituição, fatores esses responsáveis por sua coerência.

Tipo de Coerência

Como já mencionado na introdução deste trabalho, a coerência pode ser tanto local quanto global (Dijk & Kintsch, 1983). A coerência local está relacionada com as partes do texto, como sentenças e sequências menores. Já a coerência global é aquela que se refere ao texto em sua totalidade. Os dois tipos de coerência podem estar presentes em diversos níveis:

- Semântico: relação entre significados dos elementos das frases em sequência em um texto (local) ou entre os elementos do texto como um todo (global). Por exemplo, o trecho da Figura 2.1 é incoerente semanticamente, pois a primeira e a segunda parte são contraditórias, ou seja, a posição da frente da casa e a parte que diz o que a avó faz à tarde são contraditórias, já que o sol não se põe a leste, mas a oeste.

- Sintático: refere-se as formas sintáticas de expressão da coerência semântica, representada pelo uso de recursos coesivos, tais como conectivos, pronomes, referências anafóricas, sintagmas nominais. Por exemplo, veja as seguintes frases (Koch & Travaglia, 2002, p. 44):

  1. João foi à festa, todavia ela não fora convidada.

> *A frente da casa de vovó é voltada para o leste e tem uma varanda grande. Todas as tardes ela fica na varanda em sua cadeira de balanço apreciando o pôr do sol.*

Figura 2.1: Trecho de texto com incoerência semântica (Koch & Travaglia, 2002, p. 43)

2. João foi à festa, todavia ele não fora convidado.

A sentença (1) é considerada problemática porque houve falha no uso do pronome, pois o pronome "ela" teria que se referir anaforicamente a "festa" e "a festa não pode ser convidada", dentro de um universo real. Mas, na sentença (2), os recursos sintáticos foram usados adequadamente para expressar a ideia.

- Estilístico: refere-se a elementos linguísticos (léxico, tipos de estruturas, sentenças, etc) pertencentes ou constituintes do mesmo estilo ou registro linguístico. Este nível é uma noção que tem utilidade na explicação de fenômenos de quebra estilística, por exemplo, o uso de gírias em textos acadêmicos, sobretudo orais (conferências) ou o uso de palavras de baixo calão em conversas "polidas" ser precedido de um "com o perdão da palavra".

- Pragmático: refere-se fundamentalmente à situação comunicativa em que o texto se insere, sendo este concebido como uma sequência de atos de fala (ações realizadas por um locutor através de um enunciado, visando intencionalmente obter algo da pessoa a que o enunciado se dirige) entre interlocutores. Para a sequência de atos ser percebida como apropriada, os atos de fala que a constituem devem satisfazer as mesmas condições presentes em uma situação comunicativa. Caso contrário, surgirá a incoerência. Por exemplo, uma pessoa faz um pedido a outra; seria esperada uma das seguintes sequências de atos:

  - pedido/atendimento;
  - pedido/promessa;
  - pedido/jura;
  - pedido/solicitação de esclarecimento;
  - pedido/recusa/justificativa;
  - pedido/recusa;
  - atendimento ou promessa.

  As seguintes sequências de atos não são esperadas quando uma pessoa faz um pedido a outra:

  - pedido/ameaça;

– pedido/declaração de algo que não tem nenhuma relação com o conteúdo do pedido.

No seguinte diálogo, um exemplo de incoerência pragmática é mostrado:

A: Você me empresta o seu livro de PLN?

B: Ontem o jogo foi empolgante.

No exemplo acima, o diálogo é considerado incoerente, devido ao ato de fala envolvido no diálogo não ser apropriado, pois, quando uma pessoa realiza um pedido a outra, espera-se que um dos atos de fala tradicionais seja utilizado.

Esses aspectos, como afirmam Koch & Travaglia (2002), precisam ser considerados, uma vez que influenciam no estabelecimento da interpretabilidade de um texto, seja ao compreendê-lo ou ao produzi-lo.

2.1.3 Relação entre Coesão e Coerência

Segundo Charolles (apud Koch & Travaglia (2002)), a coerência se relaciona com a linearidade do texto, ou seja, a coerência se relaciona com a coesão do texto. Pois, por coesão, entende-se a ligação, a relação, os nexos que se estabelecem entre os elementos que constituem a superfície textual.

Diferente da coerência, que é implícita, a coesão é explicitamente propagada por meio de marcas linguísticas, índices formais na estrutura da sequência linguística e superficial do texto, o que leva a uma característica linear, uma vez que se manifesta no decorrer do texto.

Apesar da coesão ajudar na formação da coerência, ela não garante a obtenção de um texto coerente. De acordo com Charolles (apud Koch & Travaglia (2002)), os elementos linguísticos da coesão não são nem necessários, nem suficientes para que a coerência seja formada, pois sempre haverá a necessidade de recursos exteriores ao texto (conhecimento do mundo, dos interlocutores, da situação, de normas sociais, etc.). Assim, podem haver textos sem elementos coesivos, mas cuja textualidade pode ocorrer no nível da coerência, por exemplo, o texto (sequência de palavras) dado pela Figura 2.2. Esse exemplo mostra que há uma sequência de nomes que poderia ser um amontoado aleatório se não constituísse uma lista de convidados para uma festa, o que os relaciona, criando assim, uma unidade.

Por outro lado, sequências linguísticas coesas podem existir, contudo, não chegam a formar um sentido global que as façam coerentes, como mostra o exemplo na Figura 2.3. No exemplo, a sequência é coesiva, mas o significado desse trecho está desconexo, levando à falta de entendimento da informação que este propunha transmitir.

Assim, a coesão ajuda a estabelecer a coerência na interpretação dos textos, porque surge como uma manifestação superficial da coerência no processo de produção desses mesmos textos. Desta forma, um texto é coerente porque as frases que o compõem guardam entre si determinadas relações. E é por meio dessas relações, que também são dadas por teorias e métodos linguístico-computacionais, e também pelo relacionamento entre a coesão e a coerência textual, que este trabalho propôs criar modelos que pudessem capturar um padrão dessas relações em

Lista de convidados para festa de aniversário

- *João da Silva*
- *José Gregório e esposa*
- *Tereza Mardin e noivo*
- *Cecília Machado*
- *Tios, tias e primos*
- *Meus irmãos*

Figura 2.2: Texto sem coesão, mas coerente (Koch & Travaglia, 2002, p. 22)

João vai à padaria. A padaria é feita de tijolos. Os tijolos são caríssimos. Também os mísseis são caríssimos. Os mísseis são lançados no espaço. Segundo a teoria da Relatividade o espaço é curvo. A geometria rimaniana dá conta desse fenômeno.

Figura 2.3: Trecho de texto sem Coerência (Marcuschi, 1983, p. 31)

textos considerados coerentes, com o intuito de avaliar a coerência local em sumários multidocumento.

Os modelos de coerência desenvolvidos nesta tese analisam a coerência de um sumário por meio do seu texto, onde a coesão se manifesta. Caso haja problemas na coesão de um texto, a coerência desse texto pode ser prejudicada e, consequentemente, tal texto pode ser considerado incoerente pelos modelos de coerência.

Um aspecto interessante dos modelos de coerência é que os mesmos não foram desenvolvidos focados em um tipo de coerência específico (como os apresentados neste capítulo), mas em um modelo de texto coerente. Esse texto coerente apresenta informações que podem ser utilizadas na sua distinção com textos considerados incoerentes, tais informações podem ser distribuição de entidades, distribuição de informação sintática e nesta tese a distribuição de relações discursiva é uma outra informação utilizada para distinção, como pode ser visto no Capítulo 5.

2.2 Correferência

Para Halliday & Hasan (apud Koch (1998)), a correferência é um mecanismo de coesão e também é considerado o elemento essencial dentro da coesão referencial. A correferência é definida por Koch (1998) como aquela em que um componente do texto (forma referencial, remissiva) faz remissão a outro(s) elemento(s) (elemento de referência, referente textual ou antecedente) do universo textual.

2.2.1 Tipos de Correferência

De acordo com Halliday & Hasan (apud Koch (1998)), a correferência pode ser **situacional (exofórica)** e **textual (endofórica)**. A correferência é dita **exofórica** quando o referente está fora do texto. Por exemplo, em "Você ajudará no dever de casa", o termo sublinhado (Você) refere-se a uma entidade fora do texto. Na correferência **endofórica**, o referente se acha expresso no próprio texto. Por exemplo, em "Maria é uma excelente professora. Ela se formou na Universidade de São Paulo", o pronome pessoal sublinhado está relacionado a um elemento identificado no próprio texto, no caso, "Maria".

A correferência pode ser feita para trás e/ou para frente, formando assim uma **anáfora** e/ou uma **catáfora**, respectivamente. Se o referente precede o item coesivo, tem-se a anáfora. Por exemplo, em "O José está viajando. Por isso que não o encontrei", o pronome pessoal "o" (sublinhado) é o termo anafórico, referencialmente dependente, que retorna o valor do grupo nominal "O José". Entretanto, se o referente vem após o item coesivo, tem-se a catáfora. Por exemplo, em "Ela era tão boa, a minha esposa!", o elemento referente ("a minha esposa!") vem após o item coesivo ("Ela").

A anáfora pode ser: direta, indireta, associativa e nova no discurso (Rossi et al., 2001). Tais tipos são definidos e exemplificados, a seguir:

- Anáforas diretas: são aquelas antecedidas por uma expressão (definida ou não, ou seja, expressões precedidas por artigos definidos ou não) que tem o mesmo nome-núcleo (substantivo) e referem-se à mesma entidade no discurso, por exemplo:

  "O time da Espanha foi derrotado pelo Brasil na final da Copa das Confederações. Mas o time mostrou um bom futebol".

- Anáforas indiretas: são aquelas antecedidas por uma expressão (definida ou não) que não têm o mesmo nome-núcleo do seu antecedente. As sentenças a seguir exemplificam esse tipo de anáfora.

  "O Flamengo e o Vasco fizeram um grande clássico. Os times mostraram raça e dedicação".

- Associativas: são as que introduzem um referente novo no discurso, o qual possui uma relação semântica com algum antecedente já introduzido. Desta forma, a descrição definida tem seu significado "amarrado" em uma entidade, o que impossibilita classificá-la como nova no discurso. Abaixo, um exemplo é apresentado.

  "O carro dos bandidos foi todo destruído no acidente na tentativa de fuga. Somente as rodas ficaram intactas".

- Novas no discurso: são aquelas que introduzem um novo referente no texto que não se relaciona com nenhum antecedente no discurso, ou seja, não tem uma "âncora" em que possa se apoiar semanticamente. Em sua maioria, ocorrem no início do texto ou com sintagmas nominais seguidos de sintagmas preposicionais. Veja o exemplo a seguir:

  "O presidente da Fifa, Joseph Blatter, demonstrou preocupação com as manifestações populares do Brasil.".

Além disso, a correferência pode ser: pessoal (realizada com a utilização de pronomes pessoais e possessivos), demonstrativa (feita por meio de pronomes demonstrativos e advérbios indicativos de lugar), e comparativa (praticada por via indireta, por meio de identidades e similaridades) (Halliday & Hasan, 1976). Alguns exemplos de correferência são mostrados em seguida:

- Romário e Bebeto foram ótimos jogadores. Eles formaram a dupla de atacantes na copa do mundo de 1994. (Correferência pessoal).

- Comprei todos os produtos, menos este: o filtro de água para a minha casa. (Correferência demonstrativa).

- É um ser inteligente igual a nós. (Correferência comparativa).

- Por que você está decepcionado? Esperava algo diferente? (Referência comparativa).

Para Koch (1998), a substituição é uma forma de correferência (anáfora indireta) que consiste na colocação de um item em lugar de outro(s) do texto, ou mesmo de uma oração inteira.

Por exemplo: “Meu irmão comprou um computador e eu também”, sendo que a palavra também está substituindo o evento de comprar um computador.

Desta forma, a correferência é uma peça importante nesse quebra-cabeça linguístico na busca de sumários automáticos coerentes. A informação de correferência será útil neste trabalho como um aspecto da qualidade linguística que influencia diretamente a coerência local, ou seja, um sumário coerente possui todos os seus termos relacionados aos seus respectivos antecedentes.

A correferência é tratada pelos modelos de coerência desta tese por meio do agrupamento de todos os sintagmas nominais de mesmo núcleo, pois alguns modelos necessitam de todos os elementos correferentes para um melhor desempenho. Essa medida é difundida na literatura quando não se dispõe de uma ferramenta robusta que possa tratar a correferência de forma automática, como foi o caso deste trabalho.

2.3 Sumarização

Um sumário é a versão mais curta de um ou mais textos (Mani, 2001). Os sumários, também conhecidos por resumos, estão cada vez mais presentes e corriqueiros no cotidiano das pessoas. Os sumários podem ser, por exemplo: manchetes de jornais escritas em uma linguagem concisa e direta sobre uma determinada notícia; *trailers* ou prévias de um filme, de um show artístico ou até mesmo de uma peça de teatro; narrativa de uma pessoa para outra sobre um evento ocorrido, onde esta tende a ser breve e sem muitos detalhes; etc.

Sumários, de forma geral, envolvem diversas pressuposições e características, como conteúdos e correspondências com suas fontes de origem diversificadas (Martins et al., 2001). Por exemplo, um sumário jornalístico esportivo pode considerar que um bom título de destaque para um texto fonte, que descreve a conquista da seleção brasileira de futebol da copa do mundo de 1994 nos Estados Unidos, é mencionar o grande destaque do atacante Romário na conquista do Brasil. Assim, um título possível para o sumário seria “Romário leva o Brasil ao Tetracampeonato”. Para o mesmo evento, outro sumário humano pode priorizar a coletividade da equipe brasileira na conquista da copa do mundo, não levando em conta somente a atuação do jogador Romário. Neste caso, outro título possível para o sumário seria “O time do Brasil vence a copa do mundo nos EUA”.

O exemplo anterior mostra que um determinado evento pode ser resumido de acordo com a proposta do autor, ou seja, um evento pode ser abstraído por meio de sumários com vários focos e informações veiculadas No primeiro título, o referido sumarizador quer chamar a atenção do leitor para o Romário, dando a ideia de que ele foi o único jogador decisivo para o título. Já no segundo, a equipe como um todo é o foco, ou seja, o título quer ilucidar que o futebol é um esporte coletivo e não individual. Neste exemplo, é interessante mencionar que os títulos podem também sumarizar seus respectivos textos.

As características na sumarização humana, como a variação de conteúdo informativo, grande quantidade de sentenças ou formas dos sumários, são pontos presentes na construção

de sumários automáticos, possibilitando assim, a produção de mais de um sumário para o(s) mesmo(s) texto(s) de origem (Martins et al., 2001).

Com estas variações na obtenção de um sumário, o processo automático de sumarização se mostra problemático, principalmente na questão de modelar de forma mais adequada um determinado sumário, de modo que este processo reflita a variedade de sumários sem que estes percam sua interdependência com os seus respectivos textos fonte (Martins et al., 2001).

De acordo com Martins et al. (2001), outras características oriundas da análise do processo de construção de sumários por humanos irão interferir no desenvolvimento de sumarizadores automáticos. São elas: 1) sumários direcionam a eventos ou a textos fonte dos mesmos e 2) sumários devem ser construídos sem que haja perda do significado essencial original, mesmo contendo poucas informações e apresentando diferentes estruturas, em relação a sua fonte.

Segundo Mani & Maybury (1999), sumários podem ser classificados com base na função que exercem: informativos, indicativos ou críticos. Os sumários informativos possuem as informações principais dos textos fonte e que possuem todas as características de "textualidade", podendo até mesmo substituir a leitura dos textos de origem. Os sumários indicativos, ao contrário dos informativos, não substituem os textos originais, mas apenas dizem do que se tratam. Por exemplo, índices podem ser classificados como sumários indicativos. Os sumários críticos apresentam juízos além do resumo em si. Exemplos de sumários críticos são as resenhas de livros.

A sumarização também pode ser monolíngue ou multilíngue. A monolíngue processa textos fonte em uma língua e produz um sumário nessa mesma língua. Já na multilíngue, os textos fonte podem estar em duas ou mais línguas e o sumário poderá ser em qualquer uma das línguas dos textos de origem. A maioria dos sumarizadores automáticos são monolíngues, entretanto, os sumários multilíngues vêm obtendo espaço por causa do crescimento de informação na web. Caso um leitor queira saber o que as principais agências de notícias internacionais relataram sobre um determinado assunto, o mesmo terá que recorrer a um sistema de sumarização multilíngue.

Os sumários também podem ser classificados em relação a audiência a que se destinam, assim, podem ser genéricos ou focados nos interesses dos leitores. A extração das informações mais relevantes dos textos fonte correspondentes, sem levar em conta os interesses particulares dos leitores, define a sumarização genérica. Já a sumarização focada nos interesses dos leitores prepara as informações que traz baseada nos conhecimentos dos leitores. Por exemplo, um leitor leigo em um determinado assunto do texto fonte precisará de um sumário com mais informações contextuais; já um leitor que detém um certo conhecimento sobre o contexto do texto original espera que o sumário contenha informações adicionais ou novas para o mesmo.

Extratos ou *abstracts* é uma outra classificação de sumários baseada na sua construção (Jones, 1993). Extratos são sumários formados por trechos não modificados do texto fonte (cópia e cola). Os *abstracts* apresentam partes ou são, como um todo, reescritos, ou seja, existe um nível de alteração na estrutura e/ou significado dos trechos extraídos do texto fonte.

A construção de sumários está relacionada a duas abordagens linguísticas: abordagem su-

perficial e profunda (Mani, 2001). A mescla da abordagem superficial com a abordagem profunda dá origem à chamada abordagem híbrida.

A abordagem superficial faz pouco ou nenhum uso do conhecimento linguístico para produzir sumários, entretanto, diferentes elementos podem ser representados em diferentes níveis. Por exemplo, palavras podem ser analisadas em nível semântico, mas sentenças serão analisadas, na maioria das vezes, no nível sintático. Essa abordagem limita-se a extrair partes importantes dos textos fonte e então organizá-las e apresentá-las de uma maneira mais eficaz. Conforme Mani (2001), a principal vantagem desta abordagem é a robustez.

Já a abordagem profunda faz grande uso de conhecimento linguístico, partindo de teorias e modelos formais da língua na criação de sumários, como léxicos, *wordnets*[1], gramáticas, análises sintático-semânticas e de discurso. É considerada a abordagem mais complexa, principalmente na construção de sumários de maneira automática, devido ao grande número de variáveis cognitivas e linguísticas a serem consideradas.

De forma geral, a Tabela 2.1 sintetiza as classificações possíveis para a sumarização em função dos critérios discutidos.

Tabela 2.1: Classificação dos Sumários

Critério	**Classificação**
Função	Indicativo, Informativo ou Critíco
Língua	Monolíngue ou Multilíngue
Audiência	Genérico ou Focado nos Interesses do Leitor
Formação	Extrato ou *Abstract*
Abordagem	Superficial, Profunda ou Híbrida
Número de textos fonte	Monodocumento ou Multidocumento

Um outro conceito importante na sumarização é a taxa de compressão. Essa taxa é a quantidade de informação inclusa no sumário pelo sumarizador. Ela é calculada pela razão entre o tamanho do sumário e o tamanho dos textos fonte (Mani, 2001). No caso da sumarização multidocumento, normalmente adota-se o tamanho do maior texto ou um número fixo de palavras. Assim, um sumário com taxa de compressão de 70% apresenta tamanho equivalente a 30% do tamanho do texto-fonte, geralmente medido em número de palavras.

De acordo com Mani (2001), a tarefa de sumarização multidocumento não é natural para humanos, mas Mckeown et al. (2005) demostraram que, apesar dos obstáculos, tanto os sumários produzidos pela SAM quanto os produzidos por humanos se mostraram muito úteis em experimentos que simulavam a compreensão da informação por humanos.

Na SAM, além de buscar a informação importante e necessária no conjunto de textos, há a necessidade de eliminar informação redundante do sumário, ordenar (de forma temporal ou não) os segmentos textuais que comporão os sumários, juntar segmentos textuais com informações complementares, realizar a manutenção da coerência do sumário, etc, lembrando que, os

[1]Um grande banco de dados léxicos para o Inglês, como a *wordnet* de Princeton (Miller, 1995)

sumários multidocumento são formados de fontes diferentes, sendo esses textos fonte escritos por pessoas diferentes e, consequentemente, têm estilos variados (Pardo, 2008).

Para exemplificar alguns dos desafios citados anteriormente (além dos já exemplificados no capítulo de Introdução desta tese), um sumário gerado por um sumarizador automático multidocumento a partir de três textos fonte é mostrado na Figura 2.4.

"A vaia e aplauso são dois momentos de reação de ser humano. A única que eu, particularmente, fico triste é que eu fui preparado para uma festa... como se eu fose convidado para o aniversário de um amigo meu, chegasse lá e encontrasse um grupo de pessoas que não queria a minha presença lá. Eu tenho certeza de que não é esse o pensamento do Rio de Janeiro. ***Depois que terminou o evento, várias pessoas vieram dizer que tinha sido organizado, que gente tinha recebido o convite. A mim, não me interessa o que aconteceu, já aconteceu. O importante é que foi uma abertura extraordinária dos Jogos Pan-Americanos"****, afirmou. O presidente Luiz Inácio Lula da Silva afirmou nesta segunda-feira, durante o programa de rádio "Café com o Presidente", que ficou triste com as vaias que recebeu durante a abertura oficial da 15^{TM} edição dos Jogos Pan-Americanos, realizada no estádio do Maracanã, no Rio de Janeiro.* ***"Depois que terminou o evento, várias pessoas vieram dizer que tinha sido organizado, que gente tinha recebido o convite. A mim, não me interessa o que aconteceu, já aconteceu. O importante é que foi uma abertura extraordinária dos Jogos Pan-Americanos."*** *O presidente Luiz Inácio Lula da Silva classificou de "reação do ser humano" as vaias que recebeu, na última sexta-feira, durante a abertura dos Jogos Pan-Americanos do Rio, no Maracanã.*

Figura 2.4: Sumário multidocumento gerado automaticamente

O sumário da Figura 2.4 possui pronomes pessoais ("eu", "mim", "me" e "Eu") e pronomes possessivos ("meu" e "minha") sublinhados, sem uma entidade como referência, ou seja, não se sabe a quem estes pronomes estão se referenciando, deixando este parágrafo do sumário incompreensivo. Outro problema na SAM são as informações redundantes presentes no sumário, pois a mesma informação pode estar presente em vários textos fonte e, sendo assim, não deveria aparecer no sumário mais de uma vez, como acontece nos trechos em negrito da Figura 2.4. Desta forma, um sistema de avaliação de coerência deveria avaliar tal sumário como incoerente (ou menos coerente), devido aos problemas apresentados.

2.3.1 Sumarizadores Automáticos Multidocumento para o Português do Brasil

Nessa Seção, as ferramentas de SAM que produziram os sumários utilizados nesta tese serão apresentadas. Tais ferramentas poderão ser beneficiadas no futuro com os resultados obtidos por esta tese.

O trabalho de Pardo et al. (2003) produziu um sumarizador automático mono e multidocumento chamado GistSumm (*GIST SUMMarizer*). Este sumarizador extrativo faz uso de técnicas para caracterizar o argumento principal, o *gist*, dos textos que serão sumarizados. Ele identifica a idéia principal do texto, e logo em seguida, agrega informações complementares.

Desta maneira, inicialmente o GistSumm busca a sentença que melhor expressa o argumento principal (*gist sentence*) e, por meio dessa sentença, seleciona as demais sentenças para formar o extrato. Entretanto, este sistema de sumarização não possui um tratamento específico e nenhuma avaliação automática de coerência textual, pois, o GistSumm foi avaliado por juízes humanos, os quais deram os pareceres sobre a informatividade.

O trabalho de Castro Jorge & Pardo (2012) focou na seleção de conteúdo, o qual resultou em um sumarizador multidocumento para o Português do Brasil denominado CSTSumm (*CST SUMMarizer*). Este sumarizador usa a CST (*Cross-document Structure Theory*) com base em um conjunto de 24 relações semântico-discursivas que representam fatores envolvidos na sumarização multidocumento (ver Seção 2.4.2). Esta teoria é utilizada no trabalho de Castro Jorge & Pardo (2012), inicialmente, para relacionar as unidades informativas presentes nos textos (sentenças), sendo que este relacionamento foi realizado de forma manual devido a falta de um analisador automático; em seguida, um grafo é construído a partir do relacionamento CST entre as unidades do texto. Desse grafo, um ranque das unidades informativas é obtido, isto é, quanto mais relevante for a unidade informativa, mais próximo do topo do ranque ela deve estar.

Para o ranque inicial, a relevância das unidades informativas depende do número de relações CST que elas apresentam, isto é, unidades com mais relações CST são consideradas mais relevantes. Assim, a partir do ranque inicial e da preferência do usuário, um ranque mais apurado é produzido, de tal forma que as unidades informativas mais relevantes, segundo o critério especificado pelo usuário, melhorem de posição no ranque e, consequentemente, ganhem preferência para estar no sumário. Por fim, as sentenças são selecionadas respeitando o ranque refinado e a taxa compressão dada.

O trabalho de Castro Jorge (2010) não faz nenhum tipo de abordagem para tratar a coerência, ou seja, o CSTSumm pode gerar sumários com problemas de coerência.

Ribaldo (2013) desenvolveu um sistema de sumarização automática multidocumento extrativo que segmenta cada texto de uma coleção em subtópicos. Ele utiliza uma versão adaptada do *TextTiling*[2] (Hearst, 1997) e agrupa os subtópicos com medidas de similaridade.

Com o agrupamento feito, um grafo de relacionamentos é formado e o conteúdo relevante é selecionado por meio do percurso caminho denso segmentado (Salton et al., 1997). No grafo,

[2]É uma técnica para subdividir um texto em unidades de multi-parágrafos que representam passagens ou subtópicos.

os nós representam sentenças e as arestas são relacionamentos entre eles. No caminho denso segmentado, selecionam-se as sentenças mais importantes de cada subtópico.

Logo após a escolha da primeira sentença, uma sentença de transição é necessária antes da escolha da próxima sentença mais relevante de outro subtópico. Essa sentença deve ser cronologicamente posterior a sentença precedente no sumário para que a passagem de um subtópico para outro se dê de forma coerente. Esse processo é realizado até atingir a taxa de compressão. De acordo com o autor, esse sistema ficou conhecido como RSumm.

Outro trabalho voltado para a sumarização automática multidocumento para o Português do Brasil é de Cardoso (2014), a qual procurou investigar métodos de seleção de conteúdo que priorizam a importância das informações e a representatividade dos tópicos em conjunto com os fenômenos multidocumento. A princípio, dois métodos de seleção de conteúdo para sumarização automática multidocumento foram propostos. Para isso, as teorias discursivas RST e CST foram utilizadas.

Um dos métodos de seleção de conteúdo proposto parte das sentenças com mais relações CST para depois aplicar a poda das informações adicionais indicadas pelas relações RST. Cardoso (2014) considerou que as relações CST indicarão as sentenças mais relevantes do conjunto de textos e as relações RST apontarão as proposições [3] mais importantes de cada texto.

O segundo método proposto por Cardoso (2014) utiliza o método de Marcu (1997) para buscar as unidades textuais mais relevantes por meio da RST. Em seguida, as sentenças são selecionadas de acordo com o número de relacionamentos CST. Tal método foi denominado pela autora de RC-4.

De acordo com a autora, esses métodos de sumarização automática multidocumento são os primeiros que utilizam a RST neste cenário multidocumento, sendo que a mesma foi bastante explorada para sumarização monodocumento. Nesse trabalho assume-se que a relevância de uma sentença é influenciada pela sua saliência, dada pela RST, e pela sua correlação com os fenômenos multidocumento, indicada pela CST. Utilizando da RST para remover segmentos satélites (segmentos não importantes), estratégia que funciona bem no cenário monodocumento. Quando essa mesma estratégia se aplica para um conjunto de textos, o resultado pode ser um sumário com problemas de coerência. Esse trabalho utiliza as relações CST para reconhecer e tratar informações redundantes com o foco de melhor a qualidade dos sumários extrativos formados e, consequentemente, a coerência do sumário.

Castro Jorge (2015) propõe uma abordagem gerativa estatística para a sumarização multidocumento. Especificamente, Castro Jorge (2015) quer formular a tarefa de sumarização multidocumento usando um modelo *Noisy-Channel*[4] (Shannon, 2001), por meio da exploração de fatores como a redundância, a complementaridade e a contradição. Tal trabalho investiga por meio de uma abordagem estatística gerativa os fatores envolvidos na geração de um sumário multidocumento, fazendo uso das relações semânticas da CST para representar os fatores ante-

[3] No contexto de discurso, proposições ou segmentos discursivos correspondem ao conteúdo de uma oração, de um segmento textual qualquer, de uma sentença, ou mesmo de um trecho maior de texto, dependendo do assunto que se discute.

[4] O esquema Noisy-Channel surgiu dentro da a?rea de Teoria da Informac?a?o como um teorema de codificac?a?o de dados na linha telefo?nica

riormente mencionados.

O trabalho de Castro Jorge (2015) avaliou a coerência dos sumários por meio de um modelo de língua (modela boas construções da língua de forma estatística). Visto que as possíveis métricas, BLEU e ROUGE, estão mais relacionadas com a avaliação da informatividade do que da coerência, a autora utilizou um dos modelos de coerência local desenvolvidos nesta tese (ver Seção 5.1) como um modelo de língua.

Para Castro Jorge (2015), o modelo de coerência melhorou os resultados dos sumários, em termos de medida ROUGE, em comparação com os resultados sem o uso desse modelo. Mesmo com a dificuldade de analisar o real impacto do modelo de coerência, já que os efeitos do mesmo não seriam detectados pela medida ROUGE, observou-se uma influência positiva do modelo de coerência na seleção de conteúdo, nas duas formas em que ele foi incorporado.

Como visto nessa Seção, alguns trabalhos apresentaram certas soluções que podem melhorar a qualidade de cada sumário gerado. Mas, tais soluções ainda não são suficientes para todos os tipos de erros que podem prejudicar a coerência textual. Assim, esta pesquisa pretende cobrir essa lacuna e ajudará os trabalhos, aqui apresentados, a melhorar seus respectivos sumarizadores na geração de sumários mais coerentes.

2.4 Conhecimento Discursivo

Nesta seção, as teorias discursivas que este trabalho acredita ser úteis na avaliação da coerência serão apresentadas.

2.4.1 *Rhetorical Structure Theory* - RST

A RST de Mann & Thompson (1987) considera que cada texto possui uma estrutura retórica subjacente e que tal estrutura permite recuperar o caráter comunicativo que o escritor do texto pretendeu atingir ao escrevê-lo. A estrutura RST é composta por unidades elementares do discurso (*Elementary Discourse Unit* ou EDUs), inter relacionadas por meio de relações retóricas. As EDUs são unidades mínimas de significado que compõem um texto. As relações retóricas indicam os tipos de relações existentes entre tais unidades, visando a organização coerente de um texto ou discurso.

De acordo com a RST, EDU é o conteúdo expresso pelo segmento textual que se relacionam por meio de relações retóricas. A cada EDU é atribuído um papel de núcleo ou satélite. O núcleo, ou unidade nuclear, expressa a informação principal em uma relação, sendo considerado mais relevante do que o satélite. O satélite apresenta informação adicional, a qual exerce influência na interpretação do leitor sobre a informação apresentada no núcleo.

Normalmente, os núcleos são compreensíveis independentemente dos satélites, mas o contrário não é verdadeiro, já que na maioria dos casos torna-se impossível a compreensão do satélite sem o seu respectivo núcleo. Há casos também em que as unidades de uma relação retórica podem ser nucleares, ou seja, ambas apresentam informações importantes. Desta forma,

as relações RST são divididas em duas classes: hipotáticas e paratáticas. As relações hipotáticas relacionam pares de EDUs que apresentam diferentes graus de importância, sendo uma nuclear e a outra satélite. Essas relações denominam-se mononucleares. As relações paratáticas relacionam EDUs que apresentam o mesmo grau de importância e são denominadas relações multinucleares.

Os autores afirmam que as relações retóricas da RST são capazes de representar todas as possíveis relações de significado entre os segmentos discursivos de uma grande quantidade de textos. O conjunto de relações originais pode ser visto na Tabela 2.2 .

Tabela 2.2: Relações Retóricas da RST

Relação Retórica	**Tipo de Relação**
ANTITHESIS	Mononuclear
BACKGROUND	Mononuclear
CIRCUMSTANCE	Mononuclear
CONCESSION	Mononuclear
CONDITION	Mononuclear
CONTRAST	Multinuclear
ELABORATION	Mononuclear
ENABLEMENT	Mononuclear
EVALUATION	Mononuclear
EVIDENCE	Mononuclear
INTERPRETATION	Mononuclear
JOINT	Multinuclear
JUSTIFY	Mononuclear
MOTIVATION	Mononuclear
NON-VOLITIONAL CAUSE	Mononuclear
NON-VOLITIONAL RESULT	Mononuclear
OTHERWISE	Mononuclear
PURPOSE	Mononuclear
RESTATEMENT	Mononuclear
SEQUENCE	Multinuclear
SOLUTIONHOOD	Mononuclear
SUMMARY	Mononuclear
VOLITIONAL CAUSE	Mononuclear
VOLITIONAL RESULT	Mononuclear

Mann & Thompson (1987) agruparam as relações segundo as suas semelhanças. Assim, cada grupo consiste de relações que compartilham de características e diferem em 1 ou 2 atributos.

A Tabela 2.3 mostra as relações RST agrupadas segundo Mann & Thompson (1987).

Com o objetivo de melhorar o entendimento das relações e também para anotar textos que precisavam de novas relações, Marcu (1997) e Pardo & Nunes (2008) modificaram e/ou complementaram as relações da RST. Marcu (1997) acrescentou relações ao conjunto original, destacando-se as chamadas relações estruturais, as quais conectam proposições que foram

Tabela 2.3: Relações RST agrupadas (Mann & Thompson, 1987)

Circumstance
Solutionhood
Elaboration
Background
Enablement and Motivation
Enablement
Motivation
Evidence and Justify
Evidence
Justify
Relations of Cause
Volitional Cause
Non-Volitional Cause
Volitional Result
Non-Volitional Result
Purpose

Antithesis and Concession
Antithesis
Concession
Condition and Otherwise
Condition
Otherwise
Interpretation and Evaluation
Interpretation
Evalutation
Restatement and Summary
Restatement
Summary
Other Relations
Sequence
Contrast

quebradas no fluxo do texto.

A relação PARENTHETICAL é um exemplo de relação estrutural, a qual indica que o satélite apresenta uma informação relacionada ao núcleo, que não está expressa no fluxo principal do texto, aparecendo geralmente entre parênteses, colchetes ou chaves. As relações que conectam segmentos encaixados, as quais são introduzidas por orações subordinas relativas, também foram incluídas por Marcu (1997). Essas relações são indicadas por "-e" (*embedded* em inglês) no final de seu nome e apresentam o mesmo significado das relações tradicionais. A Tabela 2.4 mostra as 32 relações e seus tipos, onde os asteriscos (*) identificam as relações multinucleares.

Para exemplificar o relacionamento dado pela RST, considere o texto na Figura 2.5. A sentença [1] ilustra a idéia central do discurso, que é a de que o medo determinava o modo como o personagem (Almir) agia. Entretanto, os segmentos textuais [2] e [3] indicam, respectivamente, o fato de que poucas pessoas conhecem essa característica do personagem e o fato de que essa característica é verdadeira. Desta forma, há três proposições distintas, expressas por [1], [2] e [3]. O relacionamento entre as proposições ([1], [2] e [3]) ocorre na medida que elas são reconhecidas nesse discurso, sendo assim, as proposições [2] e [3] são identificadas como constituintes de uma relação de elaboração da afirmação expressa em [1]. Na RST, tal relacionamento é expresso pela relação retórica ELABORATION, com a proposição correspondente à sentença [1] sendo o núcleo da relação ELABORATION, enquanto as duas proposições correspondentes aos segmentos [2] e [3], juntas, constituem o satélite da relação, conforme ilustra a Figura 2.6.

De acordo com a Figura 2.6, cada número representa uma proposição do discurso estruturado, que, no texto exemplo da Figura 2.5, é indicada pelos segmentos textuais numerados. Cada relação RST é representada por um arco direcionado, sendo sua direção do satélite para o núcleo. Assim, a proposição na ponta da seta é sempre o núcleo. Relações que não são re-

[1] Muitas das atitudes "corajosas" de Almir, o Pernambuquinho, eram ditadas pelo medo. [2] Poucos sabem disso, [3] mas é verdade. [4] Quem o via de punhos cerrados, dentes trincados, desafiando adversários mais fortes, não imaginava que, por trás da valentia, escondia-se o medo de parecer covarde. [5] Certa vez ele foi suspenso por uma jogada violenta [6] que inutilizou Hélio, do América. [7] À medida que ia se aproximando o fim da suspensão, [8] Almir começou a queixar-se de uma estranha dor muscular na perna direita. [9] Dr. Valdir Luz e todo o departamento médico do Vasco já não sabiam o que fazer para curar a inexplicável "distensão". [10] Acabou-se a suspensão, [11] mas permaneceu a dor. [12] Até que o técnico Yustrich chamou o jogador para uma conversa: [13] "Você não tem nada, garoto. [14] É o medo de que alguém vingue o Hélio [15] que faz você sentir a dor".

Figura 2.5: Texto Segmentado (Ribeiro & Rino, 2005, p. 2)

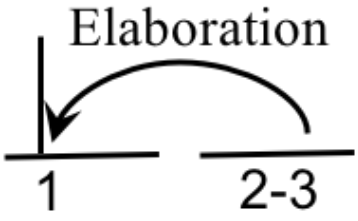

Figura 2.6: Relação ELABORATION entre as proposições 1 e 2-3 (Ribeiro & Rino, 2005, p. 2)

Tabela 2.4: Relações RST modificadas e/ou complementadas

Relação	Relação
ANTITHESIS	MOTIVATION
ATTRIBUTION	NON-VOLITIONAL CAUSE
BACKGROUND	NON-VOLITIONAL RESULT
CIRCUMSTANCE	OTHERWISE
COMPARISON	PARENTHETICAL
CONCESSION	PURPOSE
CONCLUSION	RESTATEMENT
CONDITION	SOLUTIONHOOD
ELABORATION	SUMMARY
ENABLEMENT	VOLITIONAL CAUSE
EVALUATION	VOLITIONAL RESULT
EVIDENCE	CONTRAST *
EXPLANATION	JOINT *
INTERPRETATION	LIST *
JUSTIFY	SAME-UNIT *
MEANS	SEQUENCE

presentadas por arcos direcionados são relações multinucleares, como pode ser visto na Figura 2.8.

Estruturas retóricas (estruturas RST) dão origem a árvores cujas folhas correspondem às proposições elementares e cujos nós internos às relações retóricas. A Figura 2.7 mostra a estrutura RST completa do texto da Figura 2.5. Assim, a RST é uma teoria que estabelece um conjunto de relações, a partir das quais se podem reconhecer os níveis de relevância das informações de um discurso e representar sua estrutura hierarquicamente, mediante a delimitação de suas proposições elementares.

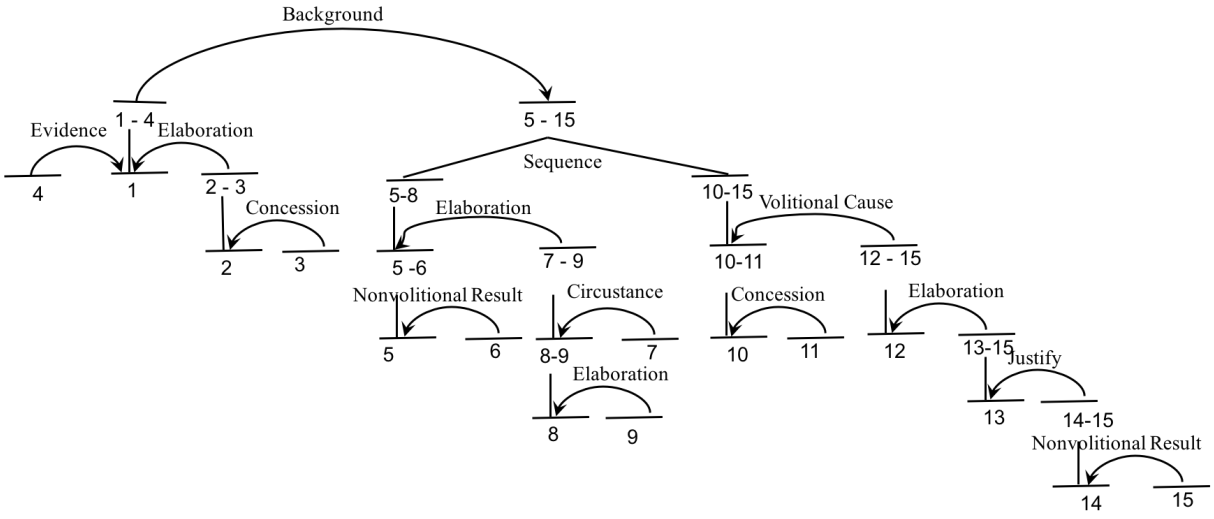

Figura 2.7: Estrutura RST do texto da Figura 2.5

De acordo com os autores, as proposições são essenciais para a coerência do texto, de tal

forma que, se um texto for coerente, será sempre possível extrair a sua estrutura retórica. Por esse motivo, as relações retóricas também são chamadas relações de coerência.

Para determinar qual a estrutura retórica correspondente a um texto é preciso distinguir cada uma de suas proposições elementares, associando-as a um núcleo ou satélite de uma relação retórica, além de reconhecer a própria relação. Isso é uma tarefa de interpretação realizada por humanos especialistas e pode ser difícil de ser feita, pois envolve a questão da interpretação que cada pessoa pode ter de forma distinta na determinação da relação retórica de uma proposição.

Além de identificar proposições pelo seu grau de importância, ou seja, proposições nucleares ou satélites, as proposições que se encontram no mesmo nível de importância também podem ser identificadas. Desta forma, as relações multinucleares são as que envolvem mais de duas proposições de mesmo nível de importância. Um exemplo de relação RST multinuclear é a relação CONTRAST, cuja definição contrapõe as proposições envolvidas. A Figura 2.8 mostra a representação da relação CONTRAST, correspondente ao segmento textual S1 a seguir (Ribeiro & Rino, 2005).

Segmento S1:

[1] Linguagens de programação de alto nível permitem ao programador uma maior naturalidade na forma de programar. [2] Entretanto, essas linguagens são mais lentas que linguagens de baixo nível durante sua execução.

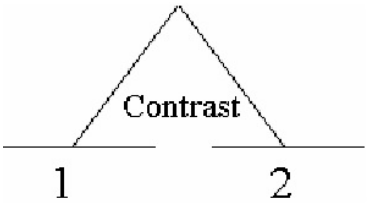

Figura 2.8: Relação CONTRAST Multinuclear

Trabalhar com essa teoria não é tão simples, já que envolve a distinção entre o que o escritor considera mais relevante para alcançar seu objetivo comunicativo, com seu discurso. O sucesso dessa distinção, pelo leitor, depende de sua observação empírica e subjetiva, atribuindo um grau de relevância maior ao que ele associa a um núcleo de uma relação RST do que ao que ele associa a seu satélite correspondente.

A noção de relevância, no contexto da RST, pode ser entendida da seguinte forma: em geral, os núcleos expressam informações que, se retirados, farão o texto resultante correspondente incoerente. Entretanto, ao retirar os satélites vinculados aos núcleos, o texto continuará coerente mesmo com uma quantidade menor de informações.

Esta é a característica que torna a RST interessante no tocante a avaliação da coerência: ao distinguir informações essenciais (nucleares) das complementares (satélites) é possível elaborar modelos que utilizem a estrutura RST de um texto coerente para distinguir textos incoerente ou menos coerente.

Para utilizar o conhecimento discursivo é necessário a construção de recursos, como analisadores discursivos e córpus anotados. A anotação de relações discursivas em um texto pode ser feita de forma manual ou automática. Anotação manual requer humanos treinados, tornando o

processo trabalhoso e demorado. A ferramenta RSTTool[5] de O'Donnell (2000) foi criada para facilitar a anotação manual de relações RST. Com esta ferramenta é possível segmentar um texto em proposições, conectá-los com relações RST e visualizar graficamente a estrutura de árvore, representando a análise final.

Já a anotação automática é realizada por ferramentas que detectam automaticamente as relações RST entre os segmentos de um texto e constroem sua estrutura discursiva. Há analisadores discursivos automáticos para RST, por exemplo, no Inglês (Marcu, 2000), no Espanhol (Cunha et al., 2010), no Português (Pardo & Nunes, 2008) e (Maziero et al., 2015), etc. O *parser* discursivo DiZer [6] para o Português tem o desempenho médio de 56,8% na segmentação textual, 62,5% na detecção de relações e 81% na determinação da nuclearidade. O DiZer utiliza padrões extraídos de um córpus de textos científicos e a sua aplicação em um outro domínio textual pode ter desempenho inferior. O trabalho de Maziero et al. (2015) utiliza a abordagem de aprendizado sem fim semissupervisionado para identificar relação RST intra sentencial. Esse trabalho obteve 79% de acurácia na identificação das relações RST intra sentencial.

Córpus anotados com relações discursivas é outra forma de trabalhar com teoria discursiva. Existem vários corpora e de diferentes línguas com relações RST anotadas, dentre eles estão: o Discourse Treebank (Carlson et al., 2001), o Discourse Relations Reference Corpus (Taboada & Renkema, 2008) e o Penn Discourse Treebank (Prasad et al., 2008a) para a língua inglesa; o RST Spanish Treebank[7] (da Cunha Fanego, 2008) para o espanhol; o RST Basque Treebank [8] (Iruskieta et al., 2014) para o basco; o Potsdam Commentary Corpus (Stede, 2004) para o alemão; o Discourse-Annotated Dutch Corpus (Vliet et al., 2011) para o holandês; o CorpusTCC[9] (Pardo & Nunes, 2004), o Rhetalho[10] (Pardo & Seno, 2005), o Summ-it (Collovoni et al., 2007) e o CSTNews[11] (Aleixo & Pardo, 2008; Cardoso et al., 2011) para a língua portuguesa. Nesta tese, o córpus CSTNews foi utilizado e será detalhado na Seção 2.5.1.

Mesmo esta teoria sendo considerada subjetiva, o que pode levar a ambiguidade na identificação da melhor relação retórica a escolher, na segmentação das EDUs e na definição da nuclearidade das EDUs, acreditamos que um texto coerente possui padrões mais recorrentes de estruturação discursiva. Assim, caso um texto de entrada possua um padrão de relações RST diferente dos usuais em textos coerentes, este texto de entrada pode ser menos coerente.

No Apêndice A, todas as definições das relações retóricas identificadas no córpus utilizado nesta tese estão listadas, onde (N) representa o núcleo e (S) o satélite.

2.4.2 *Cross-Document Structure Theory* - CST

Devido ao desejo de identificar as relações entre vários textos, estruturando o discurso de forma a conectar sentenças provenientes de diferentes documentos e estabelecendo um ou mais

[5]http://www.wagsoft.com/RSTTool/
[6]http://www.nilc.icmc.usp.br/dizer2/
[7] http://www.corpus.unam.mx/rst/
[8]http://ixa2.si.ehu.es/diskurtsoa/en/index.php
[9]http://www.icmc.usp.br/ taspardo/CorpusTCC.zip
[10]http://www.icmc.usp.br/ taspardo/rhetalho.zip
[11]http://www.icmc.usp.br/ taspardo/sucinto/cstnews.html

tipos de relações entre elas é que Radev (2000) propôs a CST (*Cross-Document Structure Theory*).

As palavras, sintagmas, orações, sentenças, parágrafos ou blocos de texto ainda maiores podem ser relacionadas por meio das relações CST. Embora orações e sentenças (unidades discursivas) são comumente mais utilizadas nas relações CST, as unidades menores também podem ser relacionadas.

Na Figura 2.9 é mostrado o grafo de relacionamentos entre textos, representado pelas linhas tracejadas, e os subgrafos menores que reproduzem os relacionamentos dentro de cada texto, representado pelas linhas mais grossas e que podem ter relacionamentos RST, sintáticos, etc. Os documentos similares são representados numa hierarquia de palavras, sintagmas, sentenças e os próprios documentos, ou seja, todos esses níveis podem ser considerados na análise. Em cada nível da hierarquia, podem ocorrer relações CST, apesar de sentenças serem usualmente mais utilizadas nos trabalhos da área.

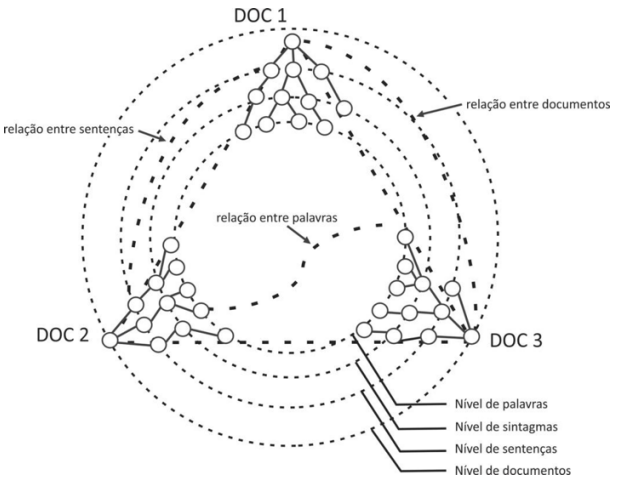

Figura 2.9: Grafo de relacionamentos CST (Radev, 2000, p.5)

Na Tabela 2.5, há o conjunto de relações CST originais (Radev, 2000), sendo que S1 representa Sentença 1 e S2 representa Sentença 2 em documentos diferentes.

Da mesma forma que aconteceu com a RST, os pesquisadores modificaram as 24 relações CST originais. Na língua inglesa, Zhang et al. (2003) verificaram que algumas relações eram ambíguas. Como resultado, os autores propuseram um refinamento para 18 relações. Aleixo & Pardo (2008) aplicaram o conjunto de Zhang et al. (2003) em textos da língua portuguesa. Ainda assim os autores notaram que algumas relações eram redundantes ou ambíguas e sugeriram um novo refinamento para 14 relações. Baseado no refinamento de Aleixo & Pardo (2008), Maziero et al. (2010) determinaram uma tipologia das relações, segundo a Figura 2.10. De acordo com a tipologia, 2 grupos maiores dividem as relações CST: o primeiro grupo abrange as relações cuja finalidade é principalmente relacionar o conteúdo de segmentos e o segundo grupo contêm as relações de apresentação e forma, as quais capturam os estilos de escrita e organização dos textos. Em cada grupo ainda há a divisão por categorias. No grupo de conteúdo, as relações são classificadas em redundância, complemento ou contradição, representando os fenômenos multidocumento. O subgrupo redundância expressa níveis diferentes de sobreposi-

Tabela 2.5: Relações CST

1	*Identity*	O mesmo texto aparece em mais de um local.
2	*Equivalence (paraphrasing)*	Duas sentenças possuem a mesma informação contida.
3	*Translation*	Mesma informação, contida em línguas diferentes.
4	*Subsumption*	S1 contém toda a informação em S2, mais informação adicional que não está em S2.
5	*Contradiction*	S1 e S2 apresentam informação conflitante.
6	*Historical background*	S1 fornecem contexto histórico da informação em S2.
7	*Cross-reference*	A mesma entidade é mencionada.
8	*Citation*	S1 explicitamente cita o documento S2.
9	*Modality*	S1 apresenta uma versão mais qualificada da informação em S2, por exemplo, “é dito que; se sabe que”.
10	*Attribution*	S1 atribui a versão da informação em S2, usando, por exemplo, “de acordo com a CNN”.
11	*Summary*	S1 resume S2.
12	*Follow-up*	S1 apresenta informação adicional que tem acontecido desde S2.
13	*Elaboration*	S2 insere informação adicional a S1.
14	*Indirect speech*	S1 indiretamente menciona algo que foi diretamente mencionado em S2.
15	*Refinement*	S1 fornece detalhes de alguma informação dada de forma mais generalizada em S2.
16	*Agreement*	S1 expressa concordância com S2.
17	*Judgment*	S1 qualifica o fato de S2.
18	*Fulfilment*	S1 afirma a ocorrência de um evento previsto em S2.
19	*Description*	S1 descreve uma entidade mencionada em S2.
20	*Reader profile*	S1 e S2 fornecem a mesma informação, porém escrita para públicos diferentes.
21	*Contrast*	S1 contrasta os fatos ou relatos de S2 ou vice-versa.
22	*Parallel*	S1 compara os fatos ou relatos de S2 ou vice-versa.
23	*Generalization*	S1 generaliza S2 ou vice-versa.
24	*Change of perspective*	A mesma entidade apresenta uma opinião diferente ou apresenta um fato por outro ângulo.

ção de conteúdo e por isso, é dividido em redundância total e redundância parcial. O subgrupo complemento relaciona informações que se complementam e se referem a fatos temporais ou não. O grupo de apresentação e forma é dividido em dois subgrupos: de fonte/autoria e estilo. Para Zhang et al. (2003), os relacionamentos CST são independentes de domínio.

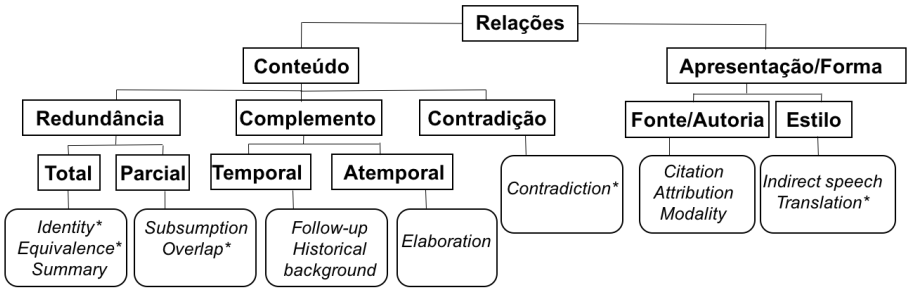

Figura 2.10: Tipologia das relações CST(Maziero et al., 2010)

No Apêndice B, há uma lista de relações CST refinadas, as quais foram identificadas e anotadas no córpus CSTNews (Cardoso et al., 2011; Aleixo & Pardo, 2008) (córpus utilizado nesta tese e que será descrito na Seção 2.5.1). Os nomes das relações foram preservados em inglês.

No exemplo dado na Figura 2.11, as sentenças S1 e S2 podem ser relacionadas pelas relações CST *Contradiction* e *Attribution*. No primeiro caso, há informações contraditórias: S1 diz que a colisão foi no 26o andar e S2 diz que foi no 25o andar. No segundo caso, a relação *Attribution* se deve ao fato de que a informação contida tanto em S1 quanto em S2 está sendo atribuída em S1 a uma jornalista, ou seja, a fonte da informação está sendo identificada.

(S1) A colisão no 26o andar ocorreu às 5:50 p.m. na quinta-feira, disse a jornalista Desideria Cavina.

(S2) O avião colidiu no 25o andar do prédio Pirelli no centro de Milão.

Figura 2.11: Exemplo de identificação de relações CST (Aleixo & Pardo, 2008)

Na CST, algumas relações possuem direcionalidade, como as relações *Attribution*, *Subsumption* e *Historical Background*, entre outras (na Figura 2.10 as relações sem o asterisco possuem direcionalidade). A direcionalidade é dada pelos símbolos - (não há direcionalidade), -> (direcionalidade de S1 para S2) e <- (direcionalidade de S2 para S1). Por exemplo, as duas notícias de diferentes fontes na Figura 2.12 possuem duas relações RST: *Equivalence* e *Attribution* (Aleixo & Pardo, 2008):
Na relação de *Equivalence* não há uma direcionalidade específica, pois tanto S1 é equivalente a

(S1) Um pequeno avião chocou-se em um edifício no centro de Milão, incendiando os últimos andares do prédio, informou uma jornalista italiana da CNN.

(S2) Um pequeno avião chocou-se hoje com um edifício no centro de Milão incendiando vários andares do prédio.

Figura 2.12: Exemplo das relações *Equivalence* e *Attribution*

S2 quanto S2 é equivalente a S1. Entretanto, o mesmo não acontece na relação de *Atributtion*, em que a direcionalidade é de S1 para S2 (->), pois a atribuição do fato é dada a jornalista em S1 e o mesmo não ocorre se houver a troca da direcionalidade entre S2 e S1 (<-).

A relação *Historical Background* é outro exemplo com ambas as direcionalidades. Por exemplo, a Figura 2.13 :

(S1) Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.

(S2) Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.

Figura 2.13: Exemplo da relação *Historical Background*

Nas duas sentenças da Figura 2.13, há a relação *Historical background* com direcionalidade ->, porque é S1 que está trazendo um fato histórico de acidentes. Já nas duas sentenças da Figura 2.14, a sentença S2 é que traz o fato histórico. Portanto sua direcionalidade é <-.

(S1) O prédio da Pirelli em Milão foi atingido por um avião de pequeno porte.

(S2) O prédio foi construído em 1958 e desenhado pelos arquitetos Gio Ponti e Pier Luigi Nervi.

Figura 2.14: Outro exemplo da relação *Historical Background*

Em pesquisas de SA multidocumento, a teoria discursiva CST já foi utilizada nos textos da língua inglesa (Zhang et al., 2002) e da língua portuguesa (Castro Jorge, 2010, 2015; Ribaldo, 2013; Cardoso, 2014). Quanto aos córpus anotados com CST, existem poucos: CSTBank[12] (Radev et al., 2004), para língua inglesa, e CSTNews[13] (Aleixo & Pardo, 2008; Cardoso et al., 2011), para língua portuguesa.

[12]http://clair.si.umich.edu/clair/CSTBank/

[13]http://www.icmc.usp.br/∼taspardo/sucinto/cstnews.html

Assim, a CST e a RST foram fundamentais na criação de modelos de coerência com informações de relações discursivas presentes tanto em sentenças de um mesmo texto, quanto em sentenças que possuem relações intertextuais.

Um grande diferencial deste trabalho é a utilização das relações CST como elementos que auxiliem na distinção entre sumários coerentes e incoerentes. Acredita-se que a distribuição de relações CST em sumários multidocumento considerados coerentes (feitos por humanos) é diferente em sumários multidocumento menos coerente. Partindo desse pressuposto, este trabalho desenvolveu modelos que utilizam essa possível distribuição de relações para distinguir sumários multidocumento coerentes dos incoerentes.

2.4.3 *Centering Theory*

De acordo com a teoria de *Centering*, um discurso deve apresentar coerência na sequência de enunciados que o forma. Um discurso deve exibir coerência local (entre as declarações de um mesmo segmento) e global (entre os seus diversos segmentos). A teoria de *Centering* propõe um modelo para o tratamento da coerência local descrevendo um sistema de restrições e regras que governam as relações entre o foco de atenção do discurso e as formas escolhidas para construção das declarações que o compõem (Grosz et al., 1995).

Para o melhor entendimento dos conceitos e da metodologia empregada pela Teoria *Centering*, dois segmentos discursivos (adaptados de Grosz et al. (1995)) serão utilizados:

1. a. João foi a sua loja de música favorita para comprar um piano.

   b. Ele havia frequentado a loja por vários anos.

   c. Estava excitado porque iria finalmente poder comprar um piano.

   d. Mas quando chegou, a loja estava fechada.

2. a) João foi a sua loja de música favorita para comprar um piano.

   b) Esta era a loja que João frequentou por vários anos.

   c) Ele estava excitado porque iria finalmente poder comprar um piano.

   d) Ela estava fechada quando João chegou.

Observando os dois segmentos de discurso, os mesmos expressam a mesma informação com enunciados diferentes. Apesar disso, o discurso (1) é intuitivamente mais coerente que o discurso (2). Isso parece acontecer porque no segmento discursivo (1) foca-se apenas de um indivíduo central, “João”, enquanto que no (2), o foco principal oscila entre “João” e “a loja de música”. Isto mostra que diferentes estruturações sintáticas implicam em diferenças na inferência dos referentes anafóricos para o receptor (leitor/ouvinte). A teoria de *Centering* fornece elementos para o tratamento destas diferenças.

Nomenclatura

Para Grosz et al. (1995), o termo centro de uma sentença são todas as entidades referidas pela sentença que a liga a outra sentença no segmento discursivo que as contêm. Os mesmos enunciados presentes em diferentes situações discursivas podem ter diferentes centros. Desta forma, centros são construções discursivas; mais especificamente objetos semânticos, frases ou formas sintáticas.

Um segmento de discurso consiste de uma sequência de enunciados U_1, *U2,...*, U_n. Os enunciados possuem a propriedade de *realizar* entidades do contexto do discurso. Por exemplo, no enunciado "João foi a sua loja de música favorita comprar um piano", têm-se as entidades realizadas JOÃO, LOJA-DE-MUSICA e PIANO.

Para cada enunciado U_m é associado um conjunto ordenado de "centros prospectivos" (*Forward-looking centers*), $C_f(U_m)$, consistindo das entidades do discurso que são realizadas por este enunciado.

A ordem dos elementos de $C_f(U_m)$ segue o seguinte critério: para todo f_i, $f_j \in C_f(U_m)$, se f_i realiza um sujeito e f_j realiza um objeto, então $f_i \prec f_j$, ou seja, f_i precede (tem mais importância) que f_j. O primeiro elemento de $C_f(U_m)$ é chamado "próximo centro preferencial" (*Prefered Center*), $C_p(U_m)$.

O "centro retrospectivo" (*Backward-looking Center* - $C_b(U_m)$) estabelece uma relação coerente com o enunciado imediatamente anterior, U_{m-1}, desde que o enunciado corrente (U_m) não seja o primeiro segmento, isto é, $C_b(U_m)$ = {vazio}. O exemplo (3) mostra os termos introduzidos:

3. a) Marco$_{a1}$ possui um helicóptero$_{a2}$.

   b) Ele$_{r1}$ pilota bem.

   c) Luciano$_{a3}$ viaja com ele$_{r2}$ a trabalho$_{a4}$.

   d) Ele$_{r3}$ normalmente o$_{r4}$ solicita.

De acordo com o exemplo, as entidades do mundo são reconhecidas por meio dos substantivos que as descrevem: Marco, helicóptero, Luciano e trabalho. Os índices que aparecem nos enunciados (*a1, a2, a3 e a4*) nomeiam as construções do segmento de discurso que referenciam as entidades. Da mesma forma, nomeiam-se os elementos anafóricos encontrados (*r1, r2, r3 e r4*). Com a Identificação completada, pode-se construir a "âncora", o par <C_b, C_f>, de cada enunciado. As "âncoras" verificam as relações de centros de atenção dos enunciados, ou seja, as relações entre as entidades (substantivos) dos enunciados. Um exemplo da representação do discurso por "âncora" é dado a seguir:

4. a) <*(?), [(Marco, a_1),(helicóptero, a_2)]*>

   b) <*(Marco, a_1), [(Marco, r_1)]*>

   c) <*(Marco, r_1), [(Luciano, a_3),(Marco, r_2),(trabalho, a_4)]*>

   d) <*(Luciano, a_3), [(Luciano, r_3),(Marco, r_4)]*>

O C_b indica quem é a atual entidade central do discurso. No conjunto, C_f aponta quais entidades foram realizadas por quais elementos. O par (Luciano, r_3) da âncora (d), por exemplo, indica que a entidade "Luciano" foi realizada pelo elemento r_3. A primeira âncora (a) apresenta um (?) como C_b, pois a teoria não define como escolher o C_b do primeiro enunciado do segmento de discurso.

Transições

A teoria de *Centering* possui outro conceito importante, que são as transições entre os enunciados. Elas descrevem como estes são ligados em um segmento de discurso coerente. Brennan et al. (1987) propuseram quatro tipos. São eles:

1. *Continue*: $C_b(U_{m-1}) = C_b(U_m) = C_p(U_m)$.

   O discurso permanece centrado na mesma entidade e esta é o centro preferido a ser usado na próxima sentença.

2. *Retain*: $C_b(U_{m-1}) = C_b(U_m) \neq C_p(U_m)$.

   O discurso permanece centrado na mesma entidade, mas na próxima sentença um novo centro será o preferido.

3. *Smooth-shift*: $C_b(U_{m-1}) \neq C_b(U_m) = C_p(U_m)$.

   O discurso trocou de centro e este novo centro é o preferido para ser usado na próxima sentença.

4. *Rough-shift:*: $C_b(U_{m-1}) \neq C_b(U_m) \neq C_p(U_m)$.

   O discurso trocou de centro e outro centro será o preferido a ser usado na próxima sentença.

Para exemplificar os tipos de transições, um exemplo é mostrado a seguir:

5. a) Roberto$_{a1}$ é um ator.

   <(?), [(roberto, a_1)]>

   b) Ele$_{r1}$ visitou Cláudia$_{a2}$ ontem.

   <(roberto, a_1), [(roberto, r_1),(claudia, a_2)]>

   ci) Ele$_{r2}$ conversou muito com ela$_{r3}$

   Continue: <(roberto, r_1), [(roberto, $_{r2}$),(claudia, $_{r3}$)]>

   cii) Ela$_{r2}$ recebeu a visita dele$_{r3}$ entusiasmada.

   Retain: <(roberto, $_{r1}$), [(claudia, $_{r2}$),(roberto, $_{r3}$)]>

   ciii) Ela$_{r2}$ não gostou.

   Smooth-shift: <(claudia, $_{a2}$), [(claudia, $_{r2}$)]>

civ) Julia$_{a3}$ a$_{r2}$ viu na semana passada.

Rough-shift: <(claudia, a_2), [(julia, a_3),(claudia, $_{r2}$)]>

Além das transições, a teoria *Centering* apresenta um conjunto de restrições e regras quanto a forma como os centros de atenção podem ser utilizados para a composição de um texto coerente:

Restrições

1. Existe apenas um C_b para cada enunciado.
2. Todos elementos de $C_f(U_m)$ são realizados em U_m
3. $C_b(U_m)$ é o mais bem colocado elemento de $C_f(U_{m-1})$ que é realizado em U_m.

Regras

1. Se $f_j \in C_f(\mathrm{U}_{m-1})$ e f_j é realizado por um pronome em U_m, assim todo $f_i \in C_f(\mathrm{U}_{m-1})$ realizado em U_m tal que $f_i \prec \mathrm{f}_j$ em $C_f(\mathrm{U}_{m-1})$, deve ser realizado por um pronome em U_m. Isto implica que se existe um pronome na sentença, então C_b é realizado por um pronome.
2. *Continue* tem preferência sobre *Retain* que tem preferência sobre *Smooth-shift* o qual tem preferência sobre *Rough-shift*.

A teoria de *Centering* inspirou modelos computacionais voltados para a avaliação da coerência textual, devido a premissa de que essa teoria reconhece que a distribuição de entidades em textos coerentes localmente exibe certa regularidades (padrão de distribuição de entidades). Tal premissa pode ajudar na distinção de textos coerentes dos incoerentes.

2.5 Recursos e Ferramentas Linguístico-Computacionais

Nessa Seção, os principais recursos e ferramentas que foram utilizados nesta tese serão introduzidos.

2.5.1 Córpus CSTNews

Os trabalhos de PLN geralmente necessitam de uma amostra da língua (escrita ou falada) para estudar e buscar soluções automáticas para um fenômeno específico da língua. Neste caso, um conjunto de textos (córpus) que possua um fenômeno da língua a ser estudado é formado. Portanto, para o estudo da coerência textual e o desenvolvimento de modelos que automaticamente avaliem a coerência em sumários multidocumento, o córpus CSTNews foi utilizado.

O córpus CSTNews foi o primeiro experimento de identificação de relações CST para o Português, o qual fez parte de um projeto de mestrado que visava à construção de um analisador discursivo multidocumento automático para o Português do Brasil (Aleixo & Pardo, 2008).

Esse córpus possui 50 coleções de textos jornalísticos de domínios variados e cada coleção possui em média 3 documentos de diferentes fontes que abordam um mesmo assunto. Dados como o número exato de documentos por domínio, número de sentenças e palavras por coleção são visualizados na Tabela 2.6.

Segundo Aleixo & Pardo (2008), os textos foram coletados manualmente das páginas das agências de notícias na *web* por um período de 2 meses, entre Agosto e Setembro de 2007. As fontes dos textos foram os jornais *on-line*: Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. A escolha desses jornais foi devido a popularidade que os mesmos possuem na *web* e também por trazerem as principais notícias do dia corrente, que é o que importa para o córpus multidocumento, ou seja, uma mesma notícia publicada em fontes diferentes. Os textos jornalísticos foram escolhidos por possuírem uma linguagem clara e do dia a dia, além da facilidade de serem encontrados na *web*.

Além dos textos originais que cada um dos agrupamentos do córpus CSTNews possui, há também a identificação de expressões temporais, anotação RST, anotação CST, a segmentação dos textos originais, análise sintática e mais recentemente a anotação de aspectos internativos nos sumários humanos. O córpus ainda possui sumários feitos por humanos e de forma automática para cada um dos agrupamentos. E é por ter essas informações contidas no córpus CSTNews que esse córpus foi utilizado neste trabalho.

A anotação das relações CST foi realizada por uma equipe de 4 linguistas computacionais. A tarefa de anotação foi realizada em duas etapas: o treinamento e a anotação de fato. A etapa de treinamento durou aproximadamente três meses, período em que os anotadores estudaram a teoria e experimentaram a anotação de alguns textos não pertencentes ao córpus.

Para a anotação do córpus, uma ferramenta semi-automática foi utilizada. Tal ferramenta é chamada de CSTTool (Aleixo & Pardo, 2008).

Para medir a concordância para a tarefa de anotação das relações CST foi utilizada a medida *Kappa* de Carletta (1996). *Kappa* (K) é uma medida clássica de concordância usada em PLN, a qual depende da tarefa e que indica a correlação entre anotadores enquanto ela desconta a concordância aleatória ou sorte. A equação 2.1, apresenta-se a fórmula da medida *Kappa*, onde *P(A)* é a proporção de vezes que os anotadores concordam e *P(E)* é a proporção de vezes que os anotadores concordam ao acaso. Apesar de não existir um valor específico a partir do qual se deva considerar o valor da *Kappa* como adequado, encontram-se na literatura algumas sugestões que orientam esta decisão: valores menores do que 0,4 indicam uma anotação não confiável; se o valor de Kappa estiver entre 0,4 e 0,75, a anotação é satisfatória; e se o valor de Kappa for maior do que 0,75, a anotação é considerada muito confiável.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.1}$$

Para a tarefa de anotação das relações CST, a medida de concordância *Kappa* foi calculada

Tabela 2.6: Dados do CSTNews

Coleção	Domínio	Qt. de documentos	Qt. de sentenças	Qt. de palavras
C1	Mundo	3	24	432
C2	Política	3	51	996
C3	Cotidiano	3	50	1243
C4	Cotidiano	3	39	832
C5	Cotidiano	2	23	572
C6	Cotidiano	3	36	925
C7	Ciência	2	23	585
C8	Esportes	3	25	593
C9	Política	3	36	965
C10	Mundo	3	39	964
C11	Cotidiano	3	56	987
C12	Mundo	3	37	960
C13	Mundo	3	37	962
C14	Mundo	3	25	739
C15	Mundo	3	26	565
C16	Política	3	47	1031
C17	Política	2	41	963
C18	Mundo	3	70	1301
C19	Esportes	2	13	298
C20	Política	3	42	949
C21	Cotidiano	3	41	870
C22	Cotidiano	3	50	964
C23	Mundo	2	25	572
C24	Esportes	3	24	541
C25	Esportes	3	88	1561
C26	Mundo	3	58	1406
C27	Esportes	3	89	1543
C28	Esportes	3	35	717
C29	Mundo	3	48	1167
C30	Dinheiro	3	46	1131
C31	Esportes	2	10	217
C32	Mundo	3	66	1328
C33	Cotidiano	3	68	1638
C34	Cotidiano	3	59	1139
C35	Mundo	3	36	876
C36	Cotidiano	3	74	1357
C37	Cotidiano	2	26	475
C38	Esportes	3	26	535
C39	Cotidiano	3	35	914
C40	Política	3	28	746
C41	Esportes	3	45	958
C42	Política	2	39	1061
C43	Política	3	49	1267
C44	Política	2	26	719
C45	Cotidiano	3	47	1223
C46	Mundo	3	38	740
C47	Mundo	3	43	1373
C48	Esportes	2	43	800
C49	Cotidiano	3	23	1001
C50	Política	3	63	1546
Total de documentos		140		
Total de sentenças			2.088	
Total de palavras				47.247

levando em consideração três aspectos: concordância das relações utilizadas, concordância sobre as direcionalidades das relações e concordância das relações agrupadas (relações que pertencem a uma mesma categoria de acordo com a tipologia de Maziero et al. (2010)). A Tabela 2.7 mostra os resultados da medida Kappa para a tarefa de anotação das relações CST.

Tabela 2.7: Kappa para a tarefa de anotação CST para o córpus CSTNews

Aspectos de Concordância	**Kappa**
Relações	0,50
Direcionalidade	0,44
Relações agrupadas	0,61

Seguindo as sugestões da literatura, os valores de kappa da tarefa de anotação das relações CST mostrados na Tabela 2.7 são satisfatórios nos três aspectos considerados.

Além da medida *Kappa*, também foi utilizada uma medida de porcentagem direta para avaliar a concordância. Com esta medida são avaliados três tipos de concordância: a concordância total (quando todos os anotadores indicam a mesma relação, direcionalidade ou relações agrupadas), concordância parcial (quando a maioria dos anotadores indicam a mesma relação, direcionalidade ou relações agrupadas) e concordância nula (quando nenhum dos anotadores indicam a mesma relação, direcionalidade ou relações agrupadas). A Tabela 2.8 mostra a medida de porcentagem de concordância obtida no córpus CSTNews.

Tabela 2.8: Porcentagem de concordância no córpus CSTNews

Aspectos de Concordância	**Concordância Total (%)**	**Concordância Parcial (%)**	**Concordância Nula (%)**
Relações	54	27	18
Direcionalidade	58	27	14
Relações agrupadas	70	21	9

De acordo com a Tabela 2.8, 81% dos anotatores concordaram de forma parcial ou total com as relações, 85% concordaram parcialmente ou totalmente com a direcionalidade e 91% concordaram parcialmente ou totalmente com as relações agrupadas. Tais resultados mostraram-se melhores que os resultados obtidos por Zhang et al. (2002), que obtiveram apenas 58% de concordância parcial ou total das relações, para textos anotados da língua inglesa.

As relações de RST foram anotadas no córpus CSTNews por 8 anotadores, sendo que 4 deles tinham um conhecimento mais profundo da teoria RST e mais experiência na anotação. A tarefa de anotação foi realizada em duas etapas: treinamento e a anotação de fato.

A anotação dos textos foi realizada incrementalmente, isto é, os segmentos dentro das sentenças eram anotados inicialmente, logo, em seguida, as sentenças adjacentes dentro de um parágrafo eram anotadas e, finalmente, os parágrafos adjacentes eram anotados. A concordância entre os anotadores foi calculada usando a ferramenta RSTeval (Maziero & Pardo, 2009).

A RSTeval baseia-se na comparação de duas ou mais árvores retóricas para um mesmo texto. Para esta comparação, uma das árvores correspondentes ao texto é selecionada como "ideal" e as outras árvores são comparadas a essa árvore "ideal" com base nos seguintes elementos:

- Segmentos textuais simples;
- Segmentos textuais mais complexos (por exemplo, dois ou mais segmentos ligados por uma mesma relação);
- Nuclearidade de cada segmento;
- Relação RST entre segmentos.

As medidas de Precisão, Cobertura e Medida-F são calculadas para cada um dos elementos listados acima, em cada uma das árvores RST e, deste modo, determinar o quão similares são as árvores. A medida de Precisão (P) indica o número de elementos corretos (Corr) de uma árvore T (em comparação com a árvore "ideal"), dividido pelo numero total de elementos da árvore T (ver equação 2.2).

$$P = \frac{Corr}{|T|} \tag{2.2}$$

A medida de Cobertura (C) indica o número de elementos corretos (Corr) da árvore T, dividido pelo número de elementos da árvore "ideal" I (ver equação 2.3).

$$C = \frac{Corr}{|I|} \tag{2.3}$$

A Medida-F representa a média harmônica entre a Precisão e a Cobertura (ver equação 2.4).

$$Medida - F = \frac{2PC}{P + C} \tag{2.4}$$

A Tabela 2.9 mostra a concordância obtida na anotação RST do córpus CSTNews, usando as medidas descritas acima.

Tabela 2.9: Concordância para a tarefa de anotação RST para o córpus CSTNews

Elemento	**Precisão (%)**	**Cobertura (%)**	**Medida-F (%)**
Segmento simples	0,91	0,91	0,91
Segmento complexo	0,78	0,78	0,78
Núcleo	0,78	0,78	0,78
Relação RST	0,66	0,66	0,66

2.5.1.1 Metodologia de Criação de Novos Sumários para o CSTNews

Para que o córpus tivesse uma boa quantidade de sumários de referência para subsidiar esta e futuras pesquisas (já que originalmente o córpus CSTNews possuía apenas 1 sumário

multidocumento de referência para cada coleção de textos), foi conduzida a produção de mais 5 extratos e 5 *abstracts* para cada coleção de textos do CSTNews, totalizando 250 extratos e 250 *abstracts*.

Para tal finalidade, 20 pesquisadores de PLN (alunos e docentes das áreas da Linguística e da Ciência da Computação) foram reunidos, sendo que cada pesquisador teria a incumbência de produzir 25 sumários entre extratos e *abstracts*. A atribuição das coleções e do tipo de sumário a cada um dos pesquisadores foi feita de forma balanceada, já que cada coleção do CSTNews possui diferentes tamanhos.

A criação de sumários foi realizada diariamente, sendo que, a cada dia, os pesquisadores deveriam criar dois sumários, um extrato e um *abstract* de coleções diferentes, com o intuito de deixarem os sumários tão diversificados quanto possível na sua construção.

A tarefa foi realizada em aproximadamente 1 mês, sendo que, inicialmente, foi realizada uma reunião com todos os pesquisadores para que as instruções para a realização da tarefa fossem passadas e explicadas. Como não havia necessidade de reunir todos os pesquisadores no mesmo local para a criação dos sumários, já que a mesma necessitava apenas da subjetividade de cada pesquisador e de sua capacidade de resumir, além disso foi decidido que os sumários poderiam ser feitos a distância, desde que a entrega fosse feita por e-mail em, no máximo, 24 horas depois do prazo estipulado para cada coleção de textos. Esse tipo de restrição é importante para manter o comprometimento dos participantes e o controle sobre os prazos da tarefa.

Algumas restrições, em relação a tarefa, deveriam ser respeitas por todos os pesquisadores para manter a uniformidade dos sumários. Uma delas foi a limitação de tamanho dos sumários, já que, nesta tarefa, utilizou-se uma taxa de compressão de 70% em relação ao tamanho do maior texto da coleção em análise. Por exemplo, a coleção 23 do CSTNews possui 2 textos e o maior deles possui 405 palavras. Com a taxa de compressão de 70% sobre o maior texto, os extratos e os *abstracts* dessa coleção deveriam ter aproximadamente 122 palavras. Foi permitida uma tolerância de 10 palavras para mais ou para menos em relação ao tamanho especificado. Assim, para a coleção 23, os pesquisadores poderiam criar sumários com tamanhos que poderiam variar de 112 a 132 palavras.

Outra restrição importante foi que cada pesquisador deveria evitar ao máximo copiar qualquer parte do texto fonte quando o sumário em foco era do tipo *abstract*. No caso do extrato, os sumarizadores tiveram que selecionar sentenças completas para formar o sumário, incluindo, ao final de cada uma, sua identificação de origem, isto é, sua numeração no texto fonte. Essa identificação já estava associada a cada sentença de todos os textos fornecidos aos pesquisadores. Tal identificação, ajudou esta pesquisa e ajudará os pesquisadores na recuperação de informações presentes no córpus CSTNews, caso necessário.

A Tabela 2.10 mostra para cada coleção (Col.): (i) os tipos de sumários (TS) construídos; (ii) o tamanho médio (TM) em número de palavras dos sumários obtidos; (iii) a variação da quantidade de palavras (VP) utilizadas pelos pesquisadores (que corresponde à diferença de tamanho entre o maior e o menor sumário); (iv) a quantidade de sentenças (QS) dos textos-fonte que mais foram utilizadas na construção dos extratos de sua respectiva coleção; (v) a porcenta-

gem de extratos (%Ext) em que ocorrem a(s) sentença(s) de maior uso (dada pela coluna QS); (vi) o número médio de sentenças dos sumários (NMS); e (vii) a variação da quantidade de sentenças (VS) utilizadas pelos pesquisadores (também correspondente à diferença entre o maior e o menor sumário).

Tabela 2.10: Dados dos sumários criados

Col	TS	TM	VP	QS	%Ext	NMS	VS
C1	*Abstract*	58	13	-	-	3,2	1
	Extrato	57	9	1	60	3	2
C2	*Abstract*	128	12	-	-	5,6	1
	Extrato	129	12	8	40	5,2	2
C3	*Abstract*	182	14	-	-	8,4	3
	Extrato	174	12	2	80	7,8	2
C4	*Abstract*	106	17	-	-	4,6	3
	Extrato	105	8	1	100	4	0
C5	*Abstract*	132	14	-	-	5,8	3
	Extrato	130	16	1	100	4,4	1
C6	*Abstract*	108	10	-	-	4	2
	Extrato	107	12	3	60	5	2
C7	*Abstract*	116	18	-	-	4,8	2
	Extrato	117	7	3	60	4,2	1
C8	*Abstract*	78	14	-	-	4	2
	Extrato	78	10	1	60	3,2	1
C9	*Abstract*	127	12	-	-	4,6	3
	Extrato	130	14	1	60	4,8	4
C10	*Abstract*	142	18	-	-	7,2	3
	Extrato	138	11	1	80	4,6	1
C11	*Abstract*	161	16	-	-	6,6	3
	Extrato	160	17	5	60	8,4	4
C12	*Abstract*	102	15	-	-	4,6	2
	Extrato	106	10	1	60	3,8	2
C13	*Abstract*	109	14	-	-	4,8	2
	Extrato	111	9	1	80	4,2	1
C14	*Abstract*	86	18	-	-	4,8	3
	Extrato	77	14	1	80	2,4	1
C15	*Abstract*	83	19	-	-	4,4	1
	Extrato	78	15	1	60	3,6	1
C16	*Abstract*	154	9	-	-	6,6	5
	Extrato	151	11	2	80	6,4	1
C17	*Abstract*	175	19	-	-	6,6	3
	Extrato	177	15	3	80	6,6	3
C18	*Abstract*	206	10	-	-	11,4	5
	Extrato	202	17	1	80	9,8	4
C19	*Abstract*	53	11	-	-	3,2	1
	Extrato	49	2	1	100	2	0
C20	*Abstract*	138	15	-	-	6,6	3
	Extrato	133	14	1	100	5,2	2
C21	*Abstract*	146	20	-	-	7,4	3
	Extrato	145	9	2	80	6,6	1
C22	*Abstract*	119	13	-	-	6,6	5
	Extrato	113	9	1	80	5,8	4
C23	*Abstract*	128	5	-	-	6,2	4
	Extrato	123	14	1	100	4	0
C24	*Abstract*	88	10	-	-	4	2
	Extrato	83	9	2	60	3,6	1
C25	*Abstract*	169	20	-	-	9,4	3
	Extrato	170	13	4	80	8,2	2
C26	*Abstract*	182	18	-	-	8,8	2
	Extrato	178	17	4	60	7,2	3
C27	*Abstract*	191	18	-	-	9,4	3
	Extrato	182	17	1	60	9	5
C28	*Abstract*	86	13	-	-	4	2
	Extrato	88	16	1	60	3,4	1
C29	*Abstract*	154	8	-	-	6,4	4
	Extrato	145	9	7	40	5,2	2
C30	*Abstract*	134	10	-	-	6,2	5
	Extrato	131	14	2	60	4	2
C31	*Abstract*	46	9	-	-	2,4	1
	Extrato	45	10	1	100	2	0
C32	*Abstract*	153	11	-	-	7,2	3
	Extrato	152	16	2	80	8,6	3
C33	*Abstract*	285	12	-	-	10,8	9
	Extrato	282	16	2	80	11	9
C34	*Abstract*	164	14	-	-	8	5
	Extrato	162	20	2	80	6,8	2
C35	*Abstract*	134	18	-	-	7	3
	Extrato	132	9	3	60	5	0
C36	*Abstract*	227	20	-	-	14	7
	Extrato	228	11	4	60	11,8	4
C37	*Abstract*	82	13	-	-	5,6	2
	Extrato	85	13	2	80	4	2
C38	*Abstract*	89	11	-	-	4,2	3
	Extrato	84	17	1	80	3,2	1
C39	*Abstract*	95	18	-	-	3,4	2
	Extrato	93	19	2	40	2,6	1
C40	*Abstract*	108	14	-	-	4,2	1
	Extrato	101	18	1	40	4,6	1
C41	*Abstract*	131	18	-	-	5,8	5
	Extrato	137	13	1	100	5,4	2
C42	*Abstract*	186	19	-	-	7,4	4
	Extrato	184	14	4	60	5,8	1
C43	*Abstract*	168	18	-	-	7,2	2
	Extrato	167	13	1	80	5,2	3
C44	*Abstract*	156	15	-	-	7,6	5
	Extrato	156	15	1	100	6	2
C45	*Abstract*	161	15	-	-	8,6	4
	Extrato	168	8	1	80	7,2	2
C46	*Abstract*	89	10	-	-	6,4	2
	Extrato	82	13	1	100	3,4	1
C47	*Abstract*	162	15	-	-	7	3
	Extrato	158	10	5	40	4,2	1
C48	*Abstract*	135	20	-	-	6,8	4
	Extrato	133	18	1	100	6,6	2
C49	*Abstract*	127	18	-	-	5	2
	Extrato	131	11	3	60	5	4
C50	*Abstract*	186	16	-	-	8,8	3
	Extrato	181	10	7	40	7,2	2
Média	***Abstract***	**134**	**14,5**	**-**	**-**	**6,3**	**3,1**
	Extrato	**133**	**12,4**	**2,2**	**72**	**5,4**	**2**

De acordo com a Tabela 2.10, em 35 coleções, o tamanho médio dos abstracts foi maior do que os extratos. Tal resultado advém da maior liberdade na construção dos abstracts. Entretanto, na média geral, o tamanho dos abstracts foi similar ao dos extratos. Devido a tolerância de 10 palavras no tamanho dos sumários, calculamos a variação média do tamanho dos sumários. Podemos observar que houve variação alta de tamanho tanto para abstracts quanto para extratos (conforme coluna VP na tabela). Tal dado mostra a importância do uso da tolerância no tamanho dos sumários, principalmente para a criação dos extratos (mais restrição do que os

abstracts), já que a maioria dos sumários extrativos tiveram seus tamanhos acima da taxa de compressão utilizada em cada coleção. Isso mostra que houve uma certa dificuldade por parte dos pesquisadores em produzir bons extratos informativos dentro de um espaço reduzido.

Observa-se, nas colunas QS e %Ext, que todos os extratos produzidos para 10 coleções tiveram 1 sentença em comum, e, na maioria desses casos, foi a primeira sentença de um dos textos-fonte de suas respectivas coleções. Vê-se também que não há casos de 2 ou mais sentenças em comum em todos os sumários. Há também extremos, em que 7 ou 8 sentenças são comuns a uma parcela (não todos) dos sumários (veja, por exemplo, as coleções 2 e 50). Esses dados indicam que grande parte da informação principal estava contida no início do texto-fonte e foi utilizada para compor o extrato.

Outro dado interessante, representado pela coluna NMS, é que a maioria das coleções (42) tiveram os abstracts com uma média de sentenças superior aos extratos (6), e em apenas 2 coleções o número médio de sentenças tanto dos abstracts quanto dos extratos foi igual. O comportamento é similar quando analisamos a variação do número de sentenças (coluna VS), sendo a coleção 33 a que teve a maior variação de sentenças para abstracts e extratos (variação de 9 sentenças). Esses dados já eram esperados, uma vez que os pesquisadores tiveram uma liberdade maior na criação dos abstracts, consequentemente podendo produzir sentenças mais curtas e com altas variações na quantidade das mesmas entre os sumarizadores.

Com a criação dos novos sumários, cada coleção de textos do CSTNews contém, agora, 6 sumários abstrativos e 6 sumários extrativos, o que constitui um aumento significativo na quantidade de dados de referência em relação ao que se tinha anteriormente. Esses dados devem subsidiar novas pesquisas na área de SAM e permitiu que os modelos desenvolvidos nesta pesquisa fossem melhor treinados e, consequentemente, resultados melhores alcançados.

2.5.2 *Parser* Palavras

Considerado um dos melhores analisadores sintáticos automáticos para o Português do Brasil, o PALAVRAS foi desenvolvido por Bick (2000).

O PALAVRAS realiza análise morfossintática e sintática, ou seja, para cada palavra ele determina a sua classe morfológica e seu papel sintático. Tal análise é feita pelo analisador léxico-morfológico denominado de PALMORF. Este analisador é composto por dois módulos: o pré-processador e o analisador morfológico.

O pré-processador é responsável por identificar e resolver os seguintes fenômenos da linguagem: palavras compostas, letras maiúsculas, ênclise (colocações pronominais depois de verbos), abreviações. Já o analisador morfológico é responsável pela identificação de lexemas, flexões, derivações, incorporação de verbos, hifenização, aspas, além de outras funções.

O PALAVRAS possui três formatos de saída. O primeiro formato faz uso de uma representação gráfica de árvore, que representa a estrutura do texto. As folhas das árvores são compostas pelos componentes sentenciais e os nós internos da árvore representam a análise sintática da sentença. O segundo formato do PALAVRAS é uma versão textual da primeira representação, ou seja, traz as mesmas informações geradas na árvore. Já o terceiro formato é um arquivo XML

(*eXtensible Markup Language*) no padrão de anotação *Tiger* (Brants & Hansen, 2002). A Figura 2.15 mostra o formato do padrão de anotação *Tiger* utilizado pelo *parser* PALAVRAS para a realização da análise morfosintática. Neste formato, os dados estão modelados em grafos de sintaxe, ou seja, grafos direcionados acíclicos com uma única raíz. Inicialmente, cada sentença "s" do texto possui um identificador (id = "s10"), a sua referência (ref = "10") e a sua descrição no campo *text*. Os marcadores do tipo "*terminals*" são as folhas da representação gráfica de árvore, os quais possuem palavras, etiquetas morfossintáticas, etiquetas morfológicas e lemas como atributos. Já os elementos não terminais são os nós internos da representação gráfica da árvore. Esses elementos são os marcadores denominados "*nonterminals*", que apontam para os terminais correspondentes por meio de identificadores (*idref*) para recuperar informações contidas nos *terminals*.

```
<s id="s10" ref="10" source="Running text" forest="1" text="Então, o critério é um critério técnico.">
    <graph root="s10_500">
        <terminals>
            <t id="s10_1" word="Então" lemma="então" pos="adv" morph="--" sem="--" extra="kc"/>
            <t id="s10_2" word="," lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
            <t id="s10_3" word="o" lemma="o" pos="art" morph="M S" sem="--" extra="--"/>
            <t id="s10_4" word="critério" lemma="critério" pos="n" morph="M S" sem="ac" extra="--"/>
            <t id="s10_5" word="é" lemma="ser" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="fmc mv"/>
            <t id="s10_6" word="um" lemma="um" pos="art" morph="M S" sem="--" extra="--"/>
            <t id="s10_7" word="critério" lemma="critério" pos="n" morph="M S" sem="ac" extra="--"/>
            <t id="s10_8" word="técnico" lemma="técnico" pos="adj" morph="M S" sem="--" extra="np-close"/>
            <t id="s10_9" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
        </terminals>

        <nonterminals>
            <nt id="s10_500" cat="s">
                <edge label="STA" idref="s10_501"/>
            </nt>
            <nt id="s10_501" cat="fcl">
                <edge label="fA" idref="s10_1"/>
                <edge label="PU" idref="s10_2"/>
                <edge label="S" idref="s10_502"/>
                <edge label="P" idref="s10_5"/>
                <edge label="Cs" idref="s10_503"/>
                <edge label="PU" idref="s10_9"/>
            </nt>
            <nt id="s10_502" cat="np">
                <edge label="DN" idref="s10_3"/>
                <edge label="H" idref="s10_4"/>
            </nt>
            <nt id="s10_503" cat="np">
                <edge label="DN" idref="s10_6"/>
                <edge label="H" idref="s10_7"/>
                <edge label="DN" idref="s10_8"/>
            </nt>
        </nonterminals>
    </graph>
</s>
```

Figura 2.15: Exemplo da análise feita pelo *parser* PALAVRAS.

Assim, como um analisador sintático automático faz parte da estrutura da maioria dos modelos desta pesquisa, o PALAVRAS foi o analisador escolhido para realizar o processamento morfossintático nos sumários do córpus CSTNews. Esta escolha foi paltada na qualidade desse analisador, pois segundo Bick (2000), o PALAVRAS possui 97% de precisão nas marcações corretas das etiquetas sintáticas.

Capítulo

3

Trabalhos Relacionados

Neste capítulo serão descritos os principais trabalhos relacionados a distinção automática de textos coerência dos incoerentes. Tais trabalhos são baseados em 3 diferentes abordagens: (i) entidades; (ii) discurso; e (iii) estatística/matemática.

3.0.1 Trabalhos Baseados em Entidades

Os trabalhos baseados em entidades utilizam de forma direta ou indireta a distribuição de entidades de cada texto. Essa distribuição é utilizada na distinção de textos coerentes dos incoerentes.

O modelo de Grade de Entidades foi desenvolvido por Barzilay & Lapata (2005) e aprimorado em Barzilay & Lapata (2008), o qual captura o relacionamento textual por meio de transições entre sentenças.

A hipótese deste modelo é que a distribuição de entidades em um texto coerente localmente mostra uma certa regularidade. De acordo com as autoras, essa hipótese não é arbitrária, ou seja, algumas dessas regularidades têm sido reconhecidas na teoria de *Centering* de Grosz et al. (1995) e em teorias baseadas em entidades de discurso, como a de Givon (1987) e a de Prince (1981).

Segundo Barzilay & Lapata, a representação do discurso baseada em entidades permite aprender as propriedades de textos coerentes de um córpus, sem utilizar o recurso da anotação manual ou uma base de conhecimento predefinida. Para demonstrar a utilidade dessa representação, o seu poder preditivo foi testado em três experimentos: 1) ordenação textual, 2) avaliação automática de sumários coerentes e 3) avaliação da legibilidade.

O trabalho de Barzilay & Lapata permite automaticamente, embora com algum ruído, extrair atributos (*features*), que permitam executar uma avaliação em larga escala de diferentes modelos de coerência instanciados de forma diferente através de gêneros e aplicações.

Cada texto é representado por uma **Grade de Entidades** representada por uma matriz bidimensional que captura a distribuição das entidades discursivas nas sentenças do texto (unidades de análise). As linhas da grade correspondem às sentenças e as colunas correspondem às entidades discursivas. Para cada ocorrência de uma entidade no texto, a célula da grade correspondente conterá informações sobre a presença ou ausência em uma sentença. As entidades presentes em uma dada sentença terão informações sobre seu papel sintático - Sujeito (S), Objeto (O), ou Nenhum dos dois papéis anteriores (X), caso a informação sintática seja considerada. Além disso, a falta de entidades em uma sentença é repesentada na grade de entidades por um traço (-). A informação sobre o papel gramatical é obtido por meio de um *Parser* Sintático. Caso a informação sintática não esteja disponível e a entidade esteja presente em uma determinada sentença, a representação disso na sua respectiva célula da grade será dada pelo caractere X.

Um exemplo dessa representação é dada pela Figura 3.1, a qual representa um fragmento de uma grade de entidades para o texto da Figura 3.2.

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

Figura 3.1: Fragmento de uma grade de entidades (Barzilay & Lapata, 2008, p. 6)

1 [The Justice Department]$_S$ is conducting an [anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$.
2 [Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$.
3 [The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$.
4 [Microsoft]$_S$ claims [its tactics]$_S$ are commonplace and good economically.
5 [The government]$_S$ may file [a civil suit]$_O$ ruling that [conspiracy]$_S$ to curb [competition]$_O$ through [collusion]$_X$ is [a violation of the Sherman Act]$_O$.
6 [Microsoft]$_S$ continues to show [increased earnings]$_O$ despite [the trial]$_X$.

Figura 3.2: Texto com anotações gramaticais para a computação da grade (Barzilay & Lapata, 2008, p. 7)

O texto possui 6 sentenças, e consequentemente, este valor representa a quantidade de linhas da grade de entidades. Na grade da Figura 3.1, por exemplo, a entidade *Evidence* está presente tanto na primeira sentença com um papel sintático diferente de sujeito ou objeto (X), quanto na terceira sentença com o papel sintático Objeto (O), e ausente nas outras sentenças.

Segundo as autoras, o modelo de Grade de Entidades considera a resolução de correferência importante na construção da grade, pois a mesma entidade pode aparecer no decorrer do texto em diferentes formas linguísticas e, desta maneira, um resolvedor de correferência pode ajudar no agrupamento dessas entidades juntamente com as suas respectivas formas linguísticas. Tal procedimento faz com que a grade tenha uma melhor representatividade da distribuição das entidades presentes no texto. Por exemplo, a Figura 3.1 mostra a grade de entidades de um texto que passou por um resolvedor de correferência, o qual descobriu as outras formas linguísticas da entidade *Microsoft* que aparecem no decorrer do texto da Figura 3.2, ou seja, *Microsoft Corp.* e *the company* são essas outras formas linguísticas que estão presentes na coluna rotulada por *Microsoft* da Figura 3.1.

Uma observação importante dada por Barzilay & Lapata é relacionada às entidades que aparecem mais de uma vez em diferentes papéis gramaticais na mesma sentença. Caso isso aconteça, as autoras utilizaram um ranque gramatical para escolher o papel sintático da entidade na sentença. Tal ranque é baseado na precedência utilizada pela teoria *Centering*, ou seja, o papel sintático sujeito (S) possui preferência de escolha sobre o papel sintático objeto (O) e sobre X (papel sintático diferente de sujeito e objeto); já o papel sintático objeto (O) possui preferência de ser escolhido sobre X. Por exemplo, a entidade *Microsoft* é mencionada duas vezes na sentença 1 do texto da Figura 3.2: a primeira possui um papel sintático (X) para a entidade *Microsoft Corp.*, e a segunda possui o papel sintático sujeito (S) para a entidade *the company*. Desta forma, a entidade *Microsoft* terá o papel sintático sujeito representado na grade pela marca S (ver Figura 3.1).

Segundo Barzilay & Lapata, a grade de entidades de um texto é utilizada para formar um vetor de características, o qual é usado como instância para um algoritmo de Aprendizado de Máquina. A principal hipótese deste modelo é que a distribuição de entidades em textos coerentes possui uma certa regularidade evidenciada na topologia da grade.

De acordo com Barzilay & Lapata, as grades de textos coerentes provavelmente devem apresentar algumas colunas mais densas, ou seja, colunas com poucos "buracos", tal como a coluna da entidade *Microsoft* na Figura 3.1, e muitas colunas esparsas com muitos "buracos", por exemplo, as entidades *markets* e *earnings* na Figura 3.1, sendo que, para as colunas mais densas, os papéis sintáticos Sujeito e Objeto são os mais frequentes. Para textos com baixo nível de coerência, as referidas características são menos acentuadas.

Baseado na Teoria de *Centering*, a análise feita por Barzilay & Lapata é voltada a padrões alcançados pelas transições das entidade entre as sentenças do texto. Assim, uma transição de entidade é uma sequência $\{S, O, X, -\}^n$ que representa a ocorrência da entidade e seus papéis sintáticos em *n* sentenças adjacentes. As transições locais podem ser obtidas a partir da grade como uma subsequência contínua em cada coluna, e cada transição terá uma certa probabilidade em uma dada grade. Por exemplo, a probabilidade da transição [O -] na grade da Figura 3.1 é 0,09 (computada pela razão entre a sua frequência de ocorrência na grade, isto é, 7, pelo número total de transições de tamanho 2, ou seja, 75 transições). Cada texto pode ser visto como uma distribuição definida sobre os tipos de transições.

Desta forma, cada versão da grade *j* de um documento d_i corresponde a um vetor de característica $\Phi(x_{ij}) = (p_1(x_{ij}),p_2(x_{ij}),...,p_m(x_{ij}))$, onde *m* é o número de todas as transições de entidades pré-definidas e $p_t(x_{ij})$ a probabilidade da transição *t* na grade x_{ij}. Um exemplo do espaço de característica com transições de tamanho dois é mostrado na Figura 3.3, sendo que a segunda linha (marcada por d_1) é a representação do vetor de característica da grade da Figura 3.1.

	S S	S O	S X	S -	O S	O O	O X	O -	X S	X O	X X	X -	- S	- O	- X	- -
d_1	.01	.01	0	.08	.01	0	0	.09	0	0	0	.03	.05	.07	.03	.59
d_2	.02	.01	.01	.02	0	.07	0	.02	.14	.14	.06	.04	.03	.07	0.1	.36
d_3	.02	0	0	.03	.09	0	.09	.06	0	0	0	.05	.03	.07	.17	.39

Figura 3.3: Exemplo de um vetor de características representando um documento usando todas as transições de tamanho dois (Barzilay & Lapata, 2008, p. 8).

Normalmente, um conjunto de textos de referência (por exemplo, textos jornalísticos) são utilizados como padrões de textos coerentes e, a partir desse conjunto de textos, são produzidas várias instâncias (vetores de características, da forma que é mostrado em uma das linhas da Figura 3.3) do tipo coerente, as quais serão utilizadas para treinar um modelo por meio de um algoritmo de Aprendizado de Máquina, com o intuito de identificar um novo texto coerente por meio da predição dada pelo modelo produzido.

A preocupação de Barzilay & Lapata é determinar quais fontes de conhecimento linguístico são essenciais para a acurácia da predição da coerência e como codificar esse conhecimento linguístico de forma sucinta na representação do discurso. Assim, a exploração por parâmetros é guiada por 3 considerações: a importância linguística de um parâmetro, a acurácia de sua computação automática e o tamanho do espaço de característica resultante.

Para obter as entidades necessárias para a grade, as autoras utilizaram o sistema de resolução de correferência de Ng & Cardie (2002), para determinar quais os sintagmas nominais que se referem a mesma entidade do documento, para que haja apenas uma entrada na grade. Desta forma, o sistema decide se dois sintagmas nominais são correferentes por explorar características léxicas (verificar a correspondência entre os núcleos dos sintagmas nominais), gramaticais (regras de resolução de correferência sintática), semânticas (compatibilidade semântica entre os sintagmas nominais) e posicionais (mede a distância em termos do número de parágrafos entre os dois sintagmas nominais).

Uma outra abordagem para a extração de entidades considerada pelas autoras é quando classes de entidades são construídas por agrupamentos de substantivos, ou seja, cada substantivo no texto corresponde a uma entidade diferente na grade e dois substantivos são considerados correferentes somente se eles forem idênticos. Por exemplo, o termo *Microsoft Corp.* da Figura 3.2, presente na sentença 1, corresponde a dois substantivos, *Microsoft* e *Corp.*, que são distintas de *company*. Para Barzilay & Lapata, esta abordagem é considerada rústica para a resolução de correferência, mas é considerada simples para uma perspectiva implementacional e produção

de resultados consistentes através de domínios e linguagens.

Com relação à obtenção dos papéis gramaticais, as autoras utilizaram o *parser* estatístico de Collins (1997) para determinar a estrutura constituinte para cada sentença, ou seja, identificar substantivos com papel sintático Sujeito (S), substantivos com papel sintático Objeto (O) e substantivos que não são sujeitos e nem objetos (X).

Este modelo de Grade de Entidades traz algumas modelagens linguísticas, por exemplo, o conceito de saliência de entidades, as funções gramaticais das entidades e formas linguísticas de suas menções subsequentes. Desta forma, entidades salientes são as que ocorrem com maior frequência ao longo do texto e em posições sintáticas de destaques (Sujeitos ou Objetos) nas sentenças. Barzilay & Lapata avaliaram o impacto da informação saliente de duas formas: a primeira trata de todas as entidades sem distinção; a segunda diferencia transições de entidades salientes das transições não salientes, sendo que as entidades consideradas salientes são as que possuem uma certa frequência de ocorrência no texto.

Com a utilização da saliência, o procedimento da geração de características sofre uma pequena alteração, ou seja, a computação das probabilidades das transições será feita separadamente e logo em seguida são combinadas em um único vetor de característica. Desta forma, para n transições com k grupos de saliência, o espaço de característica será de tamanho n x k, sendo que foram adotados por Barzilay & Lapata dois grupos ($k = 2$), um grupo saliente e outro não saliente, já que, segundo as autoras, um modelo com múltiplas classes (grupos) de saliência pode ser construído, ou seja, para cada classe pode haver uma frequência de ocorrência de entidades.

Construída a grade e o vetor de característica, Barzilay & Lapata (2008) utilizaram um algoritmo de Aprendizado de Máquina para ranquear a melhor ordem das sentenças na geração textual, avaliar a coerência em sumários, e avaliar a legibilidade textual.

A ordenação de sentenças é considerada uma etapa importante na geração de texto por conceito (Konstas & Lapata, 2012), na sumarização multidocumento e outros problemas de síntese textual.

As autoras procuraram utilizar o modelo de coerência para ranquear ordenações alternativas de sentenças, em vez de encontrar uma ordenação ótima. De acordo com as autoras, a coerência local é uma propriedade chave de textos bem formandos, ou seja, textos sem a coerência local são naturalmente incoerentes globalmente, e um modelo que leva em consideração a coerência local é capaz de discriminar textos coerentes dos incoerentes.

Na tarefa de ordenação houve a geração de versões para cada documento teste por meio de permutações aleatórias de suas sentenças, e as autoras contabilizaram a quantidade de vezes que uma permutação é melhor ranqueada do que o documento original. Um bom modelo deve preferir o documento original com mais frequência do que a sua permutação.

O conjunto de treinamento para esta tarefa possuía pares ordenados de textos gerados por meio de permutações sentenciais (x_{ij}, x_{ik}) para cada documento original d_i, onde x_{ij} é considerado mais coerente do que x_{ik}, assumindo que $j > k$. Assim, segundo Barzilay & Lapata, o objetivo do treinamento é encontrar um vetor de parâmetros w que gere uma função de ranque

que minimiza o número de violações dos ranques em pares dados no conjunto de treinamento. A Equação 3.1 é usada para encontrar um vetor de parâmetros *w*.

$$\forall(x_{ij}, x_{ik}) \in r^* : w.\phi(x_{ij}) > w.\phi(x_{ik}) \tag{3.1}$$

onde (x_{ij}, x_{ik}) $\in$ r^* se x_{ij} é melhor ranqueado do que x_{ik} para um ranque ótimo r^* nos dados de treinamentos e $\phi(x_{ij})$ e $\phi(x_{ik})$ é um mapeamento em representações de características das propriedades de coerência do processamento das representações x_{ij} e x_{ik}, sendo que as características correspondem as probabilidades das transições de entidades. Assim, uma boa função de ranqueamento, representada pelo vetor peso *w*, é a que satisfaz a seguinte condição (ver na Equação 3.2):

$$w.\left(\phi\left(x_{ij}\right) - \phi\left(x_{ik}\right)\right) > 0 \quad \forall j,i,k \quad tal\ que\ j > K \tag{3.2}$$

Sendo assim, caso a diferença das caraterísticas de um texto melhor ranqueado que outro seja maior do que zero (0), uma boa função de ranqueamento deve manter este resultado.

De acordo com Barzilay & Lapata, o problema é tratado pelo classificador chamado *Support Vector Machine* (SVM) (Cortes & Vapnik, 1995b) e pode ser resolvido usando a técnica de busca exposta em Joachims (2002) chamada SVM^{light}, sendo que tal abordagem tem sido considerada eficiente em várias tarefas.

Para treinar e testar o método, uma ampla coleção de texto foi adquirida, fazendo uso de dados sintéticos, ou seja, um conjunto de textos fonte e suas respectivas versões geradas por meio das permutações das sentenças dos textos fonte, ou seja, a cada troca de lugar entre as sentenças adjacentes de um texto fonte forma-se um novo texto, referenciado como texto permutado.

Segundo Barzilay & Lapata, a hipótese sobre a utilização desse tipo de coleção é que a ordem das sentenças originais nos documentos fontes devem ser coerentes, e desta forma, os modelos que ranqueiam os textos fontes em uma posição mais alta do que os seus respectivos textos permutados são os modelos preferidos. O córpus inclui pares de textos (documento original e um texto permutado), sendo que a qualidade das permutações pode influenciar no ranqueamento.

O conjunto de treinamento e de teste a ser utilizado no algoritmo de Aprendizado de Máquina, no caso, SVM^{light}, foi formado por *k* textos fonte, e, para cada um dos textos fonte foram geradas *n* versões permutadas, obtendo assim *k* * *n* pares de textos (instâncias).

O córpus utilizado possui dois gêneros diferentes: artigos de jornais (com tópico em Terremotos) e relatórios de acidentes (com tópico em acidentes aéreos - Acidentes). Cada texto deste córpus possui um número médio de 10,4 e 11,5 sentenças, respectivamente. Para o treinamento e teste foram utilizados 200 documentos fonte (100 textos de cada gênero) com até 20 permutações geradas aleatoriamente para cada texto fonte (4.000 pares de textos), sendo que 10 documentos (200 pares de textos) foram utilizados para o desenvolvimento do modelo.

Com o intuito de investigar a contribuição do conhecimento linguístico na performance do modelo, Barzilay & Lapata produziram representações de grades de entidades com diferentes parametrizações de espaço de característica para o processo de aprendizagem. Desta maneira,

as autoras utilizaram três fontes de conhecimento linguístico: a correferência, o papel sintático e a saliência. Com isso, os modelos com mais informações linguísticas são comparados com os modelos com pouca ou nenhuma informação linguística. Assim, considerando a presença [+] ou a ausência [-] dessas três fontes de conhecimento, oito modelos de Grade de Entidades diferentes foram obtidos por meio das combinações de correferência [+/-], papéis sintáticos [+/-] e saliência [+/-]. Por exemplo, a notação Correferência+Sintático+Saliência+ (modelo completo) faz uso de informações sobre a correferência, as sequências de transições de entidades são denotadas por papéis sintáticos e há diferenciações entre entidades salientes e não salientes.

Além dessas variações da utilização das representações linguísticas, o modelo também especifica dois outros parâmetros: a frequência usada para identificar as entidades salientes e o tamanho da sequência de transição. Modelos baseados em saliência ótima foram obtidos com frequência de ocorrência de entidades ≥ 2, e o tamanho ótimo de transição das entidades entre as sentenças do texto é de ≤ 3.

Na tarefa de ordenação foi utilizado o pacote de treinamento e teste de Joachims (2002), o SVM^{light} para a tarefa de ranqueamento, sendo que foram atribuídos a todos os parâmetros valores padrões.

Os resultados obtidos nesta tarefa de ordenação de sentenças podem ser visto na Tabela 3.1, que mostra que o modelo completo (Correferência+Sintático+Saliência+) obteve resultados melhores do que o modelo básico, ou seja, o modelo sem a presença do conhecimento linguístico (Correferência-Sintático-Saliência-).

Tabela 3.1: Acurácia medida como a porcentagem de ranqueamentos corretos entre pares de texto no conjunto de testes

Modelos	**Terremotos (%)**	**Acidentes (%)**
Correferência+Sintático+Saliência+	87,2	**90,4**
Correferência+Sintático+Saliência-	**88,3**	90,1
Correferência+Sintático-Saliência+	86,6	88,4
Correferência-Sintático+Saliência+	83,0	89,9
Correferência+Sintático-Saliência-	86,1	89,2
Correferência-Sintático+Saliência-	82,3	88,6
Correferência-Sintático-Saliência+	83,0	86,5
Correferência-Sintático-Saliência-	81,4	86,0
Latent Semantic Analysis	81,0	87,3

O modelo de Grade de Entidades da Barzilay & Lapata também foi utilizado para avaliar a coerência de sumários por meio da comparação de ranques obtidos por este modelo com ranques produzidos por julgamentos humanos feitos em sumários.

Um modelo que exibe alta concordância com os julgamentos humanos captura não apenas as propriedades de coerência dos sumários, mas, possivelmente, possa avaliar de forma automática textos gerados por máquinas, diferentes de algumas medidas automáticas já existentes, como BLEU (Papineni et al., 2002) e ROUGE (Lin & Hovy, 2003), as quais não foram criadas para

a tarefa de avaliar a coerência, porque elas focam na similaridade de conteúdo entre os textos gerados por sistemas e textos de referência.

Para Barzilay & Lapata, a avaliação da coerência de sumários pode ser também formulada como uma tarefa de aprendizado por ranqueamento, sendo que os dados utilizados foram sumários multidocumento produzidos por humanos e sistemas de sumarização, obtidos da DUC 2003 (*Document Understanding Conference*).

De forma similar à tarefa de ordenação sentencial, os dados de treinamento para a avaliação da coerência de sumários incluem pares de sumários (x_{ij}, x_{ik}) do(s) mesmo(s) documento(s) d_i, onde x_{ij} é mais coerente do que x_{ik}. Um classificador ótimo retornaria um ranque r^* que ordena os sumários de acordo com a sua coerência. Da mesma forma que foi realizado no experimento de ordenação sentencial, Barzilay & Lapata utilizaram SVM^{light} para treinar um modelo para ranquear os sumários mutidocumento.

Com o intuito de aprender um modelo de classificação, um conjunto de sumários foi utilizado e cada um dos mesmos avaliados em termos de coerência. Esses sumários foram produzidos a partir de 16 agrupamentos de documentos. Todos os sumários foram avaliados em relação a coerência por humanos, os quais atribuíram uma nota entre 1 e 7 para cada sumário.

A partir dos sumários avaliados por humanos, um conjunto de 144 pares de sumários foram usados no treinamento e um outro conjunto formando por 80 pares de sumários foi usado para teste.

Da mesma forma que foi realizado na tarefa de ordenação de sentenças, oito modelos foram utilizados: um modelo para cada configuração (Correferência[+/-] Sintática[+/-] Saliência[+/-]). Além disso, todos eles foram treinados com SVM^{light} voltado para a configuração de ranqueamento.

A Tabela 3.2 mostra a acurácia das versões do modelo de Grade de Entidades e da *Latent Semantic Analysis* para o experimento de avaliação de coerência em sumários. Observando a Tabela 3.2, um ponto interessante que pode ser exaltado é o fato de que na ausência da informação de correferência (Correferência-Sintático+Saliência+), a acurácia aumentou em relação ao modelo completo que possui a informação de correferência (Correferência+Sintático+ Saliência+). Segundo Barzilay & Lapata, há dois motivos para a melhor acurácia na ausência da correferência, sendo que estes motivos estão relacionados ao córpus de sumários, formado por muitos textos gerados por máquinas: o primeiro é a própria ferramenta de resolução de correferência, já que a mesma foi treinada em textos bem formados feitos por humanos, dessa forma, a ferramenta não iria ajudar em textos produzidos por sistemas de sumarização; o segundo motivo é que os sistemas de sumarização automática não usam expressões anafóricas tão frequentemente como ocorre em sumários feitos por humanos. Assim, um método de agrupar as mesmas entidades é melhor para sumários gerados automaticamente do que para sumários gerados por humanos.

No experimento de avaliação da Legibilidade, Barzilay & Lapata investigaram se a representação da grade de entidades pode ser empregada na classificação de estilo, ou seja, o uso da grade de entidades em um sistema que avalie a legibilidade de documentos. Segundo as

Tabela 3.2: Acurácia medida como fração do ranque de pares corretos no conjunto de testes (Barzilay & Lapata, 2008)

Modelos	**Acurácia (%)**
Correferência+Sintático+Saliência+	80,0
Correferência+Sintático+Saliência-	75,0
Correferência+Sintático-Saliência+	78,8
Correferência-Sintático+Saliência+	**83,8**
Correferência+Sintático-Saliência-	71,3
Correferência-Sintático+Saliência-	78,8
Correferência-Sintático-Saliência+	77,5
Correferência-Sintático-Saliência-	73,8
Latent Semantic Analysis	52,5

autoras, o termo "legibilidade" descreve a facilidade com que um documento pode ser lido e compreendido e, assim, alguns métodos de legibilidade focam em fatores semânticos (palavras usadas) e fatores sintáticos (tamanho e estrutura sentencial).

Para esta avaliação, Barzilay & Lapata seguiram a abordagem de Schwarm & Ostendorf (2005), que avaliam a legibilidade como uma tarefa de classificação, podendo combinar várias fontes de conhecimentos, como as tradicionais medidas do nível de leitura, passando pelos modelos de língua estatísticos e até a análise sintática. A unidade da classificação é um único texto e a tarefa do classificador é predizer se esse texto é fácil ou difícil de se ler.

Os dados utilizados neste experimento foram textos da *Encyclopedia Britannica* e *Britannica Elementary* (voltada para crianças). Desta forma, o córpus foi composto por 107 artigos de textos completos da enciclopédia e seus correspondentes textos simplificados da *Britannica Elementary* (214 textos no total). Mesmo que os textos não estivessem explicitamente anotados com níveis de legibilidade, as autoras consideraram tais textos pertencentes a duas categorias de legibilidade: "fácil" e "difícil".

Para este experimento, Barzilay & Lapata criaram duas versões do sistema: uma que usa somente as características usadas por Schwarm & Ostendorf (características sintáticas, semânticas e a combinação das duas) e a outra que faz uso da representação de grade de entidades.

Além disso, Barzilay & Lapata enriqueceram o espaço de características do Schwarm & Ostendorf com características baseadas na correferência. As autoras fizeram, também, experimentos com dois modelos que produziram boas acurácias nos experimentos anteriores: Correferência+Sintático+Saliência+ (ordenação de sentenças) e Correferência-Sintático+Saliência+ (avaliação de sumários). O tamanho da transição era ≤ 2 e as entidades eram consideradas salientes com frequência maior ou igual a 2.

Para este experimento de avaliar a legibilidade dos textos, a acurácia é calcula considerando o número de exemplos de testes preditos corretamente pelo modelo preditivo inferido a partir do SVM sobre o tamanho do conjunto de teste, sendo que, o conjunto de treinamento e de teste tiveram os mesmos números de textos para as duas classes de legibilidade.

A Tabela 3.3 mostra a acurácia alcançada para cada experimento realizado na tarefa de

avaliar a legibilidade dos textos.

Tabela 3.3: Contribuição das características baseadas na correferência para a tarefa de avaliar de forma automática a legibilidade textual

Modelos	**Acurácia (%)**
Schwarm e Ostendorf	78,56
Schwarm e Ostendorf, Correferência+Sintático+Saliência-	**88,79**
Schwarm e Ostendorf, Correferência-Sintático+Saliência+	79,49
Schwarm e Ostendorf, Latent Semantic Analysis	78,56
Correferência+Sintático+Saliência+	50,90
Correferência-Sintático+Saliência+	49,55
Latent Semantic Analysis	48,58

De acordo com Barzilay & Lapata, o córpus revelou que textos fáceis e difíceis diferem em sua distribuição de pronomes e cadeias de correferência em geral. Textos fáceis tendem a empregar menos correferência e o uso de pronomes pessoais é relativamente esparso. Assim, tal observação sugere que a informação de correferência é um bom indicador do nível de dificuldade de leitura e que a sua omissão do espaço de característica baseado em entidades produz baixa acurácia.

O modelo de grade de entidades mostra-se flexível e computacionalmente tratável. Além disso, os resultados alcançados empiricamente validam a importância das informações de saliência e sintáticas para os modelos baseados em coerência. Assim, a combinação de conhecimento sintático e o de saliência produz modelos com boa performance para todas as tarefas apresentadas.

Filippova & Strube (2007) replicaram o experimento de ordenação de sentença de Barzilay & Lapata (2008) para textos jornalísticos em Alemão, no intuito de verificar o comportamento do método em uma língua diferente do Inglês. Desta forma, o córpus TüBa-DZ (Heike Telljohann & Kübler, 2003) com anotação manual de informação sintática, morfológica e de correferência foi utilizado. Com este córpus, 100 textos foram utilizados para o treinamento, teste e desenvolvimento, além da utilização do pacote de aprendizagem SVM^{light}[1] para a tarefa de ranqueamento.

Similar ao trabalho da Barzilay & Lapata (2008), foram utilizadas as propriedades: Correferência (CORREF), Papéis Sintáticos (SINT) e Saliência(SAL). Assim, a Tabela 3.4 mostra a porcentagem de pares que foram raqueados corretamente (textos fontes com maior ranque do que as suas versões permutadas).

Além da implementação do modelo de entidades para o idioma Alemão, os autores propuseram realizar o agrupamento de entidades por relacionamento semântico. Para isso, eles utilizaram uma API (*Application Programming Interfaces*[2]) chamada WikiRelate (Strube & Ponzetto, 2006), que auxiliou no relacionamento semântico entre as entidades.

[1] http://svmlight.joachims.org/
[2] http://pt.wikipedia.org/wiki/API

Tabela 3.4: Acurácias do Modelo de Grade de Entidades para o Alemão (Filippova & Strube, 2007)

	CORREF+	CORREF-
SINT+SAL+	72%	62%
SINT+SAL-	69%	53%
SINT-SAL+	**75**%	**66**%
SINT-SAL-	71%	59%

Os experimentos com relacionamento semântico tiveram dois objetivos:

- Verificar se a informação semântica pode melhorar os melhores resultados alcançados com os conjuntos correferentes;
- Verificar se apenas o relacionamento semântico pode ser confiável para ser usado no agrupamento de entidades, caso um sistema de resolução de correferência esteja indisponível.

Segundo os autores, o córpus TüBa-DZ contém entidades nomeadas (pessoas, localizações, organizações, etc) que podem ser bem relacionadas. Assim, para os autores, a Wikipedia foi a melhor escolha para buscar o relacionamento semântico entre as entidades (entidades com significado próximo ou sinônimas), já que a mesma cobre tanto entidades nomeadas quanto substantivos comuns.

Para agrupar as entidades similares utilizou-se do seguinte procedimento: quando uma nova entidade e_i é encontrada, avalia se a mesma é relacionada com alguma outra entidade já encontrada (E). Considere $e_j \in E$, se $SemRel(e_i, e_j) > t$ então $e_i \in E$, onde t é um limiar semântico.

Desse modo, outros experimentos foram realizados considerando apenas a combinação SINT-SAL+ (que obteve os melhores resultados - ver Tabela 3.4) e diferentes valores para t. Logo, a Tabela 3.5 mostra os resultados alcançados com a utilização do relacionamento semântico entre as entidades.

Tabela 3.5: Acurácias com diferentes limites de relacionamento (Filippova & Strube, 2007)

t	SINT-SAL+CORREF+	SINT-SAL+CORREF-
Sem valor	**75**%	66%
0,1	71%	66%
0,2	72%	66%
0,3	72%	68%
0,4	73%	68%
0,5	73%	**69**%

Os resultados da Tabela 3.5 mostram que a utilização do relacionamento semântico para o agrupamento de entidades melhoraram a acurácia quando não há o uso de resolução de correferência, mas ainda não superou a acurácia quando se faz uso apenas da resolução de correferência. Um exemplo disso é a linha "Sem valor" para t da Tabela 3.5, a qual indica que não

houve uso do relacionamento semântico entre as entidades. Dessa forma, a resolução de correferência foi a informação que produziu a maior acurácia de 75% e sem o uso da resolução de correferência produziu a menor acurácia de 66%.

Esse trabalho de Filippova & Strube (2007) não fez uso de relações discursivas para gerar modelos de coerência que pudessem melhorar a acurácia dos resultados. Além disso, o possível diferencial desse trabalho que seria o uso do relacionamento semântico não teve o efeito esperado pelos autores.

O trabalho de Burstein et al. (2010) usou o modelo de Grade de Entidades de Barzilay & Lapata, na avaliação da coerência de textos produzidos por estudantes (redações), principalmente os estudantes que falam a língua inglesa e não são nativos.

Os autores procuraram combinar o modelo de Grade de Entidades com outras características voltadas para a qualidade da escrita. Desta forma, o córpus foi formado por três conjuntos de dados: o primeiro foi de redações provenientes do TOEFL[3] de pessoas adultas que falam inglês e que não são nativas; o segundo são redações provenientes do GRE (*Graduate Record Admissions Test*)[4] de pessoas adultas nativas e de pessoas também adultas que falam inglês mas não são nativas; o terceiro é formado por redações de estudantes americanos do ensino fundamental e médio, além de redações de estudantes falantes do inglês mas não nativos que submeteram ao *Criterion*[5]

Dois anotadores foram treinados para avaliar a qualidade da coerência, sendo que os mesmos puderam verificar a facilidade de leitura sem esbarrar em sentenças confusas. Os anotadores também foram instruídos a utilizarem um dos 3 pontos de escalas para avaliar uma redação em relação a sua coerência: 1) baixa coerência, 2) um pouco coerente e 3) alta coerência.

Devido à dificuldade de concordância (medida por meio da *Kappa* (Carletta, 1996)), verificada na escala "um pouco coerente", os pontos de escala passaram a ser dois: alta coerência (H - *High Coherence*) e baixa coerência (L - *Low Coherence*). Desta forma, foi obtido o valor de 0,677 de medida *Kappa* (valores mais próximo de 0 possuem baixa concordância e valores mais próximos de 1 possuem alta concordância).

Seguindo a mesma abordagem de Barzilay & Lapata para formar os vetores de características, Burstein et al. (2010) utilizaram estes vetores como instâncias para o algoritmo C5.0[6]. Para melhorar o poder preditivo do algoritmo, os autores incorporaram as características de qualidade de escrita: GUMS, *Type/Token* (TT) e *Shell nouns*. A característica GUMS descreve a qualidade técnica das redações por meio da gramática, uso e erros manuais, e características de estilo do sistema AES (Triantafillou et al., 2002). A *Type/Toke* mede a variabilidade das palavras, ou seja, uma alta probabilidade de uma transição "Sujeito-a-Sujeito" indica que o autor de um texto está repetindo uma entidade na posição "Sujeito" através das sentenças adjacentes. A característica *Shell nouns* são substantivos abstratos, cujo significado vai depender das informações que se referem a outras partes do texto (Aktas & Cortes, 2008). De acordo com os autores,

[3]É um teste de Inglês como uma língua estrangeira.

[4]Teste de Admissão de Registro de Graduação

[5]http://www.fairtest.org/facts/csrtests.html

[6]C5.0 escrito por Ross Quinlan e está disponível comercialmente pela *Rulequest Research* - http://www.rulequest.com/, uma aplicação de aprendizado de máquina de árvore de decisão.

o uso de tal característica é comum em redações e pode também afetar a coerência.

Segundo os autores, redações de pessoas falantes do idioma Inglês e não nativas podem conter muitos erros de ortografia. Assim, foi levado em consideração o impacto do uso de um verificador ortográfico (SPCR+), para verificar se a variação de ortografia afetará as probabilidades das transições na grade de entidades. Por fim, os experimentos fizeram uso de votação majoritária que combinou os melhores resultados obtidos pelas características.

Para avaliar o modelo, várias configurações de características foram testadas para os 3 conjuntos de dados. As Tabelas 3.6, 3.7 e 3.8 mostram os resultados obtidos em função de Precisão (P), Revocação (R) e Medida-F (F). Além dessas medidas, as tabelas supracitadas também mostram a medida de concordância *Kappa* (K) entre o sistema e anotadores para cada experimento.

Tabela 3.6: Dados obtidos por meio do primeiro conjunto de redações (TOEFL) e a concordância entre Anotador/Sistema (Burstein et al., 2010)

		L (n=64)			**H (n=196)**			**L + H (n = 260)**		
BASELINES: sem as características do trabalho de Barzilay & Lapata (2008)	**K**	**P**(%)	**R**(%)	**F**(%)	**P**(%)	**R**(%)	**F**(%)	**P**(%)	**R**(%)	**F**(%)
(a) E-rater	0,472	56	69	62	89	82	86	79	79	79
(b) GUMS	0,455	55	66	60	88	83	85	79	79	79
(c) SOX_TT	0,484	66	55	60	86	91	88	82	82	82
SISTEMAS: com as características do trabalho de Barzilay & Lapata (2008)										
Correferência-Sintático+Saliência+ (configuração da tarefa de analisar a coerência do sumários de Barzilay & Lapata (2008)	0,253	49	34	40	81	88	84	75	75	75
(d) Correferência-Sintático-Saliência-SPCR+M+	0,472	**76**	45	57	84	95	90	83	83	83
(e) Correferência+Sintático+ Saliência-GUMS+	0,590	68	70	69	90	89	90	**85**	85	85
(f) Correferência+Sintático+ Saliência-GUMS+O_TT_Shellnouns+	0,595	68	72	70	**91**	89	90	**85**	85	85
Votação majoritária para o Baseline: (a), (b), (c)	0,450	55	64	59	88	83	85	79	79	79
Votação majoritária: (d), (e), (f)	0,598	69	70	70	90	90	90	**85**	85	85

O *E-rater* indica o uso de um conjunto completo de características do *e-rater* (sistema *online* de avaliação de escrita) [7] e SOX_TT é a relação *Type/token* com as informações usadas no modelo completo de Barzilay & Lapata (Correferência+ Sintático+ Saliência+).

Tabela 3.7: Dados obtidos por meio do segundo conjunto de redações (GRE) e a concordância entre Anotador/Sistema (Burstein et al., 2010)

		L (n=48)			**H (n=210)**			**L + H (n = 258)**		
BASELINES: sem as características do trabalho de Barzilay & Lapata (2008)	**K**	**P**(%)	**R**(%)	**F**(%)	**P**(%)	**R**(%)	**F**(%)	**P**(%)	**R**(%)	**F**(%)
(a) E-rater	0,383	**79**	31	45	86	98	92	86	86	86
(b) GUMS	0,316	68	27	39	85	97	91	84	84	84
(c) e-rater_SOX_TT	0,359	78	29	42	86	98	92	85	85	85
SISTEMAS: com as características do trabalho de Barzilay & Lapata (2008)										
Correferência-Sintático+Saliência+(configuração da tarefa de analisar a coerência do sumários de Barzilay & Lapata (2008)	0,120	35	17	23	83	93	88	79	79	79
(d) Correferência+Sintático+Saliência-SPCR+G+	0,547	1,0	43	60	89	1,0	94	90	90	90
(e) Correferência+Sintático-Saliência-P_TT+	0,462	70	44	54	88	96	92	86	86	86
(f) Correferência+Sintático+Saliência+GUMS+ SOX_TT+	0,580	71	60	65	**91**	94	93	88	88	88
Votação majoritária para o Baseline: (a), (b), (c)	0,383	79	31	45	86	98	92	86	86	86
Votação majoritária: (d), (e), (f)	0,610	1,0	49	66	90	1,0	95	**91**	91	91

[7] http://www.ets.org/erater/about

Tabela 3.8: Dados obtidos por meio do terceiro conjunto de redações (*Criterion*) e a concordância entre Anotador/Sistema (Burstein et al., 2010)

		L (n=37)			H (n=226)			L + H (n = 263)		
BASELINES: sem as características do trabalho de Barzilay & Lapata (2008)	**K**	**P**(%)	**R**(%)	**F**(%)	**P**(%)	**R**(%)	**F**(%)	**P**(%)	**R**(%)	**F**(%)
(a) E-rater	0,315	39	46	42	**91**	88	89	82	82	82
(b) GUMS	0,350	47	41	43	90	92	91	85	85	85
(c) SOX_TT	0,263	78	19	30	88	99	93	88	88	88
SISTEMAS: com as características do trabalho de Barzilay & Lapata (2008)										
(d)Correferência-Sintático+Saliência+(configuração da tarefa de analisar a coerência do sumários da Barzilay & Lapata (2008)	0,383	79	30	43	90	99	94	89	89	89
(e) Correferência-Sintático-Saliência-SPCR+	0,590	68	70	69	90	89	90	85	85	85
(f) Correferência+Sintático+Saliência+S_TT+	0,424	67	38	43	90	97	94	89	89	89
Votação majoritária para o Baseline: (a), (b), (c)	0,324	43	41	42	90	91	91	84	84	84
Votação majoritária: (d), (e), (f)	0,471	**82**	38	52	**91**	99	94	**90**	90	90

Segundo os autores, o experimento com votação majoritária superou os três *baselines*, e que os resultados mostraram que o uso do método de Grade de Entidades para avaliar coerência em redações é promissor. Desta forma, aplicar tal metodologia em outros dados adicionais e criar um sistema automatizado de pontuação de coerência para redações são os próximos trabalhos dos autores.

De acordo com as Tabelas 3.6, 3.7 e 3.8, o uso das características do trabalho de Barzilay & Lapata juntamente com as de Burstein et al. produziu as maiores precisões para os 3 conjuntos de textos avaliados. Isso mostra que as características do trabalho Barzilay & Lapata podem ser mais exploradas com outras informações como as relações discursivas.

O trabalho de Burstein et al. faz uso de um conhecimento linguístico superficial e básico para a formação das grades. Além disso, a junção de várias informações pode deixar o modelo complexo e não tão eficiente.

Baseado na ideia de que um escritor tende a utilizar de forma apropriada as relações de correferência quando este escreve um texto coerente e de que a língua japonesa (idem a italiana) é relativamente difícil de obter a transição de entidades do discurso devido ao uso de elipse (omissão de um termo que pode ser facilmente deduzido pelo contexto da matéria) foi desenvolvido, por Iida & Tokunaga (2012), uma métrica de avaliação de coerência para textos da língua japonesa fazendo uso das relações de correferências identificadas automaticamente.

A métrica proposta leva em consideração alguns pares de entidades em um texto no intuito de capturar o relacionamento dessas entidades consideradas distantes. Para avaliar a coerência do discurso usando tal métrica, os autores utilizaram a saída de um modelo de resolução de correferência.

A hipótese é que as pessoas tendem a utilizar apropriadamente as relações de correferência quando estão escrevendo um texto, ou seja, o melhor uso das relações de correferência é um bom indicador de textos coerentes.

Por exemplo, o texto da Figura 3.4 é considerado coerente; já o seu correspondente incoerente é mostrado na Figura 3.5. No texto incoerente, o pronome "*it*" (termo anafórico) está colocado longe do seu antecedente "iPad2" e uma expressão de "distração" como "*birthday party*" é inserida entre o antecedente e o termo anafórico, assim, a interpretação do "*it*" é mais

difícil do que no texto coerente. Desta forma, aplicar um modelo de correferência para textos coerentes e incoerentes faz surgir diferenças no número de relações de correferência identificadas corretamente. Além disso, se não há diferenças em termos de números, pode haver uma diferença na pontuação de confiança (probabilidade prevista emitida pelo classificador) das relações resolvidas.

s_1: [John] bought [iPad2] as [a gift] for [Lucy].
s_2: However, [it] has [something amiss] with [the sound system].
s_3: As a result, [he] went to [[Lucy]'s birthday party] with no [gift].

Figura 3.4: Entidades entre colchetes de um texto coerente (Iida & Tokunaga, 2012)

s'_1=(s_1): [John] bought [iPad2] as [a gift] for [Lucy].
s'_2=(s_3): As a result, [he] went to [[Lucy]'s birthday party] with no [gift].
s'_3=(s_2): However, [it] has [something amiss] with [the sound system].

Figura 3.5: Texto incoerente obtido pela reordenação aleatória das sentenças do texto da Figura 3.4 (Iida & Tokunaga, 2012)

Baseados nas diferenças acima citadas, os autores propuseram uma métrica para avaliar a coerência de discurso que é calculada de acordo com os dois passos seguintes:

1. Um modelo de correferência (ou anáfora) treinado com textos coerentes anotados é aplicado ao texto alvo *T*;

2. A pontuação de coerência de *T* é calculada por meio da saída do passo 1 através da Equação (3.3).

$$coerencia(T) = \frac{1}{N}\sum_{j}^{N} pont_{ana}(i,j), \tag{3.3}$$

onde *T* é o texto alvo, *j* é a anáfora candidata em *T* e *i* é o candidato à antecedente mais provável de *j*. *N* é o número de anáforas candidatas presentes em *T*. A pontuação de confiança da relação de correferência de *i* e *j*, $pont_{ana}$ *(i,j)*, é a pontuação de saída (probabilidade predita) obtida após a aplicação do modelo de correferência no texto *T* de acordo com o passo 1.

Segundo os autores, a métrica pode ser usada como uma das características do modelo de Grade de Entidades, já que esta métrica é obtida por uma perspectiva diferente da Grade de Entidades (informação da transição de entidades no discurso).

O modelo de resolução de correferência de Iida & Poesio (2011) foi escolhido por trabalhar com a resolução de correferência da língua japonesa e o mesmo apresentou, de acordo com os autores, uma melhor performance nessa tarefa. A Equação 3.4 sistematiza tal modelo de resolução de correferência.

$$coref(i,j) = \frac{P(coref|i,j) + P(anaf|j)}{2}, \tag{3.4}$$

onde *j* é um termo anafórico candidato e *i* é o antecedente candidato mais provável de *j*. O elemento P(corref | i,j) é calculado por um classificador de correferência simples tal como o de Ng & Cardie (2002) e P(anaf | j) é a pontuação anafórica de *j*, que é usada para excluir menções não anafóricas típicas, tal como o pleonasmo[8]. Segundo Iida & Tokunaga, se o modelo de resolução de correferência julgar como anáfora, a *coref(i,j)* $\geq$ *0.5*; caso contrário, não será anáfora.

O $pont_{ana}$ *(i,j)*, necessário na Equação 3.3, é definido na Equação 3.5:

$$pont_{ana}(i,j) = -log(1 - \max_{i} coref(i,j)) \tag{3.5}$$

Nesse trabalho, foram feitos 2 experimentos. O primeiro avalia a eficiência do modelo de resolução de correferência para sintagmas nominais em textos coerentes e incoerentes com o intuito de verificar o uso dos seus resultados na tarefa de avaliação da coerência discursiva. E o segundo experimento é similar ao experimento de ordenação de sentenças de Barzilay & Lapata, além de comparar a métrica desenvolvida com o modelo de Grade de Entidades.

O córpus utilizado nesse trabalho foi o NAIST, que consiste de artigos de jornais japoneses e que contém anotação manual de relações de correferência de sintagmas nominais. A Tabela 3.9 mostra dados estatísticas sobre o córpus NAIST.

Tabela 3.9: Informações sobre o córpus NAIST (Iida & Tokunaga, 2012).

Tipo	**N.° de Artigos**	**N.° de Sentenças**	**N.° de Palavras**	**N.° de Rel. de Correferência**
treino	1.753	24.263	651.986	10.206
teste	696	9.287	250.901	4.396

O primeiro experimento procurou avaliar a eficiência na resolução de correferência de sintagmas nominais em ambos os textos, coerentes e incoerentes (versão permutada dos textos considerados coerentes). Durante a fase de treinamento, os autores usaram somente textos coerentes como instâncias de treinamento para criar um classificador. Por usar somente textos coerentes para o treinamento, era esperado que o modelo apropriadamente identificasse relações de correferência em textos coerentes, enquanto em textos incoerentes teria menos sucesso. Assim, classificadores induzidos por textos coerentes são aplicados tanto em textos coerentes quanto em incoerentes para investigar as diferenças de performance na resolução de correferência.

A Tabela 3.10 mostra os resultados alcançados pela classificação em pares na resolução de correferência de sintagmas nominais em textos coerentes e incoerentes, onde a "coerência" representa os resultados em textos coerentes e a "incoerência" representa os resultados em textos incoerentes.

[8]Uso repetitivo de um conceito ou redundância de um termo, que, se não for vicioso, pode instensificar a força expressiva do discurso [p.ex.: principal protagonista, monopólio exclusivo] Ref. http://www.aulete.com.br/pleonasmo

Tabela 3.10: Resultados usando a resolução de correferência de SN (Iida & Tokunaga, 2012).

	Revocação	**Precisão**	**Medida F**
coerente	0,624	0,508	0.560
incoerente	0.538	0.496	0.516

Para o segundo experimento os autores utilizaram como modelo *baseline* um modelo que classifica aleatoriamente um dos dois textos dados.

Para a representação da Grade de Entidades em Japonês, Iida & Tokunaga empregaram o trabalho de Yokono & Okumura (2010) que, além das três marcações do trabalho original (S, O e X), fez uso da marcação T (Tópico), para distinguir palavras chaves das palavras com papéis de sujeitos, no intuito de capturar os aspectos gramaticais da língua japonesa. Além disso, o classificador utilizado foi o SVM^{light} (Joachims, 1999) .

A Tabela 3.11 mostra que o modelo proposto por Iida & Tokunaga teve uma acurácia maior que o modelo de Grade de Entidades voltado para a língua Japonesa.

Tabela 3.11: Resultados usando a resolução de correferência de SN (Iida & Tokunaga, 2012).

Modelo	**Acurácia (%)**
aleatório	50
grade de entidades (coref-)	67,3
(a) grade de entidades (coref+)	70,7
(b) métrica proposta	76,1
(a) + (b)	**78,2**

Assim, Iida & Tokunaga visualizam um grande interesse de criar modelos que integram da melhor forma os fatores que influenciam a coerência discursiva. O maior problema desse trabalho é a complexidade da língua, que faz com que adaptações sejam feitas para captar mais informações, no intuito de produzir um modelo padrão de coerência.

Feng & Hirst (2012) desenvolveram um modelo de coerência chamado *Multiple Ranks*. Esse modelo consiste em estender o modelo de Grade de Entidades de Barzilay & Lapata por meio não só dos textos fonte, mas também de um ranque de preferências entre as versões permutadas dos textos fonte produzidas na tarefa de ordenação de sentenças.

Os autores alegam que há uma ordenação canônica para as sentenças de um texto. Assim, o grau de coerência de um texto pode ser aproximado por meio da similaridade entre a sua ordenação de sentenças atual e a ordenação de sentenças canônicas. Dessa forma, os autores utilizaram métricas de dissimilaridade entre o texto fonte e suas versões permutadas para definir um ranque de permutação. Essas métricas utilizadas são: *Kendall's τ distance* (Lapata, 2006), *Average continuity* (Zhang, 2011) e *Edit distance* (Chen & Ng, 2004).

Seja *r* o número de textos fontes e *m* o número de permutações aleatórias para cada um dos textos fonte, o número de instâncias de treinamento no modelo de Grade de Entidades é $r \times m$,

enquanto no modelo *Multiple Ranks* esse número de instâncias é de $r \times \binom{m+1}{2} \approx \frac{1}{2} r \times m^2 > r \times m$, quando $m > 2$.

O procedimento na obtenção do modelo preditivo é o mesmo do modelo de Grade de Entidades. Além disso, a tarefa de ordenação de sentenças segue nos moldes de Barzilay & Lapata com a adição de três aspectos específicos nos experimentos: atribuição de ranque, extração de entidade e geração de permutação.

A atribuição de ranque para permutação é baseada no resultado da aplicação de uma métrica de dissimilaridade escolhida. Os autores utilizaram duas abordagens diferentes para atribuir ranques às permutações. Na primeira, as permutações são ranqueadas diretamente por seus valores de dissimilaridade. Já a segunda é conhecida como estratificada, na qual *C* ranques são atribuídos às permutações.

À permutação com o menor valor de dissimilaridade é atribuído o mesmo ranque do texto fonte (zero, o mais alto) e à permutação com o maior valor é atribuído o ranque mais baixo (*C - 1*). Assim, ranques de outras permutações são uniformemente distribuídos nesse intervalo de acordo com os seus valores de dissimilaridade obtidos na primeira abordagem.

Segundo os autores, duas abordagens de extração de entidades foram empregadas: (i) o uso de uma mesma ferramenta de resolução de correferência nos textos fontes e nas versões permutadas; (ii) o não uso de resolução de correferência, ou seja, agrupar substantivos principais por meio de um casamento de *string* simples.

Com a geração de permutação, os autores criaram versões permutadas mais próximas de serem coerentes de forma que o modelo pudesse aprender bem as características de um texto coerente.

Baseado em todos os aspectos mencionados, os resultados foram obtidos por meio da tarefa de ordenação de sentenças. Além disso, o córpus utilizado foi o mesmo de Barzilay & Lapata. Assim, os melhores resultados alcançados por essa abordagem foram: 87,9% de acurácia para textos sobre Terremotos (usando a métrica *Edit distance* de dissimilaridade, resolução de correferência e a atribuição de ranque completo para permutação); 86,3% de acurácia para textos sobre Acidentes (usando a métrica *Edit distance* de dissimilaridade, resolução de correferência e a atribuição de ranque igual a 4 para permutação).

O trabalho de Feng & Hirst não propôs uma abordagem nova, mas algumas modificações que afetam o processo de aprendizagem do modelo de Grade de Entidades. Essas modificações estão mais relacionadas no que se pode esperar do modelo de Grade de Entidades, quando se tem variações das permutações dos textos fonte e no uso ou não de resolução de correferência, do que uma modificação na estrutura do modelo.

Outro trabalho voltado para a avaliação da coerência automática é o de Freitas (2013), o qual investigou a aplicabilidade do modelo de Grade de Entidades de Barzilay & Lapata na avaliação da coerência em resumos científicos escritos em Português do Brasil. Segundo o autor, o intuito era incluir o modelo de Grade de Entidades no Módulo de Análise de Coerência (MAC) da ferramenta SciPo (*Scientific Portuguese*) (Feltrim et al., 2006).

O autor replicou o experimento de ordenação de sentenças realizado por Barzilay & Lapata

(2008). Para isso, os textos originais de três córpus jornalísticos foram utilizados: CSTNews (Cardoso et al., 2011), Summ-it (Collovoni et al., 2007) e Temário (Rino & Pardo, 2006). Tais córpus foram considerados coerentes pelo autor. A Tabela 3.12 resume a variação do tamanho dos textos jornalísticos em número de sentenças utilizado.

Tabela 3.12: Informações dos córpus (Freitas, 2013).

Córpus	**Textos**	**N. mínimo**	**N. Máximo**	**Média**
CSTNews	136	3	48	16,01
Summ-it	50	4	17	16,22
Temário	100	5	69	29,12

Além dos textos jornalísticos, resumos científicos escritos por alunos de graduação que compõem seus trabalhos de conclusão de curso (TCCs) foram coletados para os experimentos, sendo que 139 resumos apresentaram problemas de quebra de sentido lógico entre sentenças adjacentes do resumo, ou seja, foram julgados como prováveis textos com problemas de coerência. A Tabela 3.13 mostra a variação do tamanho dos textos científicos em número de sentenças.

Tabela 3.13: Informações do córpus Científico (Freitas, 2013)

Córpus	**Textos**	**N. mínimo**	**N. Máximo**	**Média**
Científico	139	2	18	5,96

Assim como em Barzilay & Lapata, para cada um dos textos dos córpus jornalísticos foram gerados aproximadamente 20 versões sintéticas de permutações aleatórias da ordem das sentenças, e assumiu-se que os textos com a ordenação sentencial original são considerados mais coerentes do que os textos com sentenças permutadas. Já o córpus de resumos científicos passou por julgamentos humanos para identificar os resumos que apresentassem uma quantidade considerável de problemas na leitura em relação ao tamanho do texto. Caso os textos apresentassem tais problemas, os mesmos eram marcados "com problemas", caso contrário os resumos seriam marcados como "sem problemas".

Diferentemente do modelo original de Barzilay & Lapata, o trabalho feito por Freitas não utilizou da informação de correferência devido a falta de uma ferramenta de resolução de correferência para o Português do Brasil, e os córpus não possuírem anotações de correferência. Dessa forma, a etapa de identificação de entidades seguiu de forma similar a abordagem feita por Eisner & Charniak (2011), em que apenas os sintagmas nominais que possuíssem o mesmo núcleo são considerados correferentes.

Assim, para avaliar este modelo de grade de entidades para o Português, dois tipos de experimentos foram realizados:

- Ordenação de sentenças, no mesmo formato do trabalho de Barzilay & Lapata;

- Julgamento de juízes humanos, nos moldes dos experimentos realizados por Burstein et al..

O autor, com o primeiro experimento, verificou se o comportamento do modelo de Grade de Entidades aplicado para o Português do Brasil é semelhante a de outras línguas. Já o segundo experimento investigou a eficiência do modelo de Grade de Entidades na detecção de problemas locais de coerência em resumos científicos escritos em Português. Com isso, a fase de aprendizagem foi construída como um problema de classificação, similar ao trabalho de Burstein et al. (2010). Desta forma, os experimentos foram realizados no ambiente WEKA (Witten & Frank, 2005) com três algoritmos de Aprendizagem de Máquina (AM): SVM (Cortes & Vapnik, 1995a); C4.5 (Quinlan, 1993) e *Naïve Bayes* (Tan et al., 2005).

O *baseline* utilizado nesse trabalho foi uma implementação que utiliza a técnica LSA (*Latent Semantic Analysis*). Desta forma, para o experimento de ordenação de sentenças, o trabalho de Freitas obteve 74,44% de acurácia (distinção correta dos textos originais de suas permutações) para o córpus CSTNews; 50,29% de acurácia para o córpus Summit; 59,24% de acurácia para o córpus Temário; e 58,10% de acurácia para todos juntos. Tais resultados foram alcançados na configuração completa do modelo (Sintático+ Saliência+), tendo superado o *baseline* apenas no córpus CSTNews.

Segundo o autor, o modelo de Grade de Entidades para Português desenvolvido superou o *baseline* em quase todos córpus em alguma configuração, com exceção do córpus Temário, em que o *baseline* foi sempre melhor. A Tabela 3.14 mostra os resultados, em termos de acurácia, alcançados por esse trabalho em relação ao experimento de ordenação de sentenças.

Tabela 3.14: Acurácias obtidas para o primeiro experimento (Freitas, 2013)

Modelo	**CSTNews**	**Summit**	**Temário**	**Todos Juntos**
LSA	61,42%	56%	79%	67%
Sintático+ Saliência-	64%	48,23%	60,45%	62,10%
Sintático+ Saliência+	**74,44%**	50,29%	59,24%	58,10%
Sintático- Saliência-	69,44%	63,83%	**74,84%**	**68,57%**
Sintático- Saliência+	70,88%	**72,05%**	65,45%	67,36%

No segundo experimento (avaliação da coerência em resumos científicos, distinguindo resumos "com problemas" dos "sem problemas" de coerência), o córpus apresentou um desbalanceamento considerável, sendo que a classe majoritária - "sem problemas" - correspondeu a 84% do córpus (117 resumos) e a classe minoritária - "com problemas" - correspondeu a 16% (22 resumos).

Os resultados foram obtidos a partir da técnica de validação cruzada de 10 partições no córpus com 139 resumos científicos desbalanceados. A Tabela 3.15 mostra os resultados do modelo de Grade de Entidades aplicado no córpus de resumos científicos em termos de *Medida-F* e *Kappa - k*, sendo que o algoritmo C4.5 foi o que obteve o melhor resultado, *Medida-F* = 0,91 e K = 0,65.

Tabela 3.15: Resultados obtidos para o segundo experimento (Freitas, 2013)

	Naïve Bayes		**SVM**		**C4.5**	
TT-	*Medida-F (%)*	*Kappa*	*Medida-F (%)*	*Kappa*	*Medida-F (%)*	*Kappa*
Sintático+ Saliência-	0,66	0,21	0,76	0,00	0,81	0,25
Sintático+ Saliência+	0,74	0,05	**0,80**	**0,14**	0,88	0,51
Sintático- Saliência-	0,70	0,21	0,76	0,00	0,80	0,18
Sintático- Saliência+	**0,79**	0,16	0,76	-0,14	**0,91**	**0,65**
TT+						
Sintático+ Saliência-	0,73	**0,26**	0,76	-0,01	0,80	0,27
Sintático+ Saliência+	0,77	0,11	**0,80**	**0,14**	0,87	0,49
Sintático- Saliência-	0,74	0,22	0,76	0,00	0,80	0,18
Sintático- Saliência+	**0,79**	0,16	0,79	0,12	**0,91**	**0,65**

Seguindo a mesma abordagem de Burstein et al. (2010), o trabalho de Freitas também utilizou os atributos *Type/Token* (TT+ presente na Tabela 3.15 e na 3.16) para medir a variedade léxica das entidades que ocorrem em cada papel sintático, sendo esta informação uma tentativa de melhorar o aprendizado automático, mas segundo o autor tal informação não teve o efeito esperado.

Outro experimento realizado por Freitas foi de avaliar o efeito do desbalanceamento do córpus de resumo. Para isso, os experimentos com os três algoritmos de aprendizado foram refeitos utilizando-se de uma técnica de balanceamento chamada SMOTE (*Synthetic Monority Oversampling Technique*) (Chawla et al., 2002). Assim, a Tabela 3.16 mostra os resultados do modelo de grade de entidades para o córpus de resumos científicos balanceado com o SMOTE em termos de *Medida-F* e de *Kappa*.

Tabela 3.16: Resultados obtidos para o segundo experimento com *oversampling* (Freitas, 2013)

	Naïve Bayes		**SVM**		**C4.5**	
**_TT-*	*Medida-F (%)*	*Kappa*	*Medida-F (%)*	*Kappa*	*Medida-F (%)*	*Kappa*
Sintático+ Saliência-	0,71	0,46	0,72	0,47	0,80	0,61
Sintático+ Saliência+	0,76	0,52	0,83	0,66	0,90	0,80
Sintático- Saliência-	0,63	0,29	0,67	0,38	0,76	0,54
Sintático- Saliência+	0,61	0,29	0,58	0,25	**0,91**	0,82
**_TT+*						
Sintático+ Saliência-	0,77	0,55	0,83	0,66	0,80	0,61
Sintático+ Saliência+	0,79	0,59	0,80	0,14	0,90	0,80
Sintático- Saliência-	0,70	0,43	0,72	0,46	0,76	0,53
Sintático- Saliência+	**0,89**	**0,78**	**0,86**	**0,73**	**0,91**	**0,83**

De acordo com o autor, o C4.5 continua sendo o algoritmo que obteve os melhores resultados, tanto com quanto sem o processo de *oversampling*[9]. E a configuração *Sintático- Saliência+* foi a que obteve os resultados mais próximos dos produzidos por juízes humanos.

Como sugestão de melhoramento e avanço do trabalho, o autor vê a importância de uma compilação e anotação manual de um córpus maior e mais balanceado para a realização dos

[9]Em análise de dados, a técnica de *oversampling* ajusta a distribuição de classes de um conjunto de dados (Chawla et al., 2002)

testes, para não correr riscos de possíveis influências que a técnica de *oversampling* possa exercer. Outra questão levantada é a necessidade de uma ferramenta de resolução de correferência para o Português do Brasil, a qual poderia melhorar os resultados. Além disso, a utilização deste modelo em outros contextos de aplicação é proposto como trabalho futuro.

Segundo Guinaudeau & Strube (2013), o trabalho de Barzilay & Lapata tem algumas desvantagens, como: esparsidade dos dados, dependência de domínio e complexidade computacional (principalmente em termos de espaço de características na construção do modelo). Para superar essas desvantagens, os autores propuseram representar as entidades em um grafo bipartido e computar a coerência local pela aplicação de medidas de centralidade aos nós do grafo.

De acordo com Guinaudeau & Strube, o grafo bipartido contêm informação suficiente de transição de entidades entre as sentenças. Essa informação é necessária para a computação da coerência local, sem o uso de vetores de características e, consequentemente, da etapa de aprendizado de máquina como ocorre no modelo de Grade de Entidades de Barzilay & Lapata.

O grafo bipartido *G* é definido como uma quádrupla, $G = (V_s, V_e, L, w)$, sendo que V_s e V_e são os conjuntos de nós que representam as sentenças e as entidades do texto, *L* é um conjunto de arestas associadas com pesos *w*. No grafo bipartido, só haverá uma aresta entre um nó sentença s_i e um nó entidade e_j quando a correspondente célula na grade de entidades não é igual a " - ". Cada aresta é associada com um peso $w(e_j, s_i)$, e este depende do papel gramatical da entidade e_j na sentença. Os autores consideraram que o papel gramatical Sujeito (S) tem peso igual a 3, o Objeto (O) igual 2, qualquer outro papel gramatical (X) igual a 1 e " - " igual a 0. Assim, a chamada Matriz de Incidência é uma Matriz de Entidades dada por Barzilay & Lapata, mas com a substituição dos papéis gramaticais pelos seus respectivos pesos. A Figura 3.6 (a) exemplifica parte de uma Matriz de Entidades de um sumário multidocumento e na Figura 3.6 (b) mostra parte da Matriz de Incidência da Figura 3.6 (a).

	e1	e2	e3	e4	e5
s1	-	-	-	-	-
s2	-	X	-	-	X
s3	S	-	-	-	-
s4	-	-	O	O	-
s5	-	-	O	-	-
s6	-	-	-	-	-

(a)

	e1	e2	e3	e4	e5
s1	0	0	0	0	0
s2	0	1	0	0	1
s3	3	0	0	0	0
s4	0	0	2	2	0
s5	0	0	2	0	0
s6	0	0	0	0	0

(b)

Figura 3.6: (a) Matriz de Entidades e (b) Matriz de Incidência

A partir da Matriz de Incidência, gera-se o grafo bipartido (veja a Figura 3.7), com o qual os autores modelam as transições de entidades entre sentenças. Desse grafo, geram-se 3 tipos de grafos de projeções *one-mode* (P_U, P_W e P_{Acc}) que são utilizados para calcular o valor de coerência do texto.

No grafo de projeção *one-mode* do tipo P_U (*Projection Unweighted* - Projeção não pon-

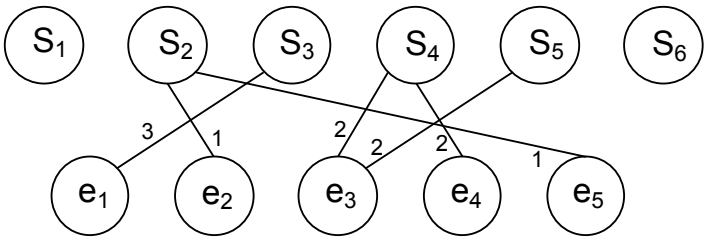

Figura 3.7: Grafo Bipartido

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0	0	0	0	0	0
S_2	0	0	0	0	0	0
S_3	0	0	0	0	0	0
S_4	0	0	0	0	**1**	0
S_5	0	0	0	0	0	0
S_6	0	0	0	0	0	0

Figura 3.8: Matriz adjacente não ponderada

derada), cria-se uma matriz de adjacências não ponderada. As linhas e colunas dessa matriz representam as sentenças do texto e as células são preenchidas por pesos que são binários e iguais a 1, caso duas sentenças tenham pelo menos uma entidade em comum. A Figura 3.8 mostra a matriz adjacente não ponderada (P_U) do grafo bipartido da Figura 3.7. Segundo o grafo bipartido da Figura 3.7, apenas as sentenças 4 e 5 possuem uma entidade em comum. Assim, a matriz de incidência não ponderada teve a célula formada por S_4 e S_5 preenchida com o valor 1, como mostra a Figura 3.8.

No grafo de projeção *one-mode* do tipo P_W (*Projection Weighted* - Projeção Ponderada) forma-se a matriz de adjacências ponderada com o preenchimento da mesma com peso, o qual é o número de entidades compartilhadas por duas sentenças. A Figura 3.9 mostra a matriz adjacente ponderada (P_W) do grafo bipartido da Figura 3.7.

Como houve apenas uma entidade em comum entre as sentenças 4 e 5, a matriz de adjacências ponderada é preenchida com o valor 1 na célula S_4 x S_5, como mostra a Figura 3.9.

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0	0	0	0	0	0
S_2	0	0	0	0	0	0
S_3	0	0	0	0	0	0
S_4	0	0	0	0	**1**	0
S_5	0	0	0	0	0	0
S_6	0	0	0	0	0	0

Figura 3.9: Matriz adjacente ponderada

Já no grafo de projeção com informação sintática (P_{Acc}), o peso utilizado no preenchimento da matriz P_{Acc} é dado seguinte pela Equação 3.6.

$$w_{ik} = \sum_{e \in E_{ik}} w(e,s_i).w(e,s_k) \tag{3.6}$$

onde E_{ik} é um conjunto de entidades compartilhadas por s_i e s_k. A distância entre sentenças pode ser usado na obtenção dos pesos das matrizes de projeções *one-mode* para diminuir a importância das ligações entre sentenças não adjacentes. Assim, os pesos dos grafos de projeções são divididos por $k - i$.

Utilizando as matrizes de projeções *one-mode*, pode-se calcular o valor da coerência local para o texto por meio da Equação 3.7.

$$\begin{aligned} LocalCoherence(T) &= AvgOutDegree(P) \\ &= \frac{1}{N} \sum_{i=1..N} OutDegree(s_i) \end{aligned} \tag{3.7}$$

onde $OutDegree(s_i)$ é a soma dos pesos associados a arestas que deixam s_i e N é o número de sentenças do texto.

Para avaliar o método, os autores utilizaram as mesmas tarefas e os mesmos dados de Barzilay & Lapata. Para a tarefa de ordenação de sentenças, tarefa utilizada nesta tese, Guinaudeau & Strube obtiveram os seguintes resultados mostrados na Tabela 3.17.

Tabela 3.17: Resultados obtidos de Guinaudeau & Strube (2013)

Modelos	Acurácia (%) com informação de correferência	Acurácia (%) sem informação de correferência
$P_U, Dist$	83,3	83,0
$P_W, Dist$	84,9	87,1
$P_{Acc}, Dist$	85,2	88,9

onde $Dist$ é a informação de distância entre sentenças. A informação de correferência é dada por um resolvedor automático de correferência de substantivos e sem informação de correferência é quando os mesmos substantivos foram agrupados na mesma coluna da matriz de entidades.

O trabalho de Silva & Feltrim (2015) combinou o modelo de Grade de Entidade de Barzilay & Lapata com informações oriundas da estrutura retórica para gerar mensagens que indiquem possíveis problemas de coerência local em regiões específicas nos resumo de trabalhos de conclusão de curso feitos por alunos de graduação. Segundo os autores, o principal problema indicado pelas mensagens é a quebra de linearidade. Tal quebra é definida como uma dificuldade em se estabelecer uma ligação clara da sentença atual com as sentenças adjacentes.

Diferentemente dos outros trabalhos baseados em Grade de Entidades de Barzilay & Lapata que realizam a análise da coerência do texto completo, o trabalho de Silva & Feltrim faz essa análise por trechos menores constituídos por um ou mais componentes retóricos. De acordo

com os autores, essa análise por trechos permite a geração de mensagens que indiquem quebras de linearidade em um componente ou grupo de componentes retóricos específicos.

A identificação dos componentes retóricos foi realizada pelo classificador AZPort (Feltrim et al., 2006), que classifica cada sentença de um resumo em uma de seis categorias retóricas: Contexto, Lacuna, Propósito, Metodologia, Resultado e Conclusão.

Segundo os autores, a grade de entidades foi construída seguindo o modelo original de Barzilay & Lapata levando em consideração as informações sintáticas das entidades e de saliência.

Dois classificadores foram criados para a quebra de linearidade: um para classificar componentes retóricos isolados e o outro para classificar resumos completos. Os classificadores foram induzidos com o algoritmo J48 disponível no ambiente Weka (Witten et al., 2011) e os resultados foram alcançados por meio do método de *10-folds cross-validation*. O treinamento e teste dos classificadores foram feitos com o CorpusTCC (Souza & Feltrim, 2012), um córpus composto por 408 resumos extraídos de monografias de conclusão de curso de graduação em Computação.

De acordo com Silva & Feltrim, 1.160 pares de componentes retóricos (compostos de no mínimo 3 sentenças) extraídos a partir dos resumos foram utilizados no treinamento do classificador de componentes. Desses pares, 580 pares eram originais e 580 pares foram gerados pela inversão das sentenças. Já para o treinamento do classificador de resumos completos, 816 resumos foram utilizados (408 resumos originais e 408 resumos gerados pela inversão da ordem das sentenças). Tanto para pares quanto para resumos, as versões geradas artificialmente foram consideradas Com Quebra enquanto os textos originais foram considerados Sem Quebra.

Segundo os autores, o classificador de componentes obteve taxa de acerto de 95,17%, e para o classificador de textos completos obteve uma taxa de acerto 85,05%. Ambos os classificadores utilizaram a grade de entidades na configuração Sintático+ Saliência+ (utilização da informação sintática e da saliência).

Para a avaliação da quebra de linearidade, os autores utilizaram um conjunto de 28 resumos originais, sendo 14 resumos Com Quebra e 14 resumos Sem Quebra. Os resumos Com Quebra foram selecionamos manualmente do CorpusTCC por dois anotadores humanos. O experimento de verificar a acurácia na identificação das quebras de linearidade de forma automática obteve uma taxa de acerto de 67,86% .

O trabalho de Silva & Feltrim está restrito a analisar a quebra de linearidade em resumos de trabalhos de conclusão de curso, tal trabalho pode ser extendido para outros elementos mais complexos que atingem a coerência local de um texto. Nesta tese foi desenvolvida outra abordagem (ver Capítulo 6) que pode identificar mais elementos que afetam a coerência.

3.0.2 Trabalhos Baseados em Discurso

Os trabalhos baseados em discurso buscam distinguir textos coerentes dos incoerentes por meio de padrões de distribuição de relações discursivas presentes nos textos.

O trabalho de Lin et al. (2011) cria um modelo que representa e avalia a coerência fazendo uso de relações discursiva presentes no texto. Lin et al. assumem que a coerência local favorece

implicitamente certos tipos de transições de relações discursivas.

Considerando que a ordem de algumas relações discursivas possam influenciar a coerência local, esse modelo busca padrões de ocorrência das relações ao longo do texto. Assim, este modelo captura a coerência de um texto baseado na distribuição das relações discursivas, mas especificamente nas transições entre as sentenças adjacentes.

As relações discursivas utilizadas foram oriundas da gramática *Discourse Lexicalized Tree Adjoining Grammar (D-LTAG)* (Webber, 2004), reconhecidas no córpus *Penn Discourse Treebank (PDTB)* (Prasad et al., 2008b). O *parser* marca cada relação explícita ou implícita com 2 níveis de tipos de relações. O trabalho de Lin et al. utilizou 4 tipos do nível 1 do PDTB: *Temporal* (Temp), *Contingency* (Cont), *Comparison* (Comp) e *Expansion* (Exp). Além dessas relações, o *parser* identifica automaticamente entidades comuns (EntRel) e sem relação (NoRel) como tipo.

Os autores consideraram duas abordagens: uma abordagem considerada simples e uma mais refinada. A primeira busca modelar diretamente as conexões entre as relações por meio do uso das sequências das transições de relações discursivas entre as sentenças, sendo utilizado um classificador para distinguir textos coerentes dos incoerentes. Tal abordagem, entretanto, revelou alguns problemas: as relações discursivas em textos curtos são poucas, dificultando assim o julgamento automático da coerência; outro problema dessa abordagem é que esta não consegue distinguir a ordenação sentencial de uma determinada relação. Já a abordagem mais refinada busca eliminar esses problemas com uma exploração melhor da saída do *parser* para prover evidências mais circunstanciais para a decisão do sistema de julgamento da coerência. Para isso, uma estrutura baseada na Grade de Entidades de Barzilay & Lapata (2008), denominada Matriz de Papéis Discursivos, é construída.

Similar ao trabalho de Barzilay & Lapata, a Matriz de Papéis Discursivos se diferencia da Grade de Entidades somente no seu preenchimento e no uso de termos radicalizados (termos na sua forma de radical), isto é, em vez de papéis sintáticos, usam-se as relações discursivas e a sinalização de argumentos das relações identificadas. Assim, a Matriz de Papéis Discursivos é composta de sentenças (linhas) e termos radicalizados (colunas), como mostra a Figura 3.10. Desta maneira, a Matriz de Papéis Discursivos representa os diferentes papéis discursivos dos termos através das sentenças em textos contínuos, sendo que as sentenças são consideradas as unidades textuais e os termos são definidos como as palavras de classe aberta (substantivos, verbos, adjetivos e advérbios), em que os radicais destas palavras são colocados em cada coluna da matriz.

Baseando-se na hipótese de que a sequência de transições de papéis discursivos em um texto coerente provê indícios que o distingue de um texto incoerente, a matriz tem uma função importante para computar tais transições de papéis discursivos de uma relação sentença por termo.

As transições dizem como os papéis discursivos de um termo variam de acordo com a progressão do texto. Por exemplo, o termo "cananea" da Figura 3.10 faz parte do argumento 1 (Arg1) da relação discursiva *Comparison* (Comp.Arg1) dada pela sentença 1 (S_1); o mesmo

S#	Termos				
	copper	cananea	operat	depend	...
S_1	nil	Comp.Arg1	nil	Comp.Arg1	
S_2	Comp.Arg2 Comp.Arg1	nil	nil	nil	
S_3	nil	Comp.Arg2 Temp.Arg1 Exp.Arg1	Comp.Arg2 Temp.Arg1 Exp.Arg1	nil	
S_4	nil	Exp.Arg2	Exp.Arg1 Exp.Arg2	nil	

Figura 3.10: Exemplo de uma Matriz de Papéis Discursivos (Lin et al., 2011)

termo faz parte do argumento 2 (Arg2) da relação *Comparison* (Comp.Arg2) dada pela sentença 3 (S_3); e Exp.Arg1 e Exp.Arg2 na S_3 e S_4 respectivamente.

São 12 possíveis papéis discursivos, ou seja, 6 tipos de relações (*Temp(oral), Cont(ingency), Comp(arison), Exp(ansion), EntRel e NoRel*) e 2 marcações de argumentos (Arg1 e Arg2), além do valor *nil* (sem relação). As transições de papéis discursivos são definidas como uma subsequência de papéis discursivos para um termo em múltiplas sentenças consecutivas. Por exemplo, a transição do papel discursivo de "cananea" da S_1 para S_2 é Comp.Arg1 $\rightarrow$ *nil*. Como uma célula pode conter mais de um papel discursivo, a transição deve produzir múltiplas subsequências, por exemplo, ainda para o termo "cananea" da S_3 para S_4, o qual possui as transições Comp.Arg2 $\rightarrow$ Exp.Arg2, Temp.Arg1 $\rightarrow$ Exp.Arg2 e Exp.Arg1 $\rightarrow$ Exp.Arg2.

Cada subsequência tem uma probabilidade que pode ser calculada por meio da matriz. Para o fragmento da matriz da Figura 3.10, o total de transições de tamanho 2 é 25. Além disso, a subsequência Comp.Arg2 $\rightarrow$ Exp.Arg2, por exemplo, ocorre duas vezes. Portanto, a probabilidade da subsequência Comp.Arg2 $\rightarrow$ Exp.Arg2 é 0,08, ou seja, 2/25.

Segundo Lin et al., a principal característica da abordagem assumida por eles é que, enquanto as transições discursivas são capturadas localmente, as probabilidades das transições discursivas são agregadas globalmente, sendo esta distribuição global de um texto coerente distinguível de um texto incoerente. Assim, a diferença distribucional de cada subsequência de textos coerentes e de textos incoerentes, em treinamento, pode subsidiar o julgamento da coerência em um texto nunca visto. Portanto, para avaliar a coerência local, os autores extraíram as subsequências de papéis discursivos como características (subsequências consistindo de apenas valores *nil* foram desconsideradas) e computaram as probabilidades das subsequências com valores para o vetor de características. Dessa forma, foi utilizada a tarefa de ranqueamento de preferência do algoritmo SVM^{light} (*Support Vector Machine*) (Joachims, 1999).

O experimento realizado nesse trabalho segue a mesma metodologia utilizada no experimento de ordenação de sentenças de Barzilay & Lapata, sendo que o sistema de aprendizado deveria predizer qual texto, dos pares em teste, seria o mais coerente. Para comparação, esse trabalho fez uso do mesmo córpus utilizado no trabalho de Barzilay & Lapata (um córpus com 100 textos com foco em notícias sobre Terremotos e com 100 textos de relatos oficiais sobre Acidentes aéreos). O córpus foi utilizado tanto em treinamento quanto nos testes, sendo que

para cada texto suas sentenças foram permutadas em até 20 vezes para criar um conjunto de textos formados por permutações das sentenças do textos fontes (textos considerados incoerentes em comparação aos textos fontes). Com isso, a base de dados era formada por pares de textos, contendo um texto fonte e uma de suas versões permutadas.

O trabalho de Lin et al., em sua versão completa (presença do Tipo da Relação, da informação do Argumento e da informação de Saliência), obteve 86,50% de acurácia no córpus Terremoto e 89,38% de acurácia no córpus Acidente.

De acordo com Lin et al. (2011), a junção desse modelo com o modelo de Grade de Entidade de Barzilay & Lapata (utilização da grade entidades com informação sintática) atingiu uma melhora significativa: 89,72% de acurácia para o córpus Terremotos e 91,64% para o córpus Acidente.

Tais resultados mostram que a utilização das relações discursivas são bem promissoras na avaliação da coerência local. Dessa forma, os autores pretendem aplicar esse modelo em outras tarefas, como a sumarização, a geração textual e um sistema de pontuação para produção textual, que também necessita produzir e avaliar a coerência discursiva. Além disso, o autor não fez uso de outras relações discursivas para comparar e verificar qual relação discursiva se sairia melhor nas predições da coerência local.

Feng et al. (2014) criaram modelos que utilizam informações de discurso capazes de diferenciar textos coerentes dos incoerentes. Tais informações de discurso advêm da RST Mann & Thompson (1987), por meio de suas relações presentes nos textos.

Os autores se basearam no modelo de Grade de Entidades de Barzilay & Lapata e no modelo de grade de relações discursivas de Lin et al. para criarem modelos de grade de relações RST. Dessa forma, os autores desenvolveram dois modelos chamados de modelo de RST Completo e o modelo de RST Superficial.

O modelo de RST Completo é similar ao modelo de Lin et al.. Entretanto, os autores utilizaram as relações RST em vez das relações PDTB e as informações de argumentos (Arg1 e Arg2) foram substituídas pelas informações de nuclearidade (Núcleo e Satélite). Assim, o modelo de RST Completo cria uma grade com entidades representadas nas colunas e sentenças nas linhas, onde o preenchimento dessa grade é com relações RST em que cada entidade participa. Desta forma, uma relação RST é colocada na grade do modelo RST Completo quando uma entidade está presente em uma sentença e esta faz parte de uma relação RST (como núcleo ou como satélite).

Considerando a representação de uma árvore RST dos textos, os autores utilizaram as EDUs principais para verificar quais EDUs seriam consideradas na relação raiz dessa representação. Segundo os autores, as EDUs principais são obtidas no percorrer das sub-árvores discursivas em que a relação de interesse constitui o nó raiz, seguindo os nós núcleo até os nós folhas.

Já o modelo de RST Superficial é similar ao modelo de RST Completo, mas, nesse modelo, os autores consideraram apenas as chamadas relações RST Superficiais, ou seja, relações RST entre duas EDUs que estão na mesma sentença ou em duas sentenças adjacentes.

Para efeito de comparação com os modelos desenvolvidos, Feng et al. reimplementaram

os modelos de Barzilay & Lapata e de Lin et al.. Além disso, os autores fizeram uso de duas tarefas: Ordenação de Sentenças e Pontuação de Redações.

Para a tarefa de Ordenação de Sentenças, a qual é utilizada nesta tese, Feng et al. utilizaram 735 textos fontes e 14.700 permutações (20 permutações para cada texto fonte). O modelo de RST Completo obteve 99.1% de acurácia e o modelo de RST Superficial obteve 98.5%.

Nesse trabalho de Feng et al., a utilização de relações RST se mostrou bastante eficaz para textos fonte que possuem uma estrutura de relações RST definida. No entanto, os autores não se atraveram a verificar essa mesma eficiência em outros tipos de textos, como os textos de sumários multidocumento.

3.0.3 Trabalhos Baseados em Estatística/Matemática

Os trabalhos baseados em Estística/Matemática tentam avaliar a coerência local por meio de métricas que utilizam pouco ou nenhum conhecimento linguístico.

Desenvolvido por Landauer et al. (1998), a *Latent Semantic Analysis* é um modelo estatístico/matemático completamente automático para extrair e representar o conhecimento do contexto esperado por meio das palavras no discurso. Inicialmente desenvolvida para a área de Recuperação de Informação, a LSA busca construir um espaço semântico em que a semelhança entre os termos se dá pela ocorrência em contextos comuns. Por exemplo, dadas as sentenças "O exército está fazendo a segurança." e "Os militares estão garantindo a segurança." as palavras "exército" e "militares" podem ser consideradas similares, já que ocorrem no mesmo contexto com a palavra "segurança".

Inicialmente, uma matriz é formada por meio da análise de um córpus. Essa matriz é formada por termos e suas respectivas quantidades de ocorrências nos textos (contextos) do córpus, ou seja, as linhas representam termos do córpus e as colunas os textos, como é mostrado na Tabela 3.18. Cada valor contido na matriz é submetido a uma normalização, a qual atribui um peso a cada entrada da matriz de acordo com sua importância em relação às outras entradas. O modelo TF-IDF (*term frequency - inverse document frequency*) (Sparck Jones, 1972) é o que realiza a normalização. Inicialmente, esta normalização é realizada por meio do cálculo do peso TF, ou seja, a frequência de cada termo dividida pelo total de termos do documento. Por exemplo, o peso TF de um termo que ocorre 5 vezes em um documento com 100 palavras é (5/100) = 0,05. Em seguida, a frequência inversa (IDF) dos documentos é calculada por meio da Equação 3.8:

$$IDF = log(\frac{N}{n_k}) \tag{3.8}$$

onde N é o número de documentos do córpus e n_k é o número de documentos em que o termo *k* ocorre no córpus. Por exemplo, seja um córpus com 1000 documentos e em 10 documentos ocorre o termo "futebol", o peso IDF é obtido por $log(\frac{1000}{10}) = 2$. Dessa forma, o valor de TF-IDF (TF x IDF) é utilizado para calcular o peso de cada termo nos documentos.

Uma técnica de reduzir a dimensionalidade da matriz e encontrar padrões associativos nos dados é aplicada após a normalização da matriz. Essa técnica é chamada de *Singular Value*

Tabela 3.18: Matriz de co-ocorrência de termos

	Texto 1	Texto 2	Texto 3	...	Texto N
Termo 1	2	1	0	...	...
Termo 2	0	3	1	...	...
Termo 3	1	0	2	...	...
⋮	...	...	...	...	...
Termo M	...	...	...	...	...

Decomposition (SVD) (Golub & Reinsch, 1970).

Com o SVD, uma matriz X (como a Tabela 3.18) normalizada pelo modelo TF-IDF é decomposta em um produto de outras três matrizes (X = TSD) sendo que:

- m = m $\leq$ min(t, d), número de dimensões;
- T = t x m, matriz de vetores singulares à esquerda;
- S = matriz diagonal m x m de valores singulares em ordem decrescente;
- D = m x d, matriz de vetores singulares à direita;

onde *t* é o número de termos (linhas), *d* o número de documentos (colunas) e X é uma matriz *t x d*.

Desta forma, a dimensão da matriz é reduzida por meio da eliminação das linhas e colunas correspondentes aos menores valores da matriz S, da mesma forma que as colunas da matriz T e as linhas da matriz D. Portanto, a redução da dimensão é realizada pela redução do número *m* de dimensões para um valor *k* (*k* < *m*) e, assim, a matriz reduzida S' afeta diretamente as dimensões das matrizes T e D, ou seja, o produto TS'D captura os elementos mais relevantes da matriz.

A partir da representação em forma de vetores permitida pela LSA, pode-se medir a similaridade de conceitos relacionados entre duas palavras ou sentenças. A Equação 3.9 desenvolvida por Landauer et al. (1997) calcula a similaridade:

$$sim(S_1, S_2) = cos(\mu(\vec{S_1}), \mu(\vec{S_2})) = \frac{\sum_{i=1}^{n} \mu_j(\vec{S_1})\mu_j(\vec{S_2})}{\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_1}))^2}\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_2}))^2}} \tag{3.9}$$

onde $\mu(\vec{S_1}) = \frac{1}{|S_i|}\sum \vec{u} \in S_i\vec{u}$ e, $\vec{u}$ é o vetor de palavras *u*. Uma medida de coerência textual geral pode ser obtida por meio da média dos cossenos para todos os pares de sentenças adjacentes S_i e S_{i+1}, como é visto na Equação 3.10 (Foltz et al., 1998):

$$coerencia(T) = \frac{\sum_{i=1}^{n-1} cos(S_i, S_{i+1})}{n-1} \tag{3.10}$$

Esse modelo LSA pode ser um bom modelo de comparação por algumas razões: 1) modelo completamente automático e com poucos parâmetros; 2) modela um aspecto da coerência

local, a similaridade das sentenças. Em contrapartida, tal modelo tem pontos negativos como: 1) modelo pode ser considerado caro, devido a dificuldade em determinar quantas dimensões diminuir; 2) modelo com baixa acurácia em relação aos outros trabalhos da literatura.

Para Louis & Nenkova (2012a), cada texto tem um propósito, como: explicar um conceito, narrar um evento, criticar uma ideia, etc. A partir disso, cada sentença em um texto tem uma meta comunicativa e a sequência de metas ajudam os autores a alcançarem o propósito do texto. Assim, as autoras apresentam um modelo para capturar a coerência a partir da dimensão da estrutura intencional. De acordo com Louis & Nenkova, essa estrutura intencional pode ser visualizada em produções sintáticos dos textos.

Segundo as autoras, o trabalho é baseado no fato que certos tipos de sentenças como perguntas e definições possuem estruturas sintáticas únicas e distinguíveis. Além disso, sentenças com estruturas sintáticas similares são prováveis de terem a mesma meta comunicativa e a regularidade na estrutura intencional é manifestada em produções sintáticas entre sentenças adjacentes.

A sintaxe é representada tanto como produções sintáticas quanto como uma sequência de nós (em uma representação de árvore sintática) com etiquetas morfossintáticas.

Em princípio, um estudo inicial foi realizado para confirmar que sentenças adjacentes em um texto exibem padrões de coocorrência sintática. Para isso, as autoras utilizaram árvores de análises padrão ouro do Penn Treebank (Marcus et al., 1993) e a unidade de análise foi um par de sentenças adjacentes (S_1, S_2). 99 documentos e 1727 pares de sentenças da Seção 0 de cada texto do córpus foram escolhidos para esse estudo.

Todas as produções que aparecem na análise sintática de alguma sentença foram enumeradas e todas as produções que aparecem menos do que 25 vezes foram excluídas, resultando em uma lista de 197 produções únicas. Assim, todos os pares (p_1, p_2)[10] de produções foram formados. Um exemplo de produção de uma sentença seria $S \rightarrow NP - SBJVP$.

Para cada par de produções, as autoras computaram: $c(p_1p_2)$ = número de pares de sentenças onde $p_1 \in S_1$ e $p_2 \in S_2$; $c(p_1\neg p_2)$ = número de pares onde $p_1 \in S_1$ e $p_2 \notin S_2$; $c(\neg p_1p_2)$ e $c(\neg p_1\neg p_2)$ são computados similarmente. Além disso, as autoras utilizaram o teste *Chi-square*[11] para entender se a conta observada em $c(p_1p_2)$ é significantemente (nível de confiança de 95%) maior ou menor do que o valor esperado se a ocorrência de p_1 e p_2 forem independentes.

Para o modelo de coerência, Louis & Nenkova descrevem 2 representações de estrutura sentencial utilizadas: *Productions* e *d-sequence*. Na representação de *Productions*, cada sentença é vista como o conjunto de produções gramaticais, $LadoEsquerdo \rightarrow LadoDireito$. O *LadoDireito* contém nós terminais/não terminais e o LadoEsquerdo contém somente nós não-terminais. Segundo as autoras, essa representação tem algumas desvantagens como o fato de alguma produção ter o lado direito muito longo e de conter informação somente sobre os nós que pertencem ao mesmo constituinte. Já na representação *d-sequence*, as autoras procuraram preservar mais informação sobre constituintes adjacentes da sentença. Na *d-sequence*, a árvore sintática é "truncada" na máxima profundidade *d* e as folhas da árvore resultante listadas da

[10] (p_1, p_2) e (p_2, p_1) são considerados pares distintos.
[11] http://www2.lv.psu.edu/jxm57/irp/chisquar.html

esquerda para a direita formam a representação *d-sequence*. A Figura 3.11 mostra um exemplo de truncamento no nível 2 (linha horizontal na árvore sintática) que representa *d-2 sequence* ou *depth-2 sequence*.

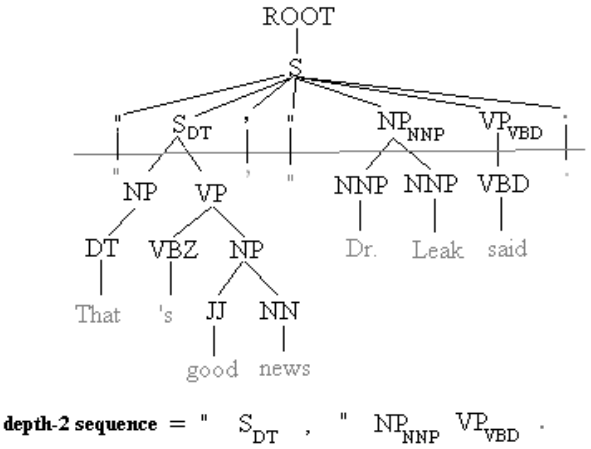

Figura 3.11: Exemplo de *d-sequence* (Louis & Nenkova, 2012a)

Todos os nós não-terminais da *d-sequence* possuem a informação do nó folha mais a esquerda que eles dominam. Essa informação é dada como sufixo desses nós não-terminais, por exemplo, os sufixos DT, NNP e VBD em S_{DT}, NP_{NNP} e VP_{VBD}. Segundo as autoras, essa informação traz a informação da sub-árvore abaixo dos nós presentes na *d-sequence*.

As sentenças são vistas como sequências de palavras sintáticas $(w_1, w_2,..., w_k)$, $k \leq p$, onde *p* é o tamanho da sequência e os w_i são os elementos que formam a *d-sequence*. Considerando o exemplo dado na Figura 3.11, as palavras sintáticas são: $w_1 =''; w_2 = S_{DT}; w_3 =, ; w_4 =''$ $; w_5 = NP_{NNP}; w_6 = VP_{VBD}$ e $w_7 = .$.

De acordo com Louis & Nenkova, o verbo principal de uma sentença é o centro para sua estrutura. Dessa forma, o parâmetro *d* é ajustado para ser maior do que o nível do verbo principal na árvore sintática.

Para Louis & Nenkova, o modelo de coerência local é baseado no estudo inicial do trabalho, ou seja, a coerência de um texto é caracterizada pela regularidade sintática em sentenças adjacentes. Assim, as autoras calcularam as probabilidades de pares de itens sintáticos pertencentes às sentenças adjacentes. Dessa forma, a coerência de um texto *T* que contém *n* sentenças (S_1,...,S_n) é computada de acordo com a Equação 3.11:

$$P(T) = \prod_{i=2}^{n} \prod_{j=1}^{|S_i|} \frac{1}{|S_{i-1}|} \sum_{k=1}^{|S_{i-1}|} p(S_i^j | S_{i-1}^k) \tag{3.11}$$

onde S_x^y indica o y-ésimo item de S_x. Os itens são Produções ou unigramas de palavras sintáticas.

As probabilidades condicionais são computadas com suavização segundo a Equação 3.12.

$$p(w_j|w_i) = \frac{c(w_i,w_j) + \delta_c}{c(w_i) + \delta_c * |V|} \tag{3.12}$$

onde w_i e w_j são itens sintáticos e $c(w_i, w_j)$ é o número de sentenças que contém o item w_i imediatamente seguido pela sentença que contém w_j. Já o tamanho do vocabulário de itens sintáticos é dado por $|V|$.

Para avaliar o modelo, pares de textos foram utilizados (um texto original e sua versão permutada). Além disso, o texto original é considerado sempre mais coerente do que a sua versão permutada. Assim, a acurácia do modelo é incrementada toda a vez que o texto fonte é identificado, ou seja, quando a probabilidade do texto original é maior do que a sua versão permutada.

O córpus utilizado para avaliar o modelo foi o mesmo utilizado em Barzilay & Lapata. Assim, 100 textos relacionados a Acidentes Aéreos e 99 textos com foco em Terremotos foram utilizados na avaliação. Além disso, 20 permutações para cada texto original foram geradas para formar os pares. Dessa forma, o melhor resultado obtido foi de 72,8% de acurácia com a utilização de *Productions* e 71,8% de acurácia para *d-sequence* com *d* = 2.

Segundo Li & Hovy (2014), a estrutura coerente de um texto é possível de ser descoberta usando representações de sentenças distribuídas aprendidas em um aprendizado profundo (*Deep Learning*).

Os autores consideraram uma abordagem chamada de *WINDOW* (Collobert et al., 2011) que agrupa sentenças. Um exemplo disso é mostrado na Tabela 3.19, onde exemplos positivos são janelas de sentenças selecionadas do texto original gerado por humanos, e exemplos negativos são gerados por substituição aleatória.

Tabela 3.19: Exemplo de textos coerente e incoerente.

João estava com muita fome. Ele não encontrou comida em casa. Assim, ele foi ao restaurante. Coerente (+): texto original	João estava com muita fome. A mãe comprou uma nova saia. ⟵ aleatório Assim, ele foi ao restaurante. Incoerente (-): substituição aleatória

As representações semânticas para termos e sentenças são obtidas por meio da otimização da estrutura de rede neural baseada nos exemplos positivos e negativos, como os mostrados na Tabela 3.19.

Os autores utilizaram duas representações de vetores para as sentenças: a representação gerada pelas redes neurais recorrentes e a representação gerada pelas redes neurais recursivas.

Assim, os autores propuseram um modelo de coerência que faz uso da abordagem *WINDOW* para treinar uma rede neural de três camadas (camada de entrada de nível de sentença, camada escondida e camada de saída) baseada em janelas de *L* sentenças.

Para avaliar a acurácia do modelo, os autores utilizaram a tarefa de ordenação de sentença de Barzilay & Lapata e definiram a pontuação de coerência para cada texto baseado em um algoritmo de convolução de sentenças. O córpus utilizado também foi o mesmo utilizado por Barzilay & Lapata, ou seja, textos relacionados a Acidentes Aéreos e a Terremotos.

Considerando as duas formas de representar as sentenças, a Tabela 3.20 mostra o resultado alcançando considerando cada assunto do corpus. A abordagem apresentada no trabalho de Li & Hovy produziu bons resultados. Entretando, obtê-los pode ser uma tarefa árdua devido a sua alta complexidade computacional. Além disso, a velocidade de treinamento e decodificação do aprendizado profundo é bastante lenta e isso é prejudicial principalmente na geração textual.

Tabela 3.20: Resultados obtidos de Li & Hovy (2014)

	Acidentes	**Terremotos**
Representação Recursiva	86,4%	97,6%
Representação Recorrente	84%	95%

O trabalho de Lin et al. (2015) apresenta um método de entropia máxima para modelar a coerência do texto. Esse método modela a coerência com características lexicais ao invés de características extraídas de todo o documento. Segundo os autores, esse método não faz uso de análise sintática e de resolução de correferência.

Os autores partem da premissa que em um texto coerente, as palavras de uma sentença são escolhidas de acordo com as sentenças anteriores. Assim, para modelar a coerência no documento *D*, o qual contém sentenças S_1, S_2,..., S_n, é necessário maximizar a função objetivo dada pela Equação 3.13:

$$p(D) = p(S_1, S_2,..., S_n) = p(S_1) * p(S_2|S_1) * ... * p(S_n|S_1, S_2,..., S_{n-1}) \tag{3.13}$$

onde *p(D)* é a probabilidade da coerência do documento *D* que é igual a probabilidade de cada sentença dada a sentença anterior do documento (p($S_n|$ S_1, S_2,..., S_{n-1})). Segundo os autores, a Equação 3.13 pode ser simplificada, como mostra a Equação 3.14:

$$p(D) \cong p(S_1) * p(S_2|S_1) * p(S_3|S_2) * ... * p(S_n|S_{n-1}) = \prod_{k=1}^{n} p(S_k|h) \tag{3.14}$$

onde *h* denota a probabilidade *a posteriore* de cada sentença do documento (história). Para melhorar a performance do modelo, a sentença também é simplificada como um vetor. Considerando a complexidade computacional, os autores usaram *bag of words*[12] para representar uma sentença no modelo. Assim, a Equação 3.14 foi simplificada como mostra a Equação 3.15:

$$p(D) \approx \prod_{k=1}^{n} p(BoWS_k|BoWh) \tag{3.15}$$

onde $BoWS_k$ é a *bag of words* da k-ésima sentença. De acordo com os autores, a *bag of words* pode ser convertida em um vetor de característica.

[12]É uma tabela, na qual as colunas representam os termos, palavras, léxicos ou outros tipos de atributos existentes nas mensagens, e os valores associados às colunas são referentes à frequência (ou presença) desses termos (ou atributos) nas sentenças.

Segundo Lin et al., um modelo de língua de entropia máxima (Rosenfeld, 1996) pode capturar mais informações. Devido a isso, os autores adicionaram características ao modelo de entropia máxima.

Para um modelo de língua de entropia máxima, a probabilidade de uma palavra *w* dado a história *h* é computada de acordo com a Equação 3.16:

$$p(w|h) = \frac{1}{Z(h)} * exp(\sum_i \lambda_i f_i(h,w)) \tag{3.16}$$

onde $Z(h)$ é o fator de normalização de história e f_i é a i-ésima função característica. Com a ideia de adicionar características ao modelo de entropia máxima, os autores combinaram as Equações 3.14, 3.15 e 3.16 para computar a probabilidade condicional de uma sentença, como mostra a Equação 3.17:

$$\begin{aligned} p(S_k|S_{k-1}) &= \frac{1}{Z(S_{k-1})} * exp(\sum_i \lambda_i f_i(S_{k-1},S_K)) \\ &= \frac{1}{Z(BoWS_{k-1})} * exp(\sum_i \lambda_i f_i(BoWS_{k-1},BoWS_K)) \end{aligned} \tag{3.17}$$

onde $p(S_k|S_{k-1})$ é a probabilidade de coerência de uma sentença S_k dada a história S_{k-1}. De acordo com os autores, a *bag of words* foi utilizada para representar uma sentença por causa da dificuldade de representá-la.

Para medir a acurácia do modelo, os autores utilizaram a tarefa de ordenação de sentença e o mesmo córpus de Barzilay & Lapata (2008). O modelo obteve 87,7% de acurácia para textos relacionados a Acidentes e 97,3% de acurácia para textos cujo foco era em terremotos.

Este capítulo mostrou os principais trabalhos relacionados a coerência local, sendo que a metodologia baseada em entidades é a mais utilizada entre os trabalhos focados na coerência local. Os trabalhos que fazem uso de pouco conhecimento linguístico e mais estatístico/matemático vem ganhando espaço na distinção entre os textos coerentes e os incoerentes. Já a abordagem discursiva, se mostrou próspera e com bom desempenho, apesar de pouca explorada. Outro ponto importante, verificado nos trabalhos relacionados, é a versatilidade do modelo de Grade de Entidades de Barzilay & Lapata. Tal versatilidade vem da junção de várias informações relacionadas a coerência, de forma a criar modelos de alta acurácia na distinção entre textos coerentes e incoerentes. Assim, o bom desempenho apresentado pela abordagem discursiva e a versatilidade do modelo de Grade de Entidades motivaram o desenvolvimento desta tese voltada especificamente para a sumarização multidocumento, algo que não foi encontrado na literatura.

3.1 Trabalhos Relacionados a Qualidade Linguística

Nessa seção serão mostrados alguns trabalhos que listaram e definiram elementos que influenciam a coerência local.

Para Koch & Travaglia (2002), a coerência é relacionada a possibilidade de encontrar um significado para um texto de acordo com alguns fatores:

- Elementos Linguísticos - serve como indicações para estimular inferências e a aquisição de orientação argumentativa obtida de expressões que compõe um texto.

- Conhecimento de Mundo - caso em que um texto possui um assunto conhecido para o leitor. Assim, o leitor compreenderá o sentido do texto, e isso faz o texto coerente.

- Conhecimento Compartilhado - conhecimento compartilhado do escritor para o leitor/ouvinte. Para um texto ser coerente é necessário um equilíbrio entre a velha e a nova informação. Um texto só com novas informações seria incompreensível, desde que o leitor não possua conhecimento sobre tais informações. Por outro lado, um texto com apenas informações velhas iria torná-lo redundante.

- Inferências - usando o conhecimento de mundo, o receptador da mensagem estabelece uma relação não explícita com o texto, onde o receptador tenta compreender e interpretar a mensagem.

Otterbacher et al. (2002) estudou os possíveis problemas relacionados a coesão de sumários multidocumento extrativos e sugere revisões (soluções) para melhorar a coesão. Os autores apresentaram uma análise baseada em córpus de sumários automáticos extrativos multidocumento gerados pelo sumarizador MEAD (Radev et al., 2002). Os sumários desse córpus foram manualmente revisados. Os autores discutiram a viabilidade de melhorar automaticamente os sumários, e eles também criaram uma taxonomia dos problemas relacionados a coesão.

A taxonomia foi dividida em 5 categorias pragmáticas relacionadas a coesão textual em sumários multidocumento, tais categorias são: Discurso, Indentificação de Entidades, Expressões Temporais, Gramática e Propriedades de Localização.

Discurso foca no relacionamento entre as sentenças dos sumário e no relacionamento entre os elementos textuais. Nessa categoria, os autores consideram alguns aspectos que podem levar a problemas de coesão em sumários multidocumento, como: Mudança de Tópico, Falta de Propósito, Contradição, Redundância e Sentenças Condicionais.

Segundo os autores, a mundança de tópico é a troca de uma sentença por outra; o tema muda de repente. Devido a isso, a adição de uma sentença transitória ou uma frase pode resolver esse problema.

Para Otterbacher et al., em sumários há sentenças que faltam propósitos. Entretanto, a adição de sentenças que motiva um propósito no segmento problemático resolveria tal problema.

A contradição está relacionada a alguma informação, em uma dada sentença, que contrasta com uma ou mais sentenças anteriores. Um marcador discursivo tal como "entretanto"ou "em contraste"modifica o marcador discursivo existente.

Redundância ocorre quando uma sentença contêm informações reportadas anteriormente. Para Otterbacher et al., uma ação possível para resolver esse problema é apagar o constituinte

redundante (elemento não nuclear dos sintagmas nominais, dos sintagmas preposicionais ou de relativas clausais).

De acordo com Otterbacher et al., eventos em uma dada sentença são condicionados por eventos em outra sentença. Assim, uma ação para resolver esse problema seria modificar as sentenças, como: Se (sentença 1), (sentença 2). Além disso, o tempo verbal pode ser modificado para o condicional.

Segundo os autores, a identificação de entidades requer a resolução de expressões referenciais, desde que o leitor necessite identificar cada entidade mencionada em um sumário. Assim, 9 problemas foram encontrados em sumários relacionados a essa categoria. Os problemas foram: Entidade não Especificada, Mal uso de quantificador, Entidade muito restrita, Entidade repetida, Anáfora descoberta, Mal uso de artigo definido, Mal uso de artigo indefinido, Falta de artigo, Falta entidade.

Entidade não especificada é uma entidade recentemente mencionada que não possui descrição ou um acrônimo sem explicação. Para resolver esse tipo de erro, adiciona-se um nome completo ou um título para uma nova entidade ou expande o acrônimo. Segundo os autores, O erro Entidade não especificada foi o mais frequente na categoria de identificação de entidades, com 38% de ocorrência nos casos.

O problema do mal uso de artigo definido pode também ser resolvido pela adição de um artigo definido se a entidade já tinha sido mencionada ou pela adição de um artigo indefinido, se a entidade é nova.

A categoria de Expressão Temporal é dada pelo relacionamento temporal correto entre eventos. Os autores identificaram 5 tipos de possíveis problemas que se enquadram nessa categoria: Ordenação Temporal, Tempo do Evento, Repetição do Evento, Sincronismo e Anacronismo.

A Ordenação Temporal é relacionada ao estabelecimento correto das relacões temporais entre eventos (ou relacionado a um evento anterior). Em caso de problema, os autores recomendam adicionar expressões de tempo, apagar expressões de tempo inapropriadas ou modificar uma expressão de tempo já existente. Com um total de 89% de ocorrências, os erros de Ordenação Temporal foram o mais frequentes na categoria Temporal.

Alguns problemas gramaticais têm sido identificados no córpus usado por Otterbacher et al.. Entre esses problemas estão: Sentença *Run-on*, Verbos incompatíveis, Falta de pontuação, Sintaxe inadequada, Parênteses, Subtítulos/títulos, Mal uso de advérbios.

Segundo Otterbacher et al., uma sentença *Run-on* é uma sentença muito longa. Assim, os autores recomendam dividir sentenças longas em duas sentenças separadas e apagar a conjunção presente na sentença longa. Esse problema foi o mais frequente da categorias de problemas gramaticais, com 35% dos erros.

O problema de parênteses está relacionado ao uso inapropriado dos próprios parênteses. Assim, os autores sugerem apagar tais parênteses.

Propriedades de localização é um tipo de revisão relacionado a localização correta dos eventos. Essas propriedades podem ser: Localização de eventos, Colocação, Mudança de localização, *Place/Source Stamp*.

Localização de eventos especifica um lugar onde o evento vai acontecer. Assim, uma possível revisão seria adicionar sintagmas preposicionais que indicam lugares (cidade, estado, país).

Colocação é relacionada a dois ou mais eventos que ocorrem nos mesmo lugar. Dessa forma, os autores sugerem adicionar um sintagma preposicional ou um advérbio que indica a colocação.

De acordo com Otterbacher et al., a categoria Discurso teve 34% de todas as revisões realizadas no córpus, seguida por Entidades, com 26%, 22% de Expressões temporais, 12% de problemas gramaticais e 6% de Propriedades de localização.

Pitler et al. (2010) avaliaram a Qualidade Linguística (QL) dos sumários gerados por sumarizadores automáticos multidocumento. Os autores analizaram a forma como os diferentes tipos de características podem ajudar o ranque dos sumários.

De acordo com Pitler et al., há alguns aspectos de QL que são relevantes para a geração de sumários automáticos e podem ser usados em avaliações manuais. Esses aspectos são: Gramaticalidade, Sem redundância, Claridade referencial, Foco e Estrutura/Coerência.

Gramaticalidade está relacionada ao sistema de formatação do texto, erros relacionados a letras maiúsculas e sentenças não gramaticais (falta de alguns componentes textual) que torna a leitura dos textos difíceis.

O aspecto de Sem redundância considera que repetições desnecessárias no sumário não poderiam ocorrer. Segundo os autores, repetições desnecessárias podem acontecer com sentenças inteiras, fatos repetidos e o uso repetido de um substantivo ou sintagma nominal, quando um pronome seria suficiente.

Claridade referencial é relacionada a identificação de quem ou do que, o pronome ou sintagma nominal refere-se no sumário.

O Foco é relacionado a existência de um assunto principal em um sumário; sentenças poderiam conter somente informações que são relacionadas ao assunto principal.

Para Pitler et al., Estrutura e Coerência de um resumo estão relacionados a este ser bem estruturado e organizado.

Além dos aspectos descritos anteriormente, Pitler et al. citaram alguns fatores que influenciaram a QL dos textos em geral, por exemplo: escolha de palavras, formas referênciais de entidades (entidades nomeadas) e coerência local (dispositivos coesivos e continuidade).

Kaspersson et al. (2012) investigou erros linguísticos que ocorrem em sumários extrativos gerados de um único documento. O foco deste trabalho foi nos erros discursivos, tais como expressões de referências sem antecedentes, e como as unidades textuais nos sumários são conectadas. Além disso, os autores também investigaram como os diferentes níveis de resumo do texto e diferentes gêneros influenciam certos tipos de erros.

De acordo com Kaspersson et al., um estudo foi realizado para encontrar tipos de erros em textos sumarizados que afetam negativamente a coesão, coerência e a legibilidade textual. Assim, os autores consideraram textos de 3 diferentes gêneros: jornalísticos, científicos e textos oficiais de governo.

Os erros encontrados foram agrupados em 3 categorias: Erro de referência anafórica, Au-

sência de coesão ou de contexto e Quebra de referência anafórica.

Erro de referência anafórica é relacionada a uma expressão anafórica no sumário que refere a um antecedente errado, dado que o antecedente correto não foi extraído do texto fonte, o qual originou o referido sumário. Essa categoria tem 3 sub-categorias: Sintagma nominal, Nomes próprios e Pronomes.

Ausência de coesão ou de contexto é relacionada a falta de algum elemento coesivo ou de contexto do sumário.

Quebra de referência anafórica acontece quando uma expressão anafórica de um sumário não tem seu antecedente, porque tal antecedente não foi extraído do texto fonte. As sub-categorias são: Sintagma nominal, Nomes próprios e Pronomes

Segundo Kaspersson et al., os erros mais significantes são: Erro de referência anafórica relacionada a pronome, Ausência de coesão ou de contexto, Quebra de referência anafórica relacionada a sintagmas nominais e Quebra de referência anafórica relacionada a pronomes.

Friedrich et al. (2014) apresentou um córpus de sumários chamado LQVSumm com vários tipos de erros de QL manualmente anotados. Esses sumários foram automaticamente criados para a tarefa compartilhada, Sumarização Guiada, da TAC 2011 (Owczarzak & Dang, 2010). Os autores identificaram 2 classes de erros: uma delas considera as menções de entidades e a outra considera cláusulas (sentenças). A primeira é relacionada a referência ou problemas de correferência. A última involve erros não gramaticais ou redundância.

Para os autores, no nível de entidades, os tipos de erro são: Primeira menção sem explicação, Menção subsequente com explicação, Sintagma nominal definido sem referência a menções anteriores, Sintagma nominal indefinido com referência a menções anteriores, Pronome sem antecedente, Pronome com antecedente enganoso, e Acrônimos sem explicação.

Primeira menção sem explicação é designada para a primeira menção de uma entidade para a qual não há uma referência clara para o leitor. Por exemplo, na sentença, "Paulo comprou brinquedos para crianças pobres." para a entidade "Paulo" não há explicação suficiente.

Menção subsequente com explicação é relacionada a menções de entidade que já foram referenciadas no texto, mas ainda apresentam uma explicação inapropriada. Por exemplo, considere a Figura 3.12, a qual mostra parte de um sumário multidocumento da coleção 21 do córpus CST-News (Aleixo & Pardo, 2008; Cardoso et al., 2011) que possui as entidades "Nelson Jobin" e "Infraero" como menções subsequentes com explicação.

(S1) O ministro da Defesa, Nelson Jobim, decidiu que será realizada uma reforma definitiva na pista principal de Guarulhos, o mais rápido possível, de acordo com a assessoria do ministério da Defesa.
(S2) De acordo com estudos apresentados pela Infraero, será possível realizar a obra definitiva em três etapas, sem que seja necessário fechar a pista neste momento.
(S3) O ministro da Defesa, Nelson Jobim, anunciou a reforma que, segundo estudos da Empresa Brasileira de Infra-Estrutura Aeroportuária (Infraero), a reforma poderá ser feita sem que a pista seja interditada.

Figura 3.12: Exemplo de Menção subsequente com explicação

Sintagma nominal definido sem referência a menções anteriores ocorre quando sintagma nominal definido é usado para referenciar a primeira menção de uma entidade no texto. Por exemplo, "a empresa Petrobras" poderia ser usado em um sumário no qual "a empresa" tenha sido mencionado.

Sintagma nominal indefinido com referência a menções anteriores ocorre quando sintagma nominal indefinido é usado em uma entidade já mencionada no texto. Por exemplo, o sintagma nominal "uma empresa Petrobras" não é apropriado se a mesma empresa já tenha sido mencionada no sumário.

Pronome sem antecedente ocorre quando não há antecedente que combina em gênero e número. Por exemplo, a Figura 3.13 mostra parte de um sumário multidocumento da coleção 16 do córpus CSTNews. Na Figura 3.13, o pronome "Ele" (sublinhado) não possui algum possível antecedente.

(S1) Depois de iniciados os processos, as renúncias não têm mais o efeito de paralisar as investigações.
(S2) Ele vai instaurar o processo contra os deputados envolvidos com a máfia dos sanguessugas amanhã, às 10h30.

Figura 3.13: Exemplo de Pronome sem antecedente

Pronome com antecedente enganoso ocorre quando uma expressão anafórica refere a um antecedente enganoso e o seu antecedente correto não está no sumário. Por exemplo, a Figura 3.14 mostra parte de um sumário multidocumento da coleção 27 do córpus CSTNews.

(S1) Aos 27, Kaká arriscou de muito longe e Ronaldinho colocou o desviou o chute.
(S2) A 20cm da linha de fundo ele deu dois dribles humilhantes no zagueiro equatoriano e cruzou para Elano, que fez o quarto, aos 37.

Figura 3.14: Exemplo de Pronome enganoso

Na Figura 3.14, o pronome "ele" (na segunda sentença e sublinhado) aparentemente refere-se a "Kaká" (na primeira sentença), mas, no texto fonte, o pronome refere-se a "Robinho", o qual não está no sumário.

Acrônimos sem explicação ocorre quando eles não são conhecidos e não são explicados. Por exemplo, a sentença "Os candidatos José Maria Eymael (PSDC) e Ruy Pimenta (PCO) não pontuaram.", que faz parte de um sumário da coleção 2 do córpus CSTNews, apesenta 2 acrônimos sem explicação "PSDC" e "PCO".

Friedrich et al. também propôs uma anotação a nível clausal ou sentencial. Esta anotação foi feita em trechos arbitrários, isto é, os tipos de erros são marcados considerando relações entre

trechos. De acordo com os autores, os erros do nível clausal são: Sentença incompleta, Inclusão de datas, Outra forma não gramatical, Sem relacionamento semântico, Informação redundante, e Sem relação no discurso.

Uma Sentença incompleta ocorre devido ao uso de compressão da sentença ou o truncamento realizado para não exceder o tamanho máximo permitido do sumário. Por exemplo, a sentença "Um foi morto em um quarto e outros foram assassinados em uma sala de aula, de acordo com o chefe da polícia do campus, W." está incompleta.

Para Friedrich et al., a Inclusão de datas em um sumário não é desejado. Por exemplo, "GEORGETOWN, Pennsylvania 05-10-2006 16:53:53".

Outra forma não gramatical considera todos os outros casos não gramaticais, tais como a falta de espaço em branco e pontuação incorreta.

Sem relacionamento semântico ocorre quando sentenças não possuem alguma relação semântica. Por exemplo, a Figura 3.15 mostra as sentenças S1 e S2 que não possuem relacionamento semântico entre si.

(S1) Na quarta-feira, o presidente da empresa, Marco Antonio Bologna, havia declarado que o avião passara por checagem justamente no dia 13 e estava em perfeito estado.
(S2) Foi um gesto de raiva e indignação com o comportamento de certo tipo de noticiário da imprensa diante da tragédia de 200 famílias, que acabou tirando conclusões precipitadas sobre o fato - tentou explicar o assessor especial da Presidência da República.

Figura 3.15: Exemplo de sentenças sem relacionamento semântico

Informação redundante ocorre quando duas ou mais sentenças expressam a mesma informação. Por exemplo, na Figura 3.16, há informação redundante nas sentenças S1 e S2.

(S1) Presença constante na cena política brasileira nas últimas quatro décadas, o senador Antonio Carlos Magalhães (DEM-BA) morreu na manhã desta sexta-feira, em São Paulo, vítima de insuficiência cardíaca.
(S2) O senador Antonio Carlos Magalhães (DEM-BA) morreu às 11h40 de hoje, aos 79 anos, em São Paulo, em decorrência de falência de múltiplos órgãos secundária à insuficiência cardíaca.

Figura 3.16: Exemplo de informação redundante

O erro Sem relação no discurso, em particular, acontece quando um conectivo discursivo explícito ("e", "mas", "porque", ...) está sendo utilizado inapropriadamente ao contexto do sumário. Por exemplo, o conectivo "E" na segunda sentença da Figura 3.17 não está apropriado.

Nessa seção, vários aspectos que podem influenciar a QL foram identificados e definidos por pesquisas da literatura. Isso mostra a relevância e a complexidade desse estudo, o qual tem

(1) O advogado de Luíz não pôde ser encontrado para falar sexta à noite.
(2) E a pessoa que coopera chega pela primeira vez a maior recompensa.

Figura 3.17: Exemplo de informação redundante

por objetivo dar suporte a sistemas de geração textual. Além disso, estas pesquisas influenciam diretamente a escolha de aspectos que podem afetar a QL de sumários automáticos multidocumento utilizados nesta tese. Assim, os aspectos que melhor se encaixam no contexto da sumarização multidocumento foram redefinidos e adaptados para a tarefa de anotação realizada neste trabalho. Tais redefinições, adaptações e a própria tarefa de anotação de erros que podem afetar a QL dos sumários multidocumentos serão descritos no Capítulo 6.

Capítulo

4

Adaptação dos Métodos da Literatura

Com o intuito de estabelecer um estudo comparativo e validar esta tese, foi necessário implementar os trabalhos da literatura considerados relevantes. Essa relevância foi baseada nas técnicas propostas e no impacto das mesmas na área de avaliação da coerência local. Nessa fase de implementações, foram consideradas adaptações dos modelos da literatura, pois todos os modelos escolhidos foram originalmente aplicados a corpora da língua inglesa.

Neste cenário, os trabalhos de Freitas (2013) e de Silva & Feltrim (2015) são os únicos conhecido para o Português, além do trabalho desta tese. Freitas (2013) adaptou o modelo de Grade de Entidades de Barzilay & Lapata para ser aplicado a textos escritos por alunos de graduação. Silva & Feltrim (2015) combinou o modelo de Grade de Entidades de Barzilay & Lapata com informações provenientes de estruturas retóricas para gerarem mensagens que indiquem quebras de linearidade em partes específicas dos resumos oriundos dos trabalhos de final de curso de alunos de graduação.

Em geral, adaptar os modelos propostos na literatura permitiu verificar a performance destes quando aplicados em um córpus de sumários multidocumento. Tais modelos serão comparados com suas versões incrementadas com conhecimento discursivo e com modelos puramente discursivos (desenvolvidos neste trabalho - ver Capítulo 5). Essa comparação será feita tanto na tarefa de ordenação de sentenças (visto no Capítulo 3) quanto na tarefa de identificação de erros da qualidade linguística (maiores detalhes no Capítulo 6) presentes nos sumários multidocumento.

A escolha dos modelos foi definida baseada nas 3 abordagens (entidades, discursivas e estatística/matemática) vista no Capítulo 3, ou seja, pelo menos um trabalho para cada abordagem foi implementado. Dentro de cada abordagem, a escolha dos modelos foi realizada levando em consideração: i) a importância do modelo na literatura; ii) um tempo factível de implementação do modelo; iii) se o mesmo era possível de ser utilizado em outro idioma; iv) se haveria recursos (*parsers* sintáticos ou discursivos, córpus, etc.) disponíveis para o português, de forma que o

modelo pudesse ser utilizado como um todo.

4.1 Modelo *Latent Semantic Analysis* (LSA)

O primeiro trabalho escolhido foi o de Foltz et al. (1998), o qual foi uma das primeiras propostas de avaliação da coerência local de forma automática. Baseado na *Latent Semantic Analysis*, esse modelo é um dos poucos que não faz uso de recursos externos (por exemplo, um *parser* sintático) como parte integrante do processo de avaliação da coerência. Tais características foram relevantes para o nosso propósito.

Esse modelo usa a similaridade entre as sentenças como um aspecto da coerência local. Para isso, os autores utilizaram a média dos cossenos para todos os pares de sentenças do texto. Assim, a coerência de um texto foi calculada seguindo a Equação 3.10 dada no Capítulo 3 e replicada em 4.1 . Desse mesmo modo, foi implementada e aplicada essa equação de coerência para cada um dos sumários do córpus CSTNews.

$$coerencia(T) = \frac{\sum_{i=1}^{n-1} cos(S_i, S_{i+1})}{n-1} \tag{4.1}$$

Para a implementação desse modelo, o pacote de rotinas do *Python* chamado *Scikit-Learn*[1] foi utilizado. Por meio desse pacote, foi possível calcular, por exemplo, a similaridade do cosseno entre as sentenças.

Nesse modelo, qualquer tipo de acentos foi retirado e os "ç" foram substituídos por "c" nas palavras que compõe as sentenças do sumário. Essa modificação foi feita devido a erros de codificação textual, os quais atrapalhavam o processamento do modelo.

A equação de coerência gera um valor, o qual foi denominada de Valor de Coerência, para cada sumário. Esse valor foi utilizado para calcular a acurácia do modelo na tarefa de ordenação de sentença.

Na Seção 4.5, os detalhes dos resultados desse modelo na tarefa de ordenação de sentenças serão mostrados.

4.2 Modelo de Grade de Entidades

Para a implementação do modelo de Grade de Entidades de Barzilay & Lapata (2008), a extração e a análise sintática de entidades feita pelo *parser* PALAVRAS Bick (2000) foram utilizadas.

A Figura 4.1 ilustra todos os passos utilizados por nós no desenvolvimento do Modelo de Grade de Entidades.

Inicialmente, um texto fonte ou um sumário é analisado morfossintática e sintaticamente pelo parser PALAVRAS. O PALAVRAS pode gerar um arquivo de análise em até três formatos. Nós utilizamos o terceiro, ou seja, o formato *Tiger*. Esse formato foi utilizado devido a sua

[1] http://scikit-learn.org/stable/

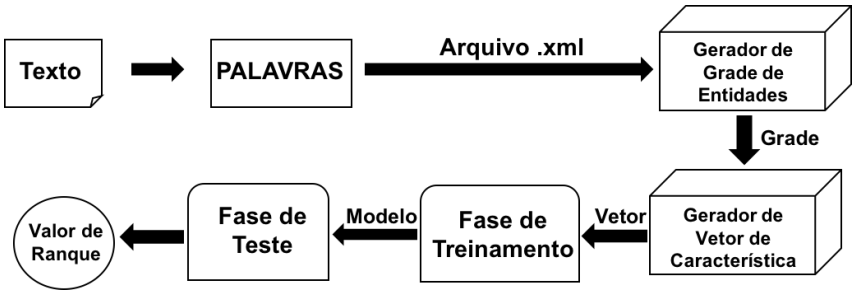

Figura 4.1: Estrutura de desenvolvimento do Modelo de Grade de Entidades

organização ser mais legível (ver Figura 2.15), a qual facilitou o processamento do Gerador de Grade de Entidades.

O Gerador de Grade de Entidades é um módulo que processa o arquivo XML produzido pelo PALAVRAS. Esse módulo extrai as entidades (núcleos dos sintagmas nominais) com suas funções sintáticas (S - Sujeito; O - Objeto; X - qualquer função sintática que não seja sujeito e nem objeto) presentes no texto e cria uma grade de entidades representada por uma matriz (linhas representam as sentenças e as colunas representam as entidades do texto).

A Figura 4.2 exemplifica parte de uma grade criada pelo Gerador de Grade de Entidades. Essa grade é de um texto cujo o assunto é sobre a Liga Mundial de Voleibol da coleção número 8 do córpus CSTNews (Cardoso et al., 2011; Aleixo & Pardo, 2008).

-	liga	ano	Tampere	classificação	Finlândia	seleção	...
1	O	-	X	-	O	S	...
2	O	-	-	-	-	-	...
3	-	-	-	X	-	-	...
4	-	-	-	X	-	S	...
5	-	-	-	-	-	-	...
6	-	S	-	-	-	-	...

Figura 4.2: Exemplo de uma grade de entidades.

Da mesma forma que o modelo original de Barzilay & Lapata (2008) faz uso de 3 informações (sintática, saliência e correferência) para criar as grades de entidades e, consequentemente, produzir diferentes variações do modelo, o Gerador de Grade de Entidades também considera tais informações.

A informação sintática está presente no modelo de Grade de Entidades (SINTÁTICO+), quando as funções sintáticas de cada entidade são usadas no preenchimento da grade. A Figura 4.2 é um exemplo de uma grade de entidades que utiliza essa informação.

Quando uma versão do modelo de Grade de Entidades não faz uso da informação sintática (SINTÁTICO-), a grade registra apenas a ocorrência de uma determinada entidade em uma sentença. Originalmente, esse registro é feito com o uso do símbolo “X”, o qual mantivemos

na nossa implementação. Por exemplo, a Figura 4.3 mostra a versão da grade exibida na Figura 4.2 sem informação sintática.

-	liga	ano	Tampere	classificação	Finlândia	seleção	...
1	X	-	X	-	X	X	...
2	X	-	-	-	-	-	...
3	-	-	-	X	-	-	...
4	-	-	-	X	-	X	...
5	-	-	-	-	-	-	...
6	-	X	-	-	-	-	...

Figura 4.3: Exemplo de uma grade de entidades sem informação sintática.

Segundo Barzilay & Lapata, uma entidade é considerada saliente quando a sua frequência é igual ou maior do que 2 em um texto. Assim, no modelo de Grade de Entidades, a informação de saliência é usada para formar uma nova grade composta somente com entidades consideradas salientes (SALIÊNCIA+). Por exemplo, a grade de saliência da Figura 4.4 é uma versão da grade sintática da Figura 4.2.

-	liga	classificação	seleção	...
1	O	-	S	...
2	O	-	-	...
3	-	X	-	...
4	-	X	S	...
5	-	-	-	...
6	-	-	-	...

Figura 4.4: Exemplo de uma grade de entidades com informação sintática e de saliência.

Segundo as autoras do modelo original, um resolvedor automático de correferência para o Inglês foi utilizado para tratar a informação de correferência dos sintagmas nominais. Infelizmente, não foi encontrada uma ferramenta robusta e disponível de resolução de correferência para o Português.

Com o intuito de evitar o aumento da esparsidade da grade (com um possível acréscimo de colunas na grade), o agrupamento de todas os núcleos iguais dos sintagmas nominais foi realizado, com o intuito de serem utilizados em uma única coluna da grade. Dessa forma, o

modelo de Grade de Entidades implementado nesta tese não possui uma versão que faz uso da informação de correferência da mesma forma que foi feito no modelo original.

Com a grade formada, o Gerador de Vetor de Característica calcula as probabilidades das transições de entidades entre as sentenças. A probabilidade é calculada por meio da razão entre a frequência de cada tipo de transição e o total de transições. Essas probabilidades irão compor o vetor de característica do texto. Como exemplo, a Figura 4.5 mostra o vetor de característica da grade de entidades da Figura 4.2.

s s	s o	s x	s -	o s	o o	o x	o -	x s	x o	x x	x -	- s	- o	- x	- -
0	0	0	0,066	0	0,033	0	0,066	0	0	0,033	0,066	0,066	0	0,033	0,63

Figura 4.5: Vetor de Característica

Na grade da Figura 4.2, há 2 (duas) transições para cada um dos tipos **[s -]**, **[o -]**, **[s -]**, **[- s]**; 1 (uma) transição para os tipos **[o o]**, **[x x]**, **[- x]**; 19 (dezenove) transições para o tipo **[- -]** e os outros tipos de transição não tiveram nenhuma ocorrência. Além disso, essa grade possui um total de 30 transições. Assim, por exemplo, a probabilidade da transição do tipo **[s -]** é de 0,066.

O vetor de característica para a grade de entidades da Figura 4.3, a qual não utiliza a informação sintática, é mostrada na Figura 4.6.

x x	x -	- -	- x
0,066	0,2	0,63	0,1

Figura 4.6: Vetor de Característica de grade sem informação sintática

Para cada texto/sumário, temos um vetor de característica e cada vetor de característica é uma instância para a Fase de Treinamento. Essa fase utiliza o pacote de aprendizado de máquina SVM^{light} com a opção de ranqueamento. Por meio desse treinamento, um modelo preditivo é gerado, e este será usado na Fase de Teste que irá contabilizar o Valor de Ranque para cada novo texto/sumário[2].

Os experimentos e os resultados obtidos desse modelo serão apresentados na Seção 4.5.

4.3 Modelo Baseado em Grafo

A principal característica desse Modelo Baseado em Grafo de citetguinaudeau2013 é a proposta de eliminar a parte de aprendizado de máquina do modelo de Grade de Entidades. Assim, com o objetivo de verificar o comportamento desse modelo no córpus CSTNews. A Figura 4.7 mostra todos os módulos criados para a implementação desse modelo.

Para esse modelo foi utilizado o parser PALAVRAS e também o Gerador de Grade de Entidades do modelo de Grade de Entidades implementado e descrito na Seção anterior.

[2]Os textos/sumários da Fase de Testes são diferentes dos usados na Fase de Treinamento

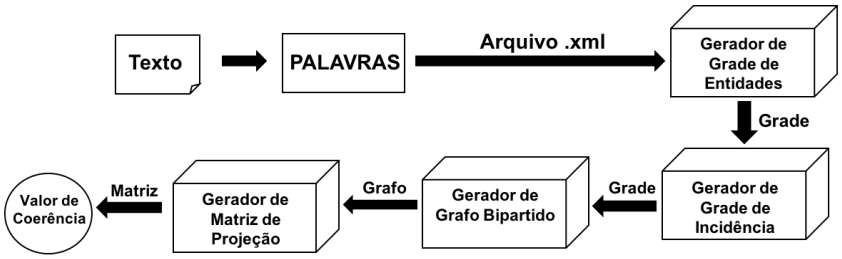

Figura 4.7: Estrutura de desenvolvimento do Modelo de Grafo

O Gerador de Grade de Incidência recebe uma grade de entidades sintática e substitui as funções sintáticas pelos seus pesos correspondentes, ou seja, **S** será substituído pelo peso **3**, **O** por **2**, **X** por **1** e **-** por **0**. Esse procedimento forma a chamada Grade de Incidência. A Figura 4.8 ilustra a transformação da grade de entidades da Figura 4.2 em sua respectiva grade de incidência.

	liga	ano	Tampere	classificação	Finlândia	seleção	...
s1	O	-	X	-	O	S	...
s2	O	-	-	-	-	-	...
s3	-	-	-	X	-	-	...
s4	-	-	-	X	-	S	...
s5	-	-	-	-	-	-	...
s6	-	S	-	-	-	-	...

	liga	ano	Tampere	classificação	Finlândia	seleção	...
s1	2	0	1	0	2	3	...
s2	2	0	0	0	0	0	...
s3	0	0	0	1	0	0	...
s4	0	0	0	1	0	3	...
s5	0	0	0	0	0	0	...
s6	0	3	0	0	0	0	...

Figura 4.8: Grade de Entidades transformada em Grade de Incidência

O módulo Gerador de Grafo Bipartido cria um grafo para a Grade de Incidência de entrada. Esse grafo gerado facilita tanto a visualização das ligações entre sentenças e entidades quanto a criação de matrizes de projeções *one mode* P_U, P_W e P_{Acc} . A Figura 4.9 mostra o grafo da grade de incidência da Figura 4.8.

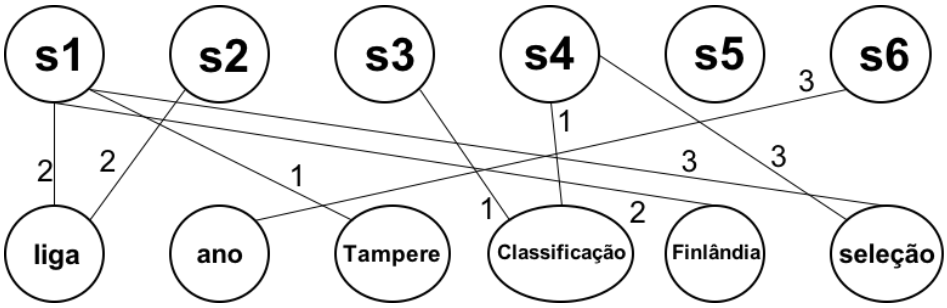

Figura 4.9: Grafo resultante do Gerador de Grafo Bipartido

Com o grafo gerado, o Gerador de Matriz de Projeção montará as respectivas matrizes de projeção *one mode*.

Por fim, um valor de coerência para cada sumário é calculado por meio do módulo Valor de Coerência, o qual faz uso da Equação 3.7 replicada em 4.2. Esse módulo também foi implementado prevendo a utilização do dado de distância entre duas sentenças com entidades em comum, segundo o modelo original.

$$\begin{aligned} LocalCoherence(T) &= AvgOutDegree(P) \\ &= \frac{1}{N} \sum_{i=1..N} OutDegree(s_i) \end{aligned} \tag{4.2}$$

4.4 Modelo Baseado em Padrões Sintáticos

O trabalho de Louis & Nenkova (2012b) propõe avaliar a coerência local por meio de padrões sintáticos entre as sentenças adjacentes. Tal abordagem foi considerada interessante, pois a mesma se distancia da abordagem do modelo de Grade de Entidades que até então era a base de vários trabalhos. Assim, esse modelo foi implementado com o objetivo de verificar a sua performance na avaliação da coerência de sumários multidocumento. Todas as etapas da implementação desse modelo são ilustradas na Figura 4.10.

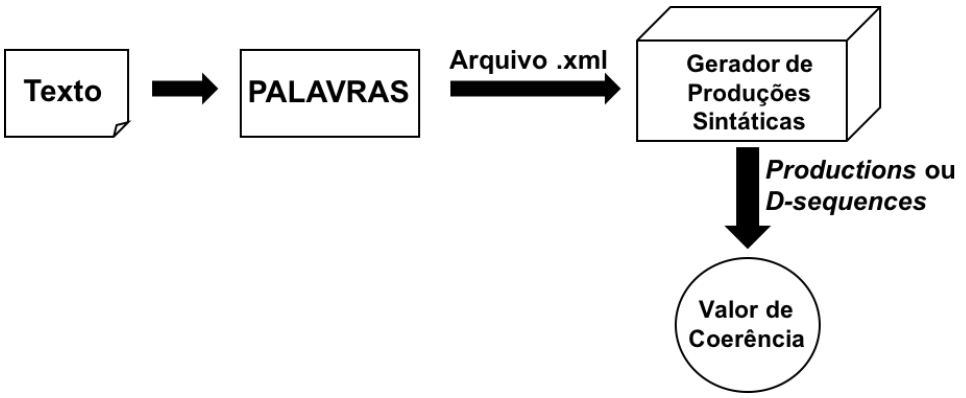

Figura 4.10: Estrutura de desenvolvimento do Modelo de Padrões Sintáticos

Para a adaptação desse modelo, a análise morfossintático do *parser* PALAVRAS foi usado para extrair as expressões sintáticas. Tal procedimento de obtenção das expressões foi implementado no Gerador de Produções Sintáticas.

O Gerador de Produções Sintáticas consegue obter as expressões tanto do tipo *Productions* (conjunto de produções gramaticais) quanto do tipo *d-sequences* (produções gramaticais que preservam mais informação sobre constituintes adjacentes da sentença). Para isso, esse gerador faz o processamento de montar as expressões sintáticas por meio do processamento na árvore sintática dada pelo arquivo do PALAVRAS. Além disso, as expressões de *d-sequence* para dois níveis foram extraídas, ou seja, expressões sintáticas até o nível 1 e o 2 da árvore sintática de cada sentença do sumário.

A Figura 4.11 mostra as expessões sintáticas do tipo *Productions* utilizadas em cada sentença de um sumário automático multidocumento obtidas pelo Gerador de Produções Sintáticas. Na Figura 4.11, cada expressão sintática (oriunda do *parser* PALAVRAS) que compõem as sentenças do sumário está relacionada a alguma palavra de uma sentença. As disposições das expressões mostrada na Figura 4.11 não estão ordenadas.

No modelo original, as autoras utilizaram, tanto para *Productions* quanto para *d-sequence*, as 25 expressões sintáticas mais frequentes do córpus. Tal restrição foi seguida na implemen-

(S1) ['S -> DN H ', 'H -> prp', 'DP -> DN H ', 'H -> prop', 'PU -> pu', 'DN -> art', 'DA -> adv', 'P -> v-fin', 'DP -> n', 'H -> n', 'STA -> S P fA fA PU ', 'DN -> DA H ', 'DP -> DN H DN ', 'fA -> H DP ', 'DN -> H DP ', 'H -> num']

(S2) ['STA -> fA PU S PU P Od PU ', 'H -> prop', 'PU -> pu', 'Vm -> v-ger', 'DN -> art', 'P -> v-inf', 'DP -> DN H ', 'S -> DN H PU DNc ', 'DP -> H DN ', 'H -> prp', 'DP -> prop', 'P -> Vaux Vm ', 'DNc -> H DP ', 'Vaux -> v-fin', 'DP -> DN H DN ', 'DN -> num', 'fA -> H DP ', 'DN -> H DP ', 'DN -> adj', 'Od -> P fA fA ', 'H -> n']

(S3) ['DN -> H DP ', 'PU -> pu', 'P -> v-fin', 'fA -> fA S fA PU fA P Od P Op Op ', 'As -> H DP ', 'STA -> S P fA PU ', 'DP -> DN DN H ', 'DP -> n', 'H -> prp', 'DNc -> v-pcp', 'DN -> art', 'Od -> P As ', 'DP -> prop', 'DP -> DN H DN ', 'Op -> H DP ', 'S -> DN H PU DNc ', 'P -> v-inf', 'DN -> num', 'S -> pron-indef', 'DP -> DN H ', 'H -> n', 'fA -> adv', 'fA -> H DP ', 'DN -> adj']

Figura 4.11: Expressões Sintáticas

tação dessa adaptação e, adicionalmente, experimentos com todas as expressões obtidas e com as 400 expressões sintáticas mais frequentes do córpus CSTNews foram realizados. Com isso, o intuito era verificar qual a quantidade ideal de expressões mais frequentes (25 ou 400) ou se todas as expressões do córpus seriam a melhor solução para uma boa performance do modelo.

Com as expressões obtidas para cada sentença do texto, o respectivo valor de coerência foi calculado por meio da utilização da Equação 3.11 replicada em 4.3. Nessa equação, a probabilidade condicional $p(S_i^j|S_{i-1}^k)$ é calculada por meio da Equação 3.12 replicada em 4.4, a qual possui um termo de suavização (δ_c). Como as autoras não deixaram claro o valor utilizado para essa suavização, três valores foram utilizados: 0,1; 0,01 e 0,001. Tais valores foram analisados, pois havia a dúvida de qual deles seria o melhor valor que impactaria menos no resultado do modelo. A suavização, na Equação 4.4, tem a única função de assegurar que não haverá divisão por 0 (zero).

$$P(T) = \prod_{i=2}^{n} \prod_{j=1}^{|S_i|} \frac{1}{|S_{i-1}|} \sum_{k=1}^{|S_{i-1}|} p(S_i^j|S_{i-1}^k) \tag{4.3}$$

$$p(w_j|w_i) = \frac{c(w_i,w_j) + \delta_c}{c(w_i) + \delta_c * |V|} \tag{4.4}$$

Os resultados dos experimentos desse modelo (listados na Seção 4.5) foram obtidos levando em consideração todos os aspectos anteriormente descritos.

4.5 Experimentos e Resultados

Os experimentos realizados nesta tese tiveram o propósito de avaliar a qualidade dos modelos aplicados em sumários multidocumento, mas também relatar os experimentos iniciais de alguns modelos adaptados que também foram aplicados em textos fonte do córpus CSTNews.

A tarefa de ordenação de sentenças de Barzilay & Lapata (2008) vem sendo usada e, nesta tese, não seria diferente, como o método para avaliar os modelos de coerência local. Essa tarefa se tornou o método para verificar e avaliar a performance dos modelos de coerência local de todos os trabalhos relacionados.

Como dito em passagens anteriores, a tarefa de ordenação de sentenças consiste em avaliar

pares de textos, no caso desta tese, avaliar pares de sumários também, ou seja, {Texto/Sumário original, Versão permutada do Texto/Sumário original}. A acurácia é medida por meio da razão entre o número de pares corretos e o total de pares de textos/sumários. O número de pares corretos é contabilizado quando o modelo em avaliação gera um valor de ranque ou de coerência maior para o texto/sumário original em relação a sua versão permutada em cada par.

Para avaliar os modelos, descritos nesse capítulo, os sumários humanos multidocumento do córpus CSTNews foram utilizados como sumários de referência, ou seja, tais sumários foram considerados coerentes. Tal proposição foi baseada na qualidade verificada dos sumários produzidos.

Com o acréscimo de 5 sumários humanos extrativos, o córpus CSTNews possui 6 sumários de referência para cada coleção, totalizando 300 sumários. Desse total, 251 sumários foram utilizados na tarefa de ordenação de sentenças. O motivo de não se utilizarem os outros 49 sumários foi por possuírem um número de sentenças igual ou menor do que 3. Essa restrição é por causa da necessidade das 20 permutações aleatórias para cada sumário original. Da mesma forma que a tarefa de ordenação de sentença virou um método padrão na área para avaliar os modelos de coerência, a quantidade de 20 permutações também se tornou um padrão na área. Portanto, a base de dados foi composta por 5.020 pares de sumários multidocumento.

Os modelos LSA de Foltz et al. (1998) e Grade de Entidades de Barzilay & Lapata (2008) foram um dos primeiros implementados e, até esse momento, não se tinha o córpus de sumários consolidado. Além disso, havia o desejo de verificar o comportamento desses modelos em textos do português brasileiro. Para isso, os textos fonte do córpus CSTNews foram usados. Desta forma, 137 dos 140 textos do CSTNews e 20 permutações aleatórias para cada um dos textos foram empregados. A quantidade total de textos do córpus não foi utilizada por causa da quantidade de sentenças que eles possuem, inviabilizando as 20 permutações. Assim, a base de dados foi composta por 2.740 pares de textos.

O modelo LSA gera um valor de coerência para cada texto fonte, sumário original e suas permutações. A Tabela 4.1 mostra a acurácia obtida pelo modelo LSA tanto na base de textos fonte quanto na base de sumários. Os resultados da Tabela 4.1 demonstram que o método baseado na similaridade entre sentenças não teve resultados expressivos. Isso pode ser explicado pela falta de informação linguística que aprimore a avaliação do modelo.

Tabela 4.1: Resultado do modelo LSA

Base	**Acurácia (%)**
Textos fonte	58,40
Sumários	55,18

Outro ponto, já esperado, é o fato do modelo ter sido melhor nos textos fonte do que nos sumários. Isso ocorre porque, os sumários extrativos multidocumento são formados por sentenças de textos diferentes, o que pode ocasionar uma baixa similaridade entre as sentenças dos sumários, prejudicando, assim, a acurácia do modelo.

Diferente do modelo original, o qual distribui os pares de textos de forma quase igualitária entre treinamento e teste, o método de validação cruzada de 10 *folds* para avaliar o modelo de Grade de Entidades foi empregado, pois se acredita que, com essa abordagem de avaliação, resultados mais confiáveis poderão ser produzidos.

Como nesta tese não foi considerada a informação de correferência para o modelo de Barzilay & Lapata (2008), 4 versões do modelo de Grade de Entidades foram avaliadas: (Sintático+Saliência+), (Sintático+Saliência-), (Sintático-Saliência+) e (Sintático-Saliência-).

Utilizando tanto a base de dados de textos fonte quanto a base de dados de sumários, todas as versões do modelo de Grade de Entidades foram avaliadas. A acurácia obtida por cada versão nas respectivas bases é mostrada na Tabela 4.2.

Tabela 4.2: Resultado do Modelo de Grade de Entidades

Modelos	**Acurácia (%)**	
	Textos Fontes	**Sumários**
Sintático+Saliência+	70,73	64,78
Sintático+Saliência-	74,10	60,21
Sintático-Saliência+	67,87	61,99
Sintático-Saliência-	**78,97**	**68,40**

Segundo a Tabela 4.2, todas a versões aplicadas nos textos fonte obtiveram melhores resultados em textos fonte do que em sumários. Normalmente, as grades de textos fonte são mais preenchidas do que nos sumários, pois as mesmas entidades são mais frequentes e melhor distribuídas ao longo do texto. Tais fatos influenciam diretamente na captura do padrão de transição de entidades presentes nos textos e nos sumários.

Outro fato interessante é que a versão completa (Sintático+Saliência+) não foi a versão que obteve os maiores valores de acurácia tanto para textos fonte quanto para os sumários. Acredita-se que a presença da informação de saliência para os textos fonte foi o fator determinante para a queda da acurácia em comparação com as versões que não fazem uso dessa informação. Uma vez que, tal informação de saliência produz uma grade somente com entidades de frequência é igual ou maior do que 2, e isso pode ocasionar uma grade com poucas entidades e com bastante lacunas (devido a natureza multidocumento dos sumários), o que pode prejudicar o aprendizado do modelo. Além disso, a informação da presença de entidades nas sentenças foi a que melhor se adequou tanto nos textos fonte quanto nos sumários. Isso se deve pelas respectivas grades com poucas lacunas e menos esparsas, o que ajudou no aprendizado do modelo.

Diferentemente dos modelos anteriormente testados, o modelo Baseado em Grafo (Guinaudeau & Strube, 2013) foi utilizado apenas em sumários multidocumento (foco desta tese). Assim, os valores de acurácia foram obtidos para cada uma das projeções *one mode* (P_U, P_W e P_{Acc}) juntamente com e sem a informação de distância. A Tabela 4.3 mostra os resultados do modelo de Grafo aplicado ao córpus de sumários.

O modelo Baseado em Grafo aplicado somente a sumários multidocumento, já que a fase de experimentos em textos fonte havia finalizada quando este modelo foi desenvolvido, teve resul-

Tabela 4.3: Resultados do Modelo baseado em Grafo

Modelos de Grafo	**Acurácia (%)**	
	Com Inf. Distância	**Sem Inf. Distância**
Projeção P_U	**52,71**	**57,69**
Projeção P_W	51,21	54,98
Projeção P_{Acc}	52,55	56,51

tados inferiores aos obtidos pelo modelo de Grade de Entidades. Mesmo o melhor resultado do modelo Baseado em Grafo ficou muito abaixo do valor de menor acurácia do modelo de Grade de Entidades.

Os resultados mostrados na Tabela 4.3 não eram o esperado, já que a motivação da criação do modelo de Grafo, segundo os autores, era a obtenção de acurácias similares ao modelo de grade de entidades mesmo sem a fase complexa (aprendizado de máquina) do modelo de Grade de Entidades. Tal comportamento pode se dever ao fato de que muitos sumários podem ter tido entidades diferentes e espalhadas em diferentes sentenças ao longo do sumário, pois os valores de coerência das projeções são baseados na ocorrência,no somatório das ocorrências e dos pesos das entidades em comuns a duas sentenças. Assim, quanto mais entidades as sentenças compartilharem melhores serão os valores de coerência, principalmente para os sumários de referência (coerentes).

E, por fim, entre os modelos adaptados e que originalmente não usaram informação discursiva, o modelo baseado em Padrões Sintáticos de Louis & Nenkova (2012b) foi avaliado. Para isso, 3 (três) conjuntos de expressões sintáticas, segundo as suas frequências, foram utilizados, ou seja, todas as expressões que tiveram frequência iguais a 1, 25 ou 400. Além disso, 3 (três) valores de suavização (0,1; 0,01; 0,001) também foram usados. Todos esses aspectos foram considerados para *Productions* e *d-sequence* com $d = 2$, ou seja, todas as expressões sintáticas dos níveis 1 e 2 da árvore sintática de cada sentença do sumário foram empregadas.

A Tabela 4.4 mostra todos os resultados obtidos com a aplicação do modelo de Padrões Sintáticos com a utilização de expressões sintáticas do tipo *Productions* para o córpus de sumários. Já a Tabela 4.5 mostra os resultados do modelo de Padrões Sintáticos com o uso de expressões sintáticos do tipo *d-sequence*.

Tabela 4.4: Resultados do modelo de Padrões Sintáticos para *Productions*

Suavização	**Acurácia (%)**		
	Frequência = 1	**Frequência = 25**	**Frequência = 400**
0,1	15,35	**15,67**	**17,43**
0,01	16,25	15,57	17,17
0,001	**19,68**	15,57	17,41

Em geral, esse modelo de Padrões Sintáticos não é o modelo mais recomendável para avaliar coerência local em sumários multidocumento, pois o melhor resultado foi de 26,19% de

Tabela 4.5: Resultados do modelo de Padrões Sintáticos para *d-sequence*

Suavização	**Acurácia (%)** *d-sequence* **= 1**		
	Frequência = 1	**Frequência = 25**	**Frequência = 400**
0,1	15,87	**16,85**	**17,43**
0,01	**20,76**	16,81	17,41
0,001	19,48	16,79	17,33
Suavização	**Acurácia (%)** *d-sequence* **= 2**		
	Frequência = 1	**Frequência = 25**	**Frequência = 400**
0,1	21,47	**22,49**	23,42
0,01	**26,19**	22,33	**23,50**
0,001	24,78	22,33	23,44

acurácia para *d-sequence* no nível 2, Frequência = 1 (todas as expressões sintáticas do nível utilizada) e suavização de 0,01.

Em média, conclui-se que o valor de suavização ideal seria de 0,01, mesmo as maiores acurácias sendo obtidas com a suavização de 0,1.

A utilização de todas as expressões sintáticas, independentes da frequência de cada uma, foi a configuração em que os modelos Baseados em Padrões Sintáticos classificaram melhor os sumários em comparação com as configurações que agruparam as expressões sintáticas de acordo com as suas frequências.

O tamanho do sumário pode ter influenciado na baixa acurácia do modelo Baseado em Padrões Sintáticos. Um bom indício disso foram as acurácias obtidas com a utilização de todas as expressões sintáticas (Frequência = 1), que foram maiores para cada tipo. Assim, quanto mais sentenças houver, consequentemente, mais expressões sintáticas permitirão ao modelo avaliar melhor a coerência local dos sumários multidocumento.

Pelos resultados de cada um dos modelos adaptados, os quais não utilizam informação de discurso, conclui-se que o modelo de Grade de Entidades foi o melhor modelo para sumários multidocumento. Além disso, tais resultados mostram a necessidade de melhorar a avaliação da coerência local em sumários multidocumento. Devido a isso, acredita-se que o conhecimento discursivo pode ser útil em modelos que avaliam a coerência local. Baseado nisso, modelos com informações discursivas foram elaborados, adaptados e criados, o quais serão descritos no Capítulo 5, comprovando a utilidade dessas informações para avaliação da coerência.

Capítulo

5

Enriquecimento de Métodos de Coerência

Nesse capítulo os modelos que dão suporte a tese deste trabalho serão descritos e avaliados. Assim, os modelos da literatura que já utilizam conhecimento discursivo foram adaptados, os modelos que não utilizam informação discursiva foram enriquecidos/incrementados com tal informação e novos modelos discursivos que avaliam a coerência local foram criados.

As relações discursivas da RST (Mann & Thompson, 1987) e da CST (Radev, 2000) anotadas no córpus CSTNews (Aleixo & Pardo, 2008; Cardoso et al., 2011) e o parser PALAVRAS (Bick, 2000) foram utilizadas na construção dos modelos. Além disso, os modelos discursivos, em sua maioria, foram aplicados em sumários multidocumento e alguns modelos foram empregados em textos fonte, mas todos foram analisados de forma que a eficiência dos mesmos pudesse ser verificada.

5.1 Modelo de Grade de Entidades com Discurso

Considerado o modelo de referência na área de avaliação da coerência local, o modelo de Grade de Entidades foi o que mais possibilitou a criação de variações com discurso, de forma que tais variações pudessem ser aplicadas em textos fonte e/ou em sumários multidocumento.

Todos os modelos discursivos baseados no modelo de Grade de Entidades ou até as suas versões enriquecidas com informação discursiva tiveram suas estruturas similares ao modelo de Barzilay & Lapata (2008) implementado nesta tese (veja a Figura 4.1). As diferenças ficaram no acréscimo de funcionalidades no módulo Gerador de Grade de Entidades e no módulo Gerador de Vetor de Características Discursivas de cada modelo e versão.

Além de extrair as entidades com suas funções sintáticas do arquivo de análise do PALAVRAS, há modelos e/ou versões em que o Gerador de Grade de Entidades também irá extrair informações discursivas anotadas nos textos fonte ou sumários multidocumento. Assim, tal

módulo poderá criar grades de entidades, grade discursivas e grade de entidades com discurso. Da mesma forma, o Gerador de Vetor de Características Discursivas poderá montar vetores de características considerando uma ou até duas grades de informações distintas.

A Figura 5.1 ilustra a estrutura das versões enriquecidas com discurso e dos modelos de discurso baseados no modelo de Grade de Entidades de Barzilay & Lapata.

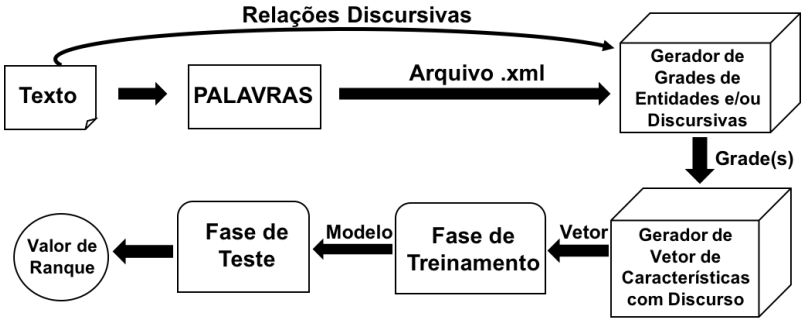

Figura 5.1: Estrutura dos Modelos de Grade de Entidades enriquecidas com discurso

A manipulação e o tipo das relações discursivas foram os aspectos que diferenciaram cada versão enriquecida, bem como os modelos discursivos baseados no modelo de Grade de Entidades.

O primeiro modelo discursivo baseado no modelo de Grade de Entidades foi direcionado para textos fonte (etapa inicial da pesquisa), tendo em vista o objetivo de investigar a sua performance em textos maiores e compará-lo com o modelo adaptado de Grade de Entidades original.

Baseado na premissa de que a partir de um texto coerente é sempre possível obter uma estrutura de relações RST e que um modelo de Grade de Entidades é capaz de capturar um padrão de transições de relações RST entre as sentenças adjacentes dos textos fonte de referência, foi implementado o modelo de Grade de Entidades denominado SINTÁTICA-SALIÊNCIA-RST+.

O modelo SINTÁTICA-SALIÊNCIA-RST+ faz uso apenas das relações RST anotadas nos textos fonte para montar a grade de entidades com informação discursiva. Dessa forma, o Gerador de Grade de Entidades e/ou Discursivas irá extrair as entidades e utilizar as relações RST anotadas nos textos para formar a grade de entidades com discurso. Para exemplificar a grade montada por esse módulo, considere as duas primeiras sentenças do texto segmentado em EDUs da Figura 2.5, como mostra a Figura 5.2.

(S1) **[1]** Muitas das atitudes “corajosas” de Almir, o Pernambuquinho, eram ditadas pelo medo.

(S2) **[2]** Poucos sabem disso, **[3]** mas é verdade.

Figura 5.2: Parte do texto da Figura 2.5

O texto da Figura 5.2 possui 3 EDUs distintas referenciadas por [1], [2] e [3]. As informações nas EDUs [2] e [3] complementam a informação contida na EDU [1], ou seja, as EDUs [2] e [3] são identificadas como constituintes de uma relação RST chamada ELABORATION da EDU [1]. Dessa forma, a EDU [1] corresponde ao núcleo e as EDUs [2] e [3] constituem o

satélite da relação ELABORATION. As EDUs [2] e [3] se relacionam por meio de uma concessão. Portanto, a EDU [2] é o núcleo e a EDU [3] é o satélite da relação CONCESSION. Tais relacionamentos são vistos na Figura 2.6 replicada em 5.3. Já a Figura 5.4 ilustra a grade de entidades com discurso para o texto da Figura 5.2.

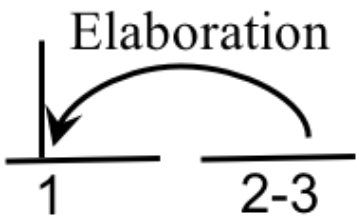

Figura 5.3: Relação ELABORATION entre as proposições 1 e 2-3 (Ribeiro & Rino, 2005, p. 2)

	atitudes	Almir	Pernambuquinho	medo	verdade
S1	Elab.Nuc	Elab.Nuc	Elab.Nuc	Elab.Nuc	-
S2	-	-	-	-	Conces.Nuc Elab.Sat

Figura 5.4: Grade de relação RST para o texto da Figura 5.2

É importante salientar que todas as relações das quais uma determina entidade faça parte de um dos contituintes são registradas na grade. Por exemplo, na grade da Figura 5.4, a entidade "verdade" ocorre em EDUs da sentença S2, que participam como núcleo da relação CONCESSION e como satélite da relação ELABORATION.

O módulo Gerador de Vetor de Características Discursivas utiliza a grade de entidades com discurso para contabilizar as probabilidades das transições de relações RST entre as sentenças adjacentes. Para isso, todas as relações RST ocorridas no córpus juntamente com a informação de nuclearidade foram consideradas. Por exemplo, a transição **[Elab.Nuc -]** na grade da Figura 5.4 tem a probalidade de 0,66, ou seja, esse valor foi obtido por meio da razão entre as 4 ocorrências da transição **[Elab.Nuc -]** e as 6 transições entre sentenças adjacentes possíveis da grade da Figura 5.4. O restante do modelo SINTÁTICA-SALIÊNCIA-RST+ segue os mesmos passos do modelo original.

Duas variações do modelo SINTÁTICA-SALIÊNCIA-RST+ também foram implementadas. Essas variações impactam na manipulação das relações RST no módulo Gerador de Grade de Entidades e/ou Discursivas e no módulo Gerador de Vetor de Características Discursivas. O objetivo dessas variações foi diminuir a complexidade da grade de entidade e do vetor de características da versão SINTÁTICA-SALIÊNCIA-RST+ e verificar se haveria algum ganho na acurácia com essas variações.

Para a Variação 1 foram utilizados os agrupamentos das relações RST feitos por Mann & Thompson (1987) (ver a Tabela 2.3) e, além disso, as informações de nuclearidade das relações RST foram ignoradas. Já na Variação 2, as relações RST foram utilizadas sem agrupá-las e

sem a informação de nuclearidade. A Figura 5.5 mostra a versão da grade da Figura 5.4 que pode ser utilizada nas duas variações, já que tanto a relação ELABORATION quanto a relação CONCESSION são consideradas um tipo de grupo na Variação 1 e também são utilizadas na Variação 2.

	atitudes	Almir	Pernambuquinho	medo	verdade
S1	Elaboration	Elaboration	Elaboration	Elaboration	-
S2	-	-	-	-	Concession Elaboration

Figura 5.5: Exemplo de grade de relação RST para as Variações 1 e 2

O modelo SINTÁTICA-SALIÊNCIA-RST+ é similar ao modelo de RST Completo de Feng et al. (2014) empregado para o Inglês. O modelo desenvolvido nesta tese foi construído concomitantemente ao modelo de RST Completo.

As versões do modelo de Grade de Entidades com informação sintática e sem saliência (SINTÁTICA+SALIÊNCIA-) enriquecidas com informações discursivas (RST e CST) são baseadas em padrões sintáticos e discursivos que os sumários mutidocumento coerentes possuem e que os diferem dos sumários considerados incoerentes. Uma dessas versões enriquecidas usa as relações CST dos sumários juntamente com a função sintática que cada entidade possui para avaliar a coerência dos sumários multidocumento. Essa versão é denominada de SINTÁTICA+SALIÊNCIA- com CST. O uso de relações CST é devido a própria natureza dos sumários que são multidocumento, o que possibilita a captura de um padrão de distribuição de relações CST para a distinção de sumários coerentes dos incoerentes. O módulo Gerador de Grade de Entidades e/ou Discursivas dessa versão gera duas grades, uma com informação sintática e outra com informação discursiva. Além disso, o Gerador de Vetor de Características com Discurso lida com as duas grades conjuntamente para formar o vetor de características.

Para exemplificar a versão SINTÁTICA+SALIÊNCIA- com CST, considere o sumário do córpus CSTNews mostrado na Figura 5.6.

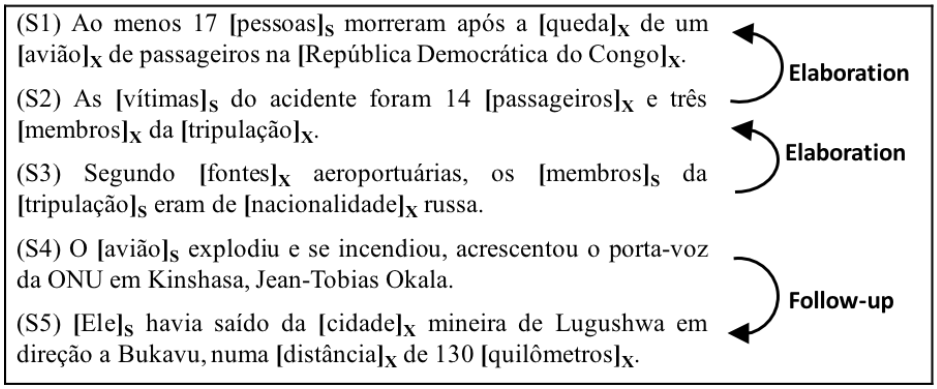

Figura 5.6: Exemplo de um sumário com relações CST

No exemplo da Figura 5.6, ocorrem 3 relações CST. A relação ELABORATION ocorre entre as sentenças S1 e S2, ou seja, a informação dada pela sentença S2 complementa o fato

principal descrito na sentença S1. O mesmo acontece entre as sentenças S2 e S3. No caso das sentenças S4 e S5, há a relação FOLLOW-UP, a qual indica que o fato descrito na sentença S4 aconteceu depois do fato descrito na sentença S5.

O sumário da Figura 5.6 é analisado pelo *parser* PALAVRAS, que gera o arquivo xml com as informações morfossintáticas necessárias para o modelo. Dessa forma, o Gerador de Grades de Entidades e/ou Discursivas utiliza as informações sintáticas e as relações CST presentes no sumário para gerar duas grades, uma grade sintática e outra discursiva, como mostra a Figura 5.7 (a) e (b), respectivamente.

	pessoas	queda	avião	R.D.Congo	vítimas	passageiros	membros	tripulação	fontes	nacionalidade	ele	cidade	distância	kilometros
S_1	S	X	X	X	–	–	–	–	–	–	–	–	–	–
S_2	–	–	–	–	S	X	X	X	–	–	–	–	–	–
S_3	–	–	–	–	–	–	S	S	X	X	–	–	–	–
S_4	–	–	S	–	–	–	–	–	–	–	–	–	–	–
S_5	–	–	–	–	–	–	–	–	–	–	S	X	X	X

(a)

	S1	S2	S3	S4	S5
S1	–	Elaboration	–	–	–
S2	–	–	Elaboration	–	–
S3	–	–	–	–	–
S4	–	–	–	–	Follow-up
S5	–	–	–	–	–

(b)

Figura 5.7: Grades (a) sintática e (b) discursiva de relações CST

A grade discursiva é representada por uma matriz composta por linhas e colunas que representam as mesmas sentenças do sumário. O preenchimento das células dessa matriz são com as relações CST entre duas sentenças do sumário, independentemente da quantidade de relações CST que exista entre duas sentenças.

Com as duas grades da Figura 5.7, o Gerador de Vetor de Características com Discurso condensa ambas as informações para criar o vetor de características para essa versão de Grade de Entidades com CST. Por exemplo, a relação ELABORATION entre as sentenças S1 e S2 co-ocorre com as transições sintáticas de entidades no mesmo par de sentenças: **[S -]**, **[X -]**, **[- S]**, **[- X]**, **[- -]** . Nesse caso, para cada célula da grade de entidades, o Gerador de Vetor de Características com Discurso contabiliza a frequência de cada transição sintática que ocorre junto com a relação CST presente no correspondente par de sentenças. Esses valores são divididos pelo número total de transições de tamanho 2 da grade de entidades, calculando, assim, a probabilidade para cada transição de entidades com as relações CST. A Figura 5.8 mostra o vetor de características relacionado às grades da Figura 5.7.

S – Elaboration	X – Elaboration	– S Elaboration	– X Elaboration	– – Elaboration	X S Elaboration	S – Follow-up	X – Follow-up	– S Follow-up	– X Follow-up	– – Follow-up	X S Follow-up	S – –	X – –	– S –	– X –	– – –	X S –
0,03	0,07	0,01	0,08	0,25	0,03	0,01	0	0,01	0,05	0,16	0	0,03	0,03	0,01	0	0,16	0

Figura 5.8: Vetor de característica da versão Grade de Entidades com CST

No vetor de características da Figura 5.8, os valores de probabilidade são obtidos por meio da razão entre a frequência de cada padrão e o número total de transições da grade de informação

sintática. Por exemplo, o padrão **[S-Elaboration]** tem o valor de probabilidade igual a 0,03. Pois, o padrão **[S-Elaboration]** ocorre 2 vezes e o número total de transições da grade de entidade sintática é 56.

Um vetor de características é criado para cada sumário de treinamento e de teste, os quais serão usados para treinar e testar, respectivamente, o modelo preditivo na avaliação da coerência local (ver Seção 5.6).

Outra versão enriquecida do modelo de Grade de Entidades é relacionada à versão SINTÁTICA+SALIÊNCIA- com CST, já que essa última pode gerar um número grande de características e, com isso, pode haver dados esparsos e, consequentemente, pode diminuir a performance da versão. Pensando nisso, a versão intitulada SINTÁTICA+SALIÊNCIA- com Categorias CST foi desenvolvida, ou seja, 5 categorias das relações CST dadas pela tipologia de Maziero et al. (2010) (ver Figura 2.10) foram utilizadas. Assim, as categorias são os tipos: Redundância, Complemento, Contradição, Fonte/Autoria e Estilo.

A modelagem da versão SINTÁTICA+SALIÊNCIA- com CST teve o mesmo princípio dos modelos de coerência discursivos anteriores, ou seja, um padrão de distribuição das categorias de relações CST ao longo do sumário pode ser utilizado na distinção de sumários coerentes dos menos coerentes (incoerentes).

O Gerador de Grade de Entidades e/ou Discursivas da versão SINTÁTICA+SALIÊNCIA- com Categorias CST reconhece a relação CST do sumário e preenche a grade de discurso com o tipo da relação, reduzindo a dimensionalidade do vetor de características. Por exemplo, a Figura 5.9 mostra a grade de discurso da versão SINTÁTICA+SALIÊNCIA- com Categorias CST da grade discursiva (b) da Figura 5.7.

	S1	S2	S3	S4	S5
S1	–	Complemento	–	–	–
S2	–	–	Complemento	–	–
S3	–	–	–	–	–
S4	–	–	–	–	Complemento
S5	–	–	–	–	–

Figura 5.9: Grade discursiva de categoria CST

A grade discursiva de categoria CST é usada juntamente com a grade de entidades com informação sintática para formar os padrões de transições de informação sintática com as categorias CST entre as sentenças dos sumários, de forma similar a versão SINTÁTICA+SALIÊNCIA- com CST. Com isso, o Gerador de Vetor de Características com Discurso da versão SINTÁTICA+SALIÊNCIA- com Categorias CST irá criar vetores de características menores e, consequentemente, diminuir a sua complexidade em comparação ao gerador da versão SINTÁTICA+SALIÊNCIA- com CST. A Figura 5.10 mostra o vetor de caraterísticas da versão da SINTÁTICA+SALIÊNCIA- com Categorias CST, relacionado à grade sintática da Figura 5.7 e à grade discursiva da Figura 5.9.

Todos os passos realizados até gerar o valor de ranque para os sumários de teste foram

S – Complemento	X – Complemento	– S Complemento	– X Complemento	– – Complemento	X S Complemento	S – –	X – –	– S –	– X –	– – –	X S –
0,03	0,07	0,01	0,09	0,25	0,03	0,03	0,03	0,01	0	0,16	0

Figura 5.10: Vetor de característica da versão Grade de Entidades com Categoria CST

realizados da mesma forma que no modelo de Grade de Entidades.

Uma nova versão com uma quantidade ainda menor de características em comparação com a versão de SINTÁTICA+SALIÊNCIA- com Categoria CST foi implementada. Com o mesmo foco de reduzir dados esparsos e, consequentemente, possibilitar o aumento do poder preditivo na avaliação da coerência local, a versão do modelo de Grade de Entidades com informação booleana de discurso (CST) foi desenvolvida. Essa informação boolena de discurso CST consiste na presença (valor=1) ou na ausência (valor=0) de relações CST entre as sentenças dos sumários. Essa versão será referenciada como SINTÁTICA+SALIÊNCIA- Booleana CST e é baseada em um padrão de distribuição das funções sintáticas das entidades juntamente com a sinalização da presença ou da ausência de relações CST entre as sentenças. Tal padrão é utilizado para distinguir sumários coerentes dos incoerentes.

O uso de informação Booleana só foi empregado nas relações CST, já que a possibilidade de obter um padrão Booleano de relações CST é maior do que um possível padrão Booleano só de relações RST em sumários multidocumento. Uma vez que, a quantidade de relações RST em sumários multidocumento não seria suficiente para obter um padrão de distribuição de relações RST, pois as sentenças que formam os sumários multidocumento, em sua maioria, vem de diferentes textos fonte.

O Gerador de Grades de Entidades e/ou Discursivas da versão SINTÁTICA+SALIÊNCIA-Booleana CST produzirá duas grades, uma grade de entidades com informação sintática, como a que é mostrada na Figura 5.7 (a), e uma grade com valores booleanos que indicam se duas sentenças estão relacionadas por relações CST. A Figura 5.11 mostra a versão boolena da grade discursiva de relações CST da Figura 5.7 (b).

	S1	S2	S3	S4	S5
S1	0	1	0	0	0
S2	0	0	1	0	0
S3	0	0	0	0	0
S4	0	0	0	0	1
S5	0	0	0	0	0

Figura 5.11: Grade booleana CST

Nesse cenário, a Figura 5.12 mostra o vetor de características gerado pelo Gerador de Vetor

de Características com Discurso da versão de SINTÁTICA+SALIÊNCIA- Booleana CST, lembrando que cada característica é a probabilidade de cada padrão de transição (linha 1 da Figura 5.12) e o cálculo da probabilidade é o mesmo realizado nas versões anteriores.

S – 1	X – 1	– S 1	– X 1	– – 1	X S 1	S – 0	X – 0	– S 0	– X 0	– – 0	X S 0
0,03	0,07	0,03	0,14	0,41	0,03	0,03	0,03	0,01	0	0,16	0

Figura 5.12: Vetor de característica booleana CST

A principal característica dos sumários multidocumento é que as sentenças que os formam podem vir de diferentes textos fonte, mas também podem vir do mesmo texto. Sentenças que vem do mesmo texto fonte podem ter relações discursivas que não são representadas pelas relações CST e sim pelas relações RST. Assim, foram usadas as possíveis relações RST das sentenças dos sumários para criar novas versões, pois se acredita que sumários coerentes podem ter mais sentenças adjacentes ordenadas vindas do mesmo texto fonte do que sumários considerados incoerentes.

A Figura 5.13 mostra um sumário humano da coleção 2 do córpus CSTNews. Esse sumário, como os outros do córpus, apresenta uma marcação no final de cada sentença que permitiu recuperar as possíveis informações discursivas de cada sentença dos textos fonte (onde foram anotadas as relações RST). Essa marcação identifica a origem de cada sentença dos sumários nos textos fonte. Por exemplo, a sentença S1 desse sumário é a primeira sentença no documento 4, na coleção 2 e cuja fonte é o jornal O Globo (<D4_C2_Globo;S1>).

(S1) O Instituto de pesquisa CNI/Ibope divulgou nesta sexta-feira, em Brasília, que o presidente Luiz Inácio Lula da Silva seria reeleito em primeiro turno caso as eleições fossem hoje. <D4_C2_Globo;S1>
(S2) De acordo com a pesquisa, Lula (PT) tem 44% das intenções de voto, contra 25% de Geraldo Alckmin (PSDB) e 11% de Heloísa Helena (PSOL). <D4_C2_Globo;S2>
(S3) Os outros candidatos que concorrem a vaga presidencial Cristovam Buarque (PDT) e Luciano Bivar (PSL) aparecem cada um com 1% das intenções de voto. <D4_C2_Globo;S3>
(S4) A CNI explica que a pesquisa não traz a comparação com pesquisas anteriores para primeiro turno porque é a primeira vez que o Ibope utiliza a lista oficial de candidatos a presidente da República. <D2_C2_Estadao;S5>
(S5) Embora não permita comparações, vale relembrar que na pesquisa de junho Lula tinha 48% das intenções de voto; Alckmin 18% e Heloísa Helena, 5%. <D2_C2_Estadao;S6>

Figura 5.13: Sumário humano com marcacões de origem das sentenças

Recuperando as possíveis informações discursivas no sumário da Figura 5.13, pode-se formar uma grade com relações CST e RST. Por exemplo, a grade de relações do sumário da Figura 5.13 é mostrada na Figura 5.14. Nessa grade, os nomes de relações com todas as letras minúsculas são relações RST e as relações CST são as que possuem a primeira letra maiúscula.

O Gerador de Vetor de Características com Discurso das versões que fazem uso das duas relações discursivas ao mesmo tempo calcula os padrões formados pelas informações sintáticas e todas as relações CST e RST anotadas no CSTNews. Essa versão foi denominada SINTÁTICA+SALIÊNCIA- com CST e RST. O intuito da incorporação das relações RST na versão SINTÁTICA+SALIÊNCIA- com CST foi de melhorar o poder de distinguir

		(S1)	(S2)	(S3)	(S4)	(S5)
		S1_D4_C2	S2_D4_C2	S3_D4_C2	S5_D2_C2	S6_D2_C2
(S1)	S1_D4_C2	-	evidence	-	Elaboration	-
(S2)	S2_D4_C2	-	-	list	-	-
(S3)	S3_D4_C2	-	-	-	-	-
(S4)	S5_D2_C2	-	-	-	-	-
(S5)	S6_D2_C2	-	-	-	-	-

Figura 5.14: Grade com relações RST e CST

os sumários coerentes dos incoerentes e tentar diminuir a esparsidade presente na versão SINTÁTICA+SALIÊNCIA- com CST.

A partir da versão SINTÁTICA+SALIÊNCIA- com CST e RST, foram criadas novas versões que seguem o mesmo formato de implementação das versões anteriores, mudando apenas a manipulação das informações discursivas no Gerador de Grade de Entidades e/ou Discursivas e no Gerador de Vetor de Características com Discurso. Tais versões são: SINTÁTICA+SALIÊNCIA- com RST, a qual utiliza a grande de entidades com informação sintática junto com a grade discursiva de possíveis relações RST; SINTÁTICA+SALIÊNCIA- Booleana RST, a qual utiliza a grande de entidades com informação sintática juntamente com a grade discursiva preenchida por 1 (presença de relação RST) e 0 (ausência de relação RST); e SINTÁTICA+SALIÊNCIA- Booleana CST e RST, a qual utiliza a grande de entidades com informação sintática juntamente com a grade discursiva preenchida por 1 (presença de relação CST e/ou RST) e 0 (ausência de relação CST e RST).

Versões discursivas para o modelo SINTÁTICA-SALIÊNCIA- da abordagem de Grade de Entidades também foram criadas, já que o modelo SINTÁTICA-SALIÊNCIA- (dentre os modelos de Grade de Entidades adaptados nesta tese) foi o que obteve o melhor desempenho na avaliação da coerência local dos textos fonte e sumários multidocumento do córpus CSTNews.

Os mesmos procedimentos realizados nas versões discursivas com informação sintática (SINTÁTICA+SALIÊNCIA-) também foram feitos na implementação das versões SINTÁTICA-SALIÊNCIA- com CST, SINTÁTICA-SALIÊNCIA- com RST e SINTÁTICA-SALIÊNCIA- com CST e RST.

O Gerador de Grades de Entidades e/ou Discursivas para as versões discursivas do modelo SINTÁTICA-SALIÊNCIA- produzirá tanto uma grade de entidades sem informação sintática, ou seja, presença (1) ou ausência (0) de uma entidade em uma determinada sentença do sumário (ver Figura 5.15, onde as entidades compostas por duas palavras são dadas pelo *parser* PALAVRAS), quanto uma grade de relações discursivas como as grades mostradas nas Figuras 5.7 (b) e 5.14, respeitando a informação discursiva considerada na versão do modelo SINTÁTICA-SALIÊNCIA-.

Todas as versões desenvolvidas, modelos e variações, foram avaliadas no córpus CSTnews. Assim, os experimentos e os seus resultados contabilizados serão mostrados mais adiante na Seção 5.6 desse Capítulo 5.

	pessoas	queda	avião	República_Democrática_Congo	vitimas	passageiros	membros	tripulação	fontes	nacionalidade	ele	cidade	distância	kilometros
S_1	S	X	X	X	–	–	–	–	–	–	–	–	–	–
S_2	–	–	–	–	S	X	X	X	–	–	–	–	–	–
S_3	–	–	–	–	–	–	S	S	X	X	–	–	–	–
S_4	–	–	S	–	–	–	–	–	–	–	–	–	–	–
S_5	–	–	–	–	–	–	–	–	–	–	S	X	X	X

Figura 5.15: Exemplo de grade de entidade sem informação sintática da grade da Figura 5.7 (a)

5.2 Modelo Baseado em Grafo com Discurso

O modelo baseado em Grafo foi enriquecido com relações CST e RST. Para isso, a estrutura mostrada na Figura 5.16 foi utilizada. Essa estrutura é similar à estrutura do modelo original adaptada nesta tese, mostrada na Figura 4.7.

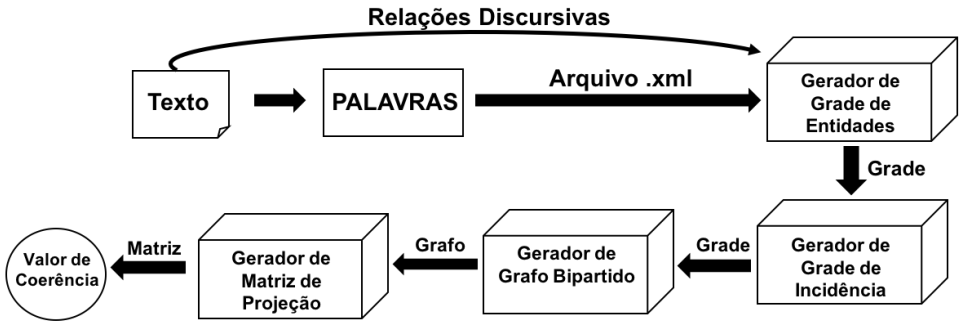

Figura 5.16: Estrutura do Modelo baseado em Grafo com Discursivo

O Gerador de Grade de Entidades utiliza as informações discursivas anotadas nos sumários do córpus CSTNews e o arquivo xml de análise morfossintática dos sumários, dada pelo *parser* PALAVRAS, para criar a grade de entidades com relações discursivas.

A grade de entidades com relações discursivas é uma matriz formada por linhas, que representam as sentenças, e as colunas representam as entidades do sumário. As células dessa grade são preenchidas com relações CST e/ou RST de cada sentença (composta por entidades), a qual se relaciona discursivamente com outra sentença. Um exemplo de grade montada pelo Gerador de Grade de Entidades desse modelo é mostrado na Figura 5.17.

Na grade da Figura 5.17, a entidade "Heloísa_Helena" ocorre na sentença S2, a qual se relaciona com as sentenças S1 (através da relação CST *evidence*) e S3 (através da relação CST *list*), sendo assim, a respectiva célula é preenchida com as duas relações pelas quais a sentença S2 se relaciona. O mesmo acontece com as entidades "Geraldo_Alkmin" e "intenção".

O Gerador de Grade de Incidência substitui cada relação discursiva da grade de entrada

	Luciano_Bibar	Heloísa_Helena	comparação	Geraldo_Alkmin	presidente	Brasília	candidato	intenção	Cristovan_Buarque	pesquisa	Ibope
S1	-	-	-	-	evidence	evidence	-	-	-	-	evidence
S2	-	evidence, list	-	evidence, list	-	-	-	evidence,list	-	-	-
S3	list	-	-	-	-	-	list	-	list	-	list
S4	-	-	-	-	-	-	-	-	-	-	-
S5	-	concession	concession	-	-	-	-	concession	-	-	-
S6	-	-	-	-	-	-	-	-	-	-	-

Figura 5.17: Parte da grade de entidade com discurso do sumário 4 da coleção 2 do CSTNews

pelo seu peso. O valor de peso igual a 1 (um) para cada relação CST e RST foi escolhido[1]; caso ocorra mais de uma relação em uma célula, somam-se os valores dos pesos de cada relação. Nas células que não há relações, o valor 0 é inserido. A Figura 5.18 mostra a grade de incidência para a grade da Figura 5.17.

	Luciano_Bibar	Heloísa_Helena	comparação	Geraldo_Alkmin	presidente	Brasília	candidato	intenção	Cristovan_Buarque	pesquisa	Ibope
S1	0	0	0	0	1	1	0	0	0	0	1
S2	0	2	0	2	0	0	0	2	0	0	0
S3	1	0	0	0	0	0	1	0	1	0	1
S4	0	0	0	0	0	0	0	0	0	0	0
S5	0	1	1	0	0	0	0	1	0	0	0
S6	0	0	0	0	0	0	0	0	0	0	0

Figura 5.18: Grade de Incidência

Com a grade de incidência, o grafo bipartido é gerado pelo módulo Gerador de Grafo Bipartido. Por exemplo, a Figura 5.19 mostra o grafo bipartido da grade de incidência da Figura 5.18.

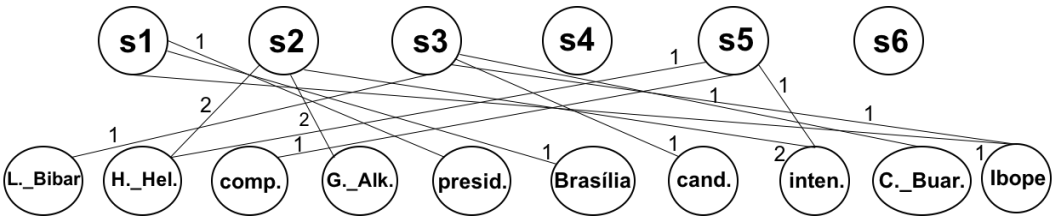

Figura 5.19: Grafo Bipartido Discursivo

O Gerador de Matriz de Projeção utiliza o grafo bipartido discursivo para gerar as matrizes de projeção *one mode* P_U e P_W e, a partir dessas matrizes, os valores de coerência para as

[1]Um estudo futuro sobre a importância de cada relação CST e RST no contexto multidocumento pode ser realizado de forma a quantificar tais relações.

respectivas projeções são calculados. A Figura 5.20 mostra as matrizes de projeções *one mode* P_U (a) e P_W (b) do grafo da Figura 5.19.

	S1	S2	S3	S4	S5	S6
S1	0	0	**1**	0	0	0
S2	0	0	0	0	**1**	0
S3	0	0	0	0	0	0
S4	0	0	0	0	0	0
S5	0	0	0	0	0	0
S6	0	0	0	0	0	0

(a)

	S1	S2	S3	S4	S5	S6
S1	0	0	**1**	0	0	0
S2	0	0	0	0	**2**	0
S3	0	0	0	0	0	0
S4	0	0	0	0	0	0
S5	0	0	0	0	0	0
S6	0	0	0	0	0	0

(b)

Figura 5.20: Matrizes de projeções *one mode* P_U (a) e P_W (b)

Como não houve uma diferenciação nos valores de pesos das relações discursivas, a versão *one mode* P_{Acc} do modelo baseado em grafo não foi enriquecida com informação de discurso, tendo em vista que os valores de coerência da versão *one mode* P_{Acc} seriam muito próximos das outras versões de projeção *one mode*. Os resultados desse modelo, aplicado em sumários multidocumento, serão descritos na Seção 5.6.

5.3 Modelo de Termo com RST

O modelo de Termo com RST é a adaptação do trabalho de Lin et al. (2011), o qual originalmente utiliza relações do Penn Discourse Treebank (PDTB) para criar um modelo de avaliação da coerência local. Esse trabalho foi o primeiro a utilizar conhecimento discursivo para tal fim.

Duas versões do modelo de Lin et al. (2011) foram adaptadas nesta tese, uma para ser utilizada nos textos fonte do córpus CSTNews e a outra nos sumários do mesmo córpus, para de compará-las com as versões criadas nesta tese. A adaptação foi mais profunda principalmente porque as relações RST foram utilizadas em vez das relações PDTB, ou seja, o preenchimento da grade discursiva (termo x sentenças) foi feito com relações RST.

O uso das relações RST foi devido à ausência de anotações de relações do PDTB no córpus CSTNews e pelas semelhanças entre ambos conjuntos de relações, já que tais conjuntos possuem, em sua maioria, relações locais (relações dentro de uma única sentença ou entre 2 sentenças adjacentes).

A implementação do modelo de Termo com RST seguiu os mesmos passos dos modelos de Grade de Entidades com Discurso (ver Figura 5.1). Entretanto, o Gerador de Grades de Entidades e/ou Discursivas desse modelo cria uma grade com os radicais dos termos, em vez de entidades (segundo o modelo original de Lin et al. (2011)).

A grade de termos é representada por uma matriz formada por linhas, que representam as sentenças do texto/sumário, e por colunas que são os termos (palavras de classe aberta) em sua forma radical. As células dessa matriz são preenchidas com relações RST que ocorrem de

forma local (versão para sumários) ou considerando todas as relações RST entre as sentenças (versão para textos fonte). A informação de nuclearidade das relações RST não foi utilizada nesse modelo para evitar o aumento da esparsidade, e também, para manter a semelhança entre as relações RST e as relações do PDTB, essas últimas não possuem a informação de nuclearidade. A Figura 5.21 ilustra parte de uma grade do modelo de Termo com RST de um sumário multidocumento.

	polic	desvi	corrupc	destin	vend	manha	lider	grup	...
S1	-	purpose	purpose	purpose	purpose	purpose	-	-	...
S2	attribution	attribution	-	-	-	-	-	attribution	...
S3	-	-	-	-	-	-	elaboration-e	-	...
S4	sequence	-	-	-	-	-	-	-	...

Figura 5.21: Grade discursiva do modelo Termo com RST

O vetor de características, utilizado na fase de aprendizado desse modelo, também segue os mesmos procedimentos feitos no modelo de Grade de Entidades com Discurso. O fato de usar termos em vez de entidades pode aumentar a esparsidade na grade discursiva do modelo Termo com RST, pois houve um aumento no número das colunas na grade discursiva desse modelo, e consequentemente, o aumento no número de células dessa grade que não necessariamente serão preenchidas por completo. Isso pode prejudicar o aprendizado dos possíveis padrões discursivos encontrados nos sumários/textos fonte. Os resultados desse modelo aplicado em sumários multidocumento e textos fonte do córpus CSTNews serão mostrados na Seção 5.6.

5.4 Modelo de Entidades com RST Local

No trabalho de Feng et al. (2014), descrito no Capítulo 3, foram desenvolvidos 2 modelos de avaliação de coerência local baseados no trabalho de Barzilay & Lapata (2008) e Lin et al. (2011). Os autores denominaram esses modelos de Completo e Superficial. O modelo Completo usa todos os relacionamentos RST que existem entre as sentenças, enquanto que o modelo Superficial só usa relacionamentos RST entre EDUs de uma única sentença ou de sentenças adjacentes (relacionamento local). O modelo Completo é similar à abordagem do modelo SINTÁTICA-SALIÊNCIA-RST+ desenvolvida nesta tese (ver Seção 5.1). Assim, o modelo SINTÁTICA-SALIÊNCIA-RST+ é considerado a adaptação do modelo Completo de Feng et al. (2014).

A adaptação do modelo Superficial é uma junção do modelo SINTÁTICA-SALIÊNCIA-RST+ (pois usa entidades e relações RST) com o modelo de Termos com RST (pois usa as relações RST locais).

Todos os módulos que compõem o modelo de Entidades com RST Local também são os mesmos que formam os modelos de Grade de Entidades enriquecido com discurso (ver Figura 5.1). A diferença ainda continua nos módulos Gerador de Grade de Entidades e/ou Discursivas e Gerador de Vetor de Características com Discurso.

O Gerador de Grade de Entidades e/ou Discursivas desse modelo cria uma grade representada por uma matriz, a qual é formada por linhas (sentenças) e colunas (entidades). As células são preenchidas com relações RST que ocorrem em uma mesma sentença ou entre sentenças adjacentes do texto/sumário. A Figura 5.22 mostra parte de uma grade produzida por esse gerador.

	madrugada	ano	enchente	rio	rede	chuva	Reino_Unido	...
S1	-	non-volitional-result same-unit	non-volitional-result same-unit	-	non-volitional-result same-unit	non-volitional-result	non-volitional-result	...
S2	-	-	-	non-volitional-result	-	-	non-volitional-result	...
S3	-	-	-	-	-	-	-	...
S4	same-unit	-	-	elaboration-e	-	same-unit	-	...

Figura 5.22: Grade discursiva do modelo de Entidades com RST Local

A grade da Figura 5.22 possui entidades que fazem parte de EDUs que, por sua vez, se relacionam com mais de uma EDU, ou seja, as células correspondentes às entidades "ano", "enchente" e "rede" com a sentença S1 possuem as relações RST *non-volitional-result* e *same-unit*. Tal fato é utilizado na formação do vetor de características dado pelo módulo Gerador de Vetor de Características com Discurso desse modelo.

De forma similar aos outros modelos, todos os resultados do Modelo de Entidades com RST Local na avaliação da coerência local de sumários multidocumento serão exibidos na Seção 5.6.

5.5 Modelo de Relações Discursivas

O modelo de Relações Discursivas considera que todo sumário multidocumento coerente possui padrões de relações discursivas (CST e RST) que o distingue dos sumários multidocumento incoerentes (menos coerentes).

Baseado no desempenho das grades discursivas, verificada nas versões dos modelos da literatura enriquecidos com discurso, o modelo de Relações Discursivas usa as relações CST e RST das sentenças dos sumários.

A Figura 5.23 mostra a estrutura do modelo de Relações Discursivas. Nessa estrutura, um sumário multidocumento com relações CST e/ou RST anotadas é a entrada do Gerador de Grade Discursiva. Por exemplo, a Figura 5.24 mostra o sumário 3 da coleção 37 do córpus CSTNews com a identificação da origem de cada sentença no final da mesma.

Nesse sumário da Figura 5.24, as sentenças S2 e S3 se relacionam por meio da relação CST *Follow-up*. A relação RST *sequence* acontece entre S1 e S4. Há também a relação RST *elaboration* entre as sentenças S1 e S3. Com a identificação das relações discursivas presentes no sumário, o Gerador de Grade Discursiva monta a respectiva grade discursiva, como pode ser visto na Figura 5.25.

Na grade da Figura 5.25, as sentenças do sumário são representadas nas linhas e nas colunas. As células dessa grade são preenchidas com relações discursivas em que as sentenças participam.

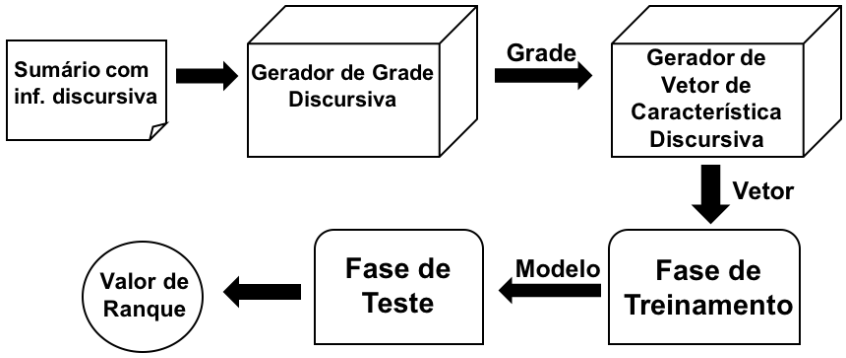

Figura 5.23: Estrutura do Modelo de Relações Discursivas

(S1) Terminou a rebelião de presos no Centro de Custódia de Presos de Justiça (CCPJ), em São Luís, no começo da tarde desta quarta-feira (17). <D2_C37_GPovo;S1>

(S2) 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. <D1_C37_OGlobo;S2>

(S3) O motim começou durante a festa do Dia das Crianças, realizada na terça-feira (16). <D2_C37_GPovo;S3>

(S4) Segundo informações da polícia, o líder da rebelião foi transferido para o Presídio de Pedrinhas, na capital maranhense. <D2_C37_GPovo;S5>

Figura 5.24: Sumário do córpus CSTNews

		(S1)	(S2)	(S3)	(S4)
		S1_D2_C37	S2_D1_C37	S3_D2_C37	S5_D2_C37
(S1)	S1_D2_C37		-	elaboration	sequence
(S2)	S2_D1_C37			Follow-up	-
(S3)	S3_D2_C37				-
(S4)	S5_D2_C37				

Figura 5.25: Grade discursiva do modelo de Relações Discursivas

Com essa grade discursiva, o Gerador de Características Discursivas utiliza as relações CST e/ou RST presentes nos sumários do CSTNews e calcula a probabilidade de cada um dos relacionamentos discursivos entre as sentenças para formar o vetor de características.

A probabilidade é calculada pela razão entre a frequência de uma relação específica e o número total de transições válidas da grade. Considere 2 sentenças S_i e S_j (onde i e j indicam a posição da sentença no sumário): se $i < j$, tem-se uma transição válida e o valor igual a 1 é adicionado ao número total de relacionamentos da grade. Considerando que as transições são visualizadas da esquerda para a direita na grade discursiva da Figura 5.25, as células em cinza não caracterizam uma transição válida, ou seja, somente a diagonal superior da grade é necessária nesse modelo para contabilizar o número total de transições válidas. Por exemplo, a probabilidade da relação RST *elaboration* na grade da Figura 5.25 é 0,16, ou seja, 1 ocorrência de *elaboration* em 6 transições possíveis. A Figura 5.26 mostra o vetor de características para a grade na Figura 5.25.

Follow-up	elaboration	sequence
0,16	0,16	0,16

Figura 5.26: Vetor de característica

Os vetores de características dos sumários serão utilizados como instâncias de treinamento do modelo preditivo ou instâncias de teste na geração do valor de ranque para o respectivo sumário, o qual servirá como parâmetro na avaliação do modelo na tarefa de ordenação de sentenças (ver Seção 5.6).

5.6 Experimentos e Resultados

A tarefa de ordenação de sentenças também foi utilizada na avaliação dos modelos com informação discursiva. Para isso, os mesmos textos fonte e sumários do córpus CSTNews, com as respectivas permutações, utilizados na avaliação dos modelos adaptados (ver Capítulo 4) também foram usados para avaliar os modelos discursivos apresentados nesse capítulo.

Todos os modelos que fazem uso de aprendizado de máquina também foram submetidos ao método de validação cruzada de 10 *folds*, cujos motivos foram explicitados na Seção 4.5 do Capítulo 4. Além disso, esses modelos utilizam o pacote de aprendizado SVM^{light} de Joachims (2002) com a opção de ranque, o que possibilitou a comparação entre os valores de ranque dos textos/sumários e de suas versões permutadas para contabilizar a acurácia.

A base de dados, como nos experimentos com modelos da literatura adaptados, foi formada por pares de textos (*texto original*, *versão permutada do texto original*) ou pares de sumários multidocumento (*sumário de referência*, *versão permutada do sumário de refência*). Os sumários formados por humanos foram considerados de referência.

Outro fato que merece ser lembrando é que alguns sumários do córpus não foram utilizados devido a quantidade de sentenças que estes possuem ser inferior a 4 (quatro), restrição esta que inviabilizava as 20 permutações para tais sumários de referência necessárias para a tarefa de ordenação de sentenças.

O cálculo da acurácia dos modelos com discurso também foi o mesmo utilizado nos experimentos dos modelos da literatura adaptados, ou seja, razão entre o número de pares corretos (valor de ranque ou de coerência do texto/sumário de referência maior do que o da versão permutada) e a quantidade total de pares.

Para o modelo SINTÁTICA-SALIÊNCIA-RST+, foi utilizado apenas o córpus de textos fonte do CSTNews, pois, nesse modelo, todos os relacionamentos RST dos textos fonte são utilizados e não apenas os relacionamentos entre sentenças adjacentes possíveis nos sumários multidocumento. A Tabela 5.1 mostra os resultados obtidos pelo modelo SINTÁTICA-SALIÊNCIA-RST+ e por suas duas variações (Variação 1 e 2).

Os resultados da Tabela 5.1 mostraram que a combinação da informação de nuclearidade

Tabela 5.1: Resultados do modelo SINTÁTICA-SALIÊNCIA-RST+ e suas Variações

Modelo	**Acurácia (%)**
SINTÁTICA-SALIÊNCIA-RST+	79,45
Variação 1	66,18
Variação 2	63,99

com as relações RST foi necessária para a obtenção da melhor acurácia do modelo SINTÁTICA-SALIÊNCIA-RST+ em relação as suas variações, principalmente em comparação a Variação 2, que usou as mesmas relações do modelo SINTÁTICA-SALIÊNCIA-RST+, mas sem a informação de nuclearidade das relações RST. Além disso, o modelo preditivo da Variação 1 não conseguiu ser o mais eficaz entre os três modelos dessa abordagem, ou seja, os padrões aprendidos das transições dos agrupamentos de relações RST presentes nos textos fonte coerentes não evidenciaram a distinção entre os textos coerentes e incoerentes como se esperava. Entretanto, em comparação com a Variação 2 (a qual não possuía informação de nuclearidade), a Variação 1 teve a acurácia superior.

Em comparação direta com o melhor modelo de Grade de Entidades adaptado nesta tese (SINTÁTICA-SALIÊNCIA-), o modelo discursivo SINTÁTICA-SALIÊNCIA-RST+ teve uma pequena melhora na acurácia da avaliação da coerência local nos textos fonte do córpus CSTNews, ou seja, de 78,97% para 79,45%.

As versões do modelo de Grade de Entidades com informação sintática e sem informação de saliência enriquecidas com informação discursiva (SINTÁTICA+SALIÊNCIA- com CST, SINTÁTICA+SALIÊNCIA- com Categoria CST, SINTÁTICA+SALIÊNCIA- Booleana CST, SINTÁTICA+SALIÊNCIA- CST e RST, SINTÁTICA+SALIÊNCIA- com RST, SINTÁTICA+SALIÊNCIA- Booleana RST e SINTÁTICA+SALIÊNCIA- Booleana CST e RST) e as versões do modelo de Grade de Entidades sem informação sintática e sem saliência enriquecidas com informação discursiva (SINTÁTICA-SALIÊNCIA- com CST, SINTÁTICA-SALIÊNCIA- com RST, SINTÁTICA-SALIÊNCIA- com CST e RST) foram avaliadas no córpus de sumários multidocumento do CSTNews. Os resultados de todas essas versões na tarefa de ordenação de sentenças são mostrados na Tabela 5.2.

De acordo com a Tabela 5.2, as versões enriquecidas somente com relações CST proporcionaram maiores valores de acurácia, tanto no modelo que usa informação sintática quanto no modelo que não usa. Tais valores são compreensíveis, pois os sumários avaliados são do tipo multidocumento e a tendência é que os padrões de relações CST se destaquem em relação aos padrões RST, até porque há mais relações CST do que RST presentes nos sumários.

Com exceção das versões Booleanas e da versão com Categoria CST, as outras versões do modelo de Grade de Entidades com o enriquecimento por meio de relações discursivas (CST e RST) tiveram resultados superiores a todos os modelos de Grade de Entidades adaptados e aplicados a sumários multidocumento. Esses resultados mostram que as informações discursivas utilizadas como agregadoras de conhecimento em modelos de Grade de Entidades para

Tabela 5.2: Resultados das versões do modelo de Grade de Entidades enriquecidas com discurso

Modelos	**Acurácia (%)**
SINTÁTICA+SALIÊNCIA- com CST	**91,31**
SINTÁTICA-SALIÊNCIA- com CST	91,13
SINTÁTICA-SALIÊNCIA- com RST	84,47
SINTÁTICA+SALIÊNCIA- com RST	81,85
SINTÁTICA-SALIÊNCIA- com CST e RST	76,80
SINTÁTICA+SALIÊNCIA- CST e RST	75,14
SINTÁTICA+SALIÊNCIA- com Categoria CST	53,41
SINTÁTICA+SALIÊNCIA- Booleana CST e RST	37,06
SINTÁTICA+SALIÊNCIA- Booleana RST	32,78
SINTÁTICA+SALIÊNCIA- Booleana CST	32,53

avaliar a coerência local podem proporcionar um ganho máximo na acurácia em relação a versão SINTÁTICA+SALIÊNCIA-, do modelo de Grade de Entidades original, de aproximadamente 52% , ou seja, o modelo SINTÁTICA+SALIÊNCIA- original adaptado obteve 60,21% (ver Tabela 4.2) de acurácia na distinção de sumários coerentes dos incoerentes, e o modelo SINTÁTICA+SALIÊNCIA- com CST (modelo enriquecido com informação discursiva) obteve 91,31% de acurácia (ver Tabela 5.2), uma diferença de 31,10%, e isso equivale a aproximadamente 52% do valor de acurácia obtido pelo modelo SINTÁTICA+SALIÊNCIA- original adaptado. Assim, o valor de 52% é a porcentagem de acurácia a mais obtido pelo modelo SINTÁTICA+SALIÊNCIA- com CST, ou seja, o valor de ganho com o uso de informação discursiva no modelo SINTÁTICA+SALIÊNCIA- aplicado a sumários multidocumento.

Os modelos que utilizaram somente relações RST tiveram bons resultados e confirmaram que muitos dos sumários de referência foram formados com sentenças do mesmo texto fonte e mantiveram a ordem original das sentenças.

Outro fato interessante foi que a junção das informações CST e RST no enriquecimento do modelo de Grade de Entidades teve um resultado mediano em comparação aos modelos que utilizaram uma ou outra informação discursiva. Além disso, os modelos de Grade de Entidades com enriquecimento de informação booleana das relações discursivas tiveram resultados bem inferiores ao esperado. Isso mostra a variabilidade que pode ocorrer com esses modelos, isto é, aspectos como o tipo de informação que está sendo utilizada para avaliar a coerência e a quantidade de informação e de exemplos (sumários) pode interferir no aprendizado do modelo preditivo e, consequentemente, no seu julgamento.

O modelo baseado em Grafo com Discurso foi avaliado considerando as projeções *one mode* P_U e P_W, além da informação de distância entre as sentenças dos sumários. A Tabela 5.3 mostra a acurácia de cada projeção na tarefa de ordenação de sentenças.

Os resultados da Tabela 5.3 mostraram que o grafo de Projeção *one mode* P_U teve a melhor acurácia entre os grafos de projeções avaliados. Entretanto, a diferença máxima entre eles foi de 1,79% (comportamento similar ao modelo adaptado), o que pode ser justificado pela própria

Tabela 5.3: Resultado do modelo baseado em Grafo com Discurso

Modelos sem inf. de distância	**Acurácia (%)**
Projeção *one mode* P_U	**80,22**
Projeção *one mode* P_W	79,66
Modelos com inf. de distância	**Acurácia (%)**
Projeção *one mode* P_U	78,50
Projeção *one mode* P_W	78,43

forma de gerar o valor de coerência do modelo, a qual é baseada na soma dos pesos das relações discursivas.

A Tabela 5.4 mostra a porcentagem de ganho proporcionado pelo uso de informações discursivas para cada uma das projeções com ou sem informação de distância do modelo baseado em Grafo.

Tabela 5.4: Valores de ganho do modelo baseado em Grafo com Discurso

Modelos sem inf. de distância	**Ganho (%)**
Projeção one mode P_U	39,05
Projeção one mode P_W	44,98
Modelos com inf. de distância	**Ganho (%)**
Projeção one mode P_U	48,92
Projeção one mode P_W	**53,15**

O trabalho de Lin et al. (2011) foi implementado e adaptado nesta tese com o nome de modelo de Termo com RST. Tal nome foi escolhido devido ao uso de termos (palavras de classe aberta) e relações RST (em vez das relações do PDTB que não estão anotadas no córpus CSTNews). Esse modelo foi avaliado em textos fonte e em sumários multidocumento do CSTNews. A Tabela 5.5 mostra as acurácias do modelo de Termo com RST.

Tabela 5.5: Resultado do modelo Termo com RST

Modelo	**Acurácia (%)**	
	Textos fonte	**Sumários**
Termo com RST	70,80	53,23

Como as relações RST em sumários multidocumento acontecem apenas localmente (intra sentencial e/ou entre sentenças adjacentes), o modelo de Termo com RST conseguiu capturar melhor padrões de relações RST dos termos nos textos fonte (considera todas as relações RST) do que nos sumários. Tal comportamento é visto nos resultados mostrados na Figura 5.5.

Outro ponto que merece destaque é que o uso de termos pode ser melhor explorado para a criação de modelos que avaliem a coerência local, já que o modelo de Termos com RST

obteve um ganho de aproximadamente 10% em relação a Variação 2 do modelo SINTÁTICA-SALIÊNCIA-RST+, o qual se diferencia apenas no uso de entidades em vez de termos.

O modelo de Entidades com RST Local de Feng et al. (2014) foi avaliado considerando apenas o córpus de sumários multidocumento do CSTNews. Todas as etapas da tarefa de ordenação de sentenças foram seguidas, como em todas as avaliações realizadas nos outros modelos, com o intuito de avaliar o modelo de Entidades com RST Local, que obteve **48,92%** de acurácia. Esse valor significa que o modelo não conseguiu obter um padrão a partir das relações RST locais de tal modo que este padrão pudesse distinguir corretamente a maioria dos pares de sumários na fase de teste do modelo.

Pelo resultado do modelo de Termos com RST aplicado também ao mesmo córpus do modelo de Entidades com RST Local, pode-se afirmar que não é recomendável o uso de apenas relações RST como informação principal, devido a baixa acurácia de duas abordagens similares de RST local.

Por fim, o modelo de Relações Discursivas desenvolvido nesta tese também foi avaliado em sumários multidocumento do córpus CSTNews, por meio da tarefa de ordenação de sentenças. Esse modelo teve o valor de acurácia de **92,69%**. O modelo de Relações Discursivas captura o padrão discursivo em função das relações CST e RST e não em função das entidades que participam das relações como acontece nas versões enriquecidas do modelo de Grade de Entidades, o que possibilitou um treinamento mais eficiente do modelo preditivo e, consequentemente, um ranqueamento mais adequado para sumários de referência, caracterizando assim, a melhor distinção entre sumários coerentes e incoerentes. A Tabela 5.6 resume os resultados de todos os modelos de coerência que fazem uso de informações discursivas para a tarefa de ordenação de sentenças em sumários automáticos multidocumeto.

Neste capítulo os modelos de coerência enriquecidos e novos modelos discursivos propostos nesta tese foram descritos, e os resultados da tarefa de ordenação de sentenças, em que cada um dos modelos foram submetidos, também foram mostrados. As grades de entidades e/ou discursivas dos modelos de coerência, mostrados neste capítulo, apresentaram certa esparsidade. Tal esparsidade foi ocasionada pelos tipos de informações consideradas nos respectivos modelos e que pode prejudicar o desempenho do modelo. Mesmo com a esparsidade, o padrão da distribuição das informações discursivas nos sumários de referência (coerentes) pode ser capturado pelos modelos preditivos, possibilitando assim, uma melhora considerável na acurácia da distinção entre os sumários coerentes e os incoerentes em relação aos modelos originais adaptados nesta tese.

Um aspecto importante que merece ser mencionado é a não utilização de testes estatísticos nos resultados. Tal medida é devido ao fato que, a maioria dos trabalhos relacionados não fazem uso de tal artifício. Além disso, há muitos questionamentos sobre a confiabilidade de tais testes, como pode ser visto em Demăr (2008) e Nuzzo (2014).

O Capítulo 6 define e exemplifica erros linguísticos que acontecem nos sumários automáticos e que podem ser relacionados aos modelos desenvolvidos nesta tese, proporcionando um cenário alternativo da utilização desses modelos.

Tabela 5.6: Resumo dos resultados de todos modelos de coerência que utilizam informação discursiva

Modelos	**Acurácia (%)**
Relações Discursivas	**92,69**
SINTÁTICA+SALIÊNCIA- com CST	91,31
SINTÁTICA-SALIÊNCIA- com CST	91,13
SINTÁTICA-SALIÊNCIA- com RST	84,47
SINTÁTICA+SALIÊNCIA- com RST	81,85
Projeção *one mode* P_U	80,22
Projeção *one mode* P_W	79,66
SINTÁTICA-SALIÊNCIA-RST+	79,45
Projeção *one mode* P_U	78,5
Projeção *one mode* P_W	78,43
SINTÁTICA-SALIÊNCIA- com CST e RST	76,8
SINTÁTICA+SALIÊNCIA- CST e RST	75,14
Variação 1 do modelo SINTÁTICA-SALIÊNCIA-RST+	66,18
Variação 2 do modelo SINTÁTICA-SALIÊNCIA-RST+	63,99
SINTÁTICA+SALIÊNCIA- com Categoria CST	53,41
Termo com RST (Lin et al., 2011)	53,23
Entidades com RST Local (Feng et al., 2014)	48,92
SINTÁTICA+SALIÊNCIA- Booleana CST e RST	37,06
SINTÁTICA+SALIÊNCIA- Booleana RST	32,78
SINTÁTICA+SALIÊNCIA- Booleana CST	32,53

Capítulo

6

Métodos de Coerência Aplicados a Sumários Automáticos Multidocumento com Erros de Qualidade Linguística

Todos os modelos adaptados e desenvolvidos nesta tese foram avaliados segundo a tarefa de ordenação de sentenças. Tal tarefa vem sendo utilizada na literatura como o método padrão para medir o desempenho dos modelos na distinção da coerência local de sumários coerentes dos incoerentes. Essa tarefa, entretanto, faz uso de permutações aleatórias de sentenças dos sumários de referência (sumários humanos) para produzir, de forma artificial, versões incoerentes utilizadas na tarefa. Já os sumarizadores automáticos multidocumento podem produzir sumários quase coerentes e não totalmente incoerentes, ou seja, sumários que podem apresentar uma variação na quantidade e nos tipos de erros que afetam a coerência. Por exemplo, os sumários das Figuras 6.1, 6.2 e 6.3 ilustram alguns erros que podem afetar a coerência.

(S1) O Instituto de pesquisa CNI/Ibope divulgou nesta sexta-feira, em Brasília, que o presidente Luiz Inácio Lula da Silva seria reeleito em primeiro turno caso as eleições fossem hoje.
(S2) Se a eleição fosse hoje, o presidente Luiz Inácio Lula da Silva, candidato à reeleição, teria 44% das intenções de voto, contra 25% do tucano Geraldo Alckmin, de acordo com a pesquisa CNI/Ibope divulgada nesta sexta-feira.
(S3) De acordo com a pesquisa, Lula (PT) tem 44% das intenções de voto, contra 25% de Geraldo Alckmin (PSDB) e 11% de Heloísa Helena (PSOL).
(S4) Heloísa Helena, candidata à Presidência pelo PSOL, aparece em terceiro, com 11% das intenções de voto, seguida por Cristovam Buarque (PDT) e Luciano Bivar (PSL), ambos com 1%.

Figura 6.1: Sumário automático da coleção 2 do córpus CSTNews

(S1) Os deputados acusados de envolvimento na máfia dos sanguessugas têm até a meia-noite desta segunda-feira para renunciar aos mandatos se quiserem escapar da cassação.
...
(S5) Termina hoje, às 20 horas, o prazo para que os deputados acusados de participar do esquema dos sanguessugas renunciem para escapar da abertura de processo por quebra de decoro parlamentar.

Figura 6.2: Parte de um sumário automático da coleção 16 do córpus CSTNews

(S1) "Estamos chocados".
(S2) O agressor também morreu.
(S3) Após a tragédia, os dois adolescentes cometeram suicídio.
(S4) "Estávamos todos conectados à internet tentando descobrir o que aconteceu."
(S5) "Quando nos liberaram temporariamente, ele voltou a atirar", continuou ela.
(S6) "Eles nos trancaram nos quartos", afirmou Kanode.
(S7) A universidade estabeleceu um local de encontro entre famílias e representantes estudantis.

Figura 6.3: Sumário automático da coleção 18 do córpus CSTNews

A Figura 6.1 mostra um sumário com informação redundante entre as sentenças. A sentença S2, por exemplo, possui a mesma informação da sentença S1, a sentença S3 tem a mesma informação da sentença S2 e a sentença S4 exibe parte da informação da sentença S3. Essas redundâncias afetam negativamente a informatividade, o tamanho e o interesse do leitor pelo sumário.

Figura 6.2 apresenta parte de um sumário com informações contraditórias, sobre o horário máximo para os deputados renunciarem, envolvendo as sentenças S1 e S5. Esse tipo de informação faz o sumário confuso e impreciso para os leitores.

O problema da ordenação de sentenças é outro problema que acontece nos sumários automáticos multidocumento, ou seja, sentenças posicionadas no sumário de modo que a ligação da informação complementar entre elas seja impossível e, consequentemente, torna um sumário incoerente. Em outras palavras, a ordenação incorreta das sentenças pode ocasionar a mudança da sequência de eventos que afeta a leitura e a compreensão do sumário. A Figura 6.3 mostra parte de um sumário com o problema de ordenação.

Os problemas acima afetam a Qualidade Linguística (QL) de um sumário automático multidocumento. Segundo Nenkova et al. (2011), sumarizadores automáticos não tratam alguns aspectos linguísticos e isso pode afetar a informatividade e a coerência de seus sumários. Por exemplo, a Figura 6.4 mostra um sumário automático multidocumento com um elemento da Qualidade Linguística (em negrito) que não foi corretamente tratado e que pode atrapalhar a compreensão do sumário.

No sumário da Figura 6.4, o acrônimo “ONU” não foi explicado no texto e isso pode prejudicar a informatividade e a compreensão do sumário. Uma simples definição do acrônimo no sumário resolveria esse problema.

Outro problema linguístico que está diretamente ligado à coesão referencial está presente

(S1) Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
(S2) Segundo uma porta-voz da **ONU**, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
(S3) Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.

Figura 6.4: Sumário automático da coleção 1 do córpus CSTNews

na primeira sentença (ver Figura 6.5) em um dos sumários automáticos da coleção 3 do córpus CSTNews. Nesse exemplo, 2 sintagmas nominais, em negritos, estão referenciando entidades que não estão no sumário. Esse tipo de problema afeta a coerência do sumário, porque o leitor não sabe qual a empresa e qual o avião que o sumário menciona.

(S1) Na quarta-feira, o presidente d**a empresa**, Marco Antonio Bologna, havia declarado que **o avião** passará por checagem justamente no dia 13 e estava em perfeito estado.
...

Figura 6.5: Parte de um sumário automático da coleção 3 do córpus CSTNews

Outro problema que pode prejudicar a QL de um sumário é apresentado na Figura 6.6. Essa figura mostra parte de um sumário da coleção 22 do CSTNews que menciona a entidade "Tom Jobim" (em negrito) sem nenhuma definição dessa entidade. Como a entidade "Tom Jobim" foi mencionada pela primeira vez no sumário, tal entidade necessita de maior clareza sobre o que ela é.

(S1) Passageiros de um vôo que deveria sair dia 17 para Belém ainda aguarda a autorização para partida da aeronave.
(S2) A madrugada foi de muita irritação no **Tom Jobim**, com muitos passageiros dormindo pelo chão do terminal.
...

Figura 6.6: Parte de um sumário automático da coleção 22 do córpus CSTNews

Os exemplos acima mostraram que os sumarizadores automáticos não trataram problemas relacionados a QL. Entretanto, antes de resolver tais problemas, é preciso compreendê-los, identificá-los e avaliá-los. Assim, um estudo sobre erros da QL em sumários automáticos multidocumento foi realizado. Com este estudo, o desempenho dos modelos desenvolvidos nesta tese também pôde ser analisado, além de um possível relacionamento dos modelos de coerência com sumarizadores automáticos e com a informatividade dos sumários.

Para que o estudo de erros da QL fosse possível, inicialmente, um levantamento dos possíveis erros juntamente com as suas definições foi realizado. Na Seção 6.1, os erros de QL serão definidos e exemplificados. Além disso, a tarefa de anotação desses erros nos sumários automáticos multidocumento realizada nesta tese será descrita.

6.1 Anotação de Erros de Qualidade Linguística

Para a tarefa de anotação, um córpus de sumários automáticos multidocumento foi formado. Esse córpus foi composto de sumários automáticos multidocumento do português brasileiro. Tais sumários foram gerados por 4 sumarizadores: GistSumm, de Pardo (2002); RSumm, de Ribaldo (2013); RC-4, de Cardoso (2014); e MTRST-MLAD, de Castro Jorge (2015). Esses sumarizadores foram escolhidos por causa de suas diferentes abordagens (superficial e profunda) e também por serem considerados os principais sumarizadores automáticos multidocumento para o português brasileiro. Cada sumarizador produziu um sumário para cada coleção do CST-News, totalizando 200 sumários automáticos multidocumento, ou seja, 50 sumários gerados por cada sumarizador.

Com o córpus pronto, os erros de Qualidade Linguística foram definidos. Tais erros foram baseados nos trabalhos de Koch & Travaglia (2002), Otterbacher et al. (2002), Pitler et al. (2010), Kaspersson et al. (2012), e Friedrich et al. (2014), que estudaram aspectos que afetam a Qualidade Linguística dos textos e os problemas oriundos da sumarização multidocumento, como informação redundante, informação contraditória e ordenação de sentenças. Baseados nos trabalhos da literatura e na análise realizada no córpus de sumários automáticos multidocumento, os erros de Qualidade Linguística foram divididos em 3 (três) categorias: erros relacionados a Menções de Entidades (nível de entidades), Violações de Gramaticalidade e Redundância (nível sentencial) e Outros (problemas que não foram listados nas categorias anteriores).

Todos os erros foram identificados no córpus por marcadores em xml. Em geral, o marcador tem o formato: **<e TYPE=***(nome_do_erro)***>***(entidade/sentença)***</e>**. Em alguns marcadores, há informação adicional colocada depois do campo *(nome_do_erro)*. Essa informação adicional será explicada quando determinado erro fizer uso da mesma. O campo "*(nome_do_erro)*" é preenchido com o nome do erro identificado no campo "*(entidade/sentença)*".

6.1.1 Erros relacionados a Menções de Entidades

Os erros relacionados a Menções de Entidades são: Primeira Menção sem Explicação (1M-EXP), Menções Subsequentes com Explicação (nM+EXP), Sintagma Nominal Definido sem Referência a Menções Anteriores (SNdef-REF), Sintagma Nominal Indefinido com Referência a Menções Anteriores (SNind+REF), Pronome sem Antecedente (PRO-ANT), Pronomes com Antecedentes Enganosos (PRO_ENG), Acrônimos sem Explicação (ACR-EXP).

Primeira menção sem explicação (1M-EXP) é atribuída a primeira menção de uma entidade nomeada[1] para a qual falta uma referência clara para o leitor. Para identificar esse erro, os anotadores não deviam utilizar conhecimento de mundo, e sim, verificar se houve a explicação explícita da entidade no texto. Tal procedimento foi adotado de forma que não houvesse nenhuma influência externa ao texto. Se aparecer, por exemplo, *Itaú*, sem dizer que é um banco,

[1]Entidades pertecentes as categorias pré definidas, tais como: pessoa, organização, lugar, entre outras (Zaccara, 2012). Estas categorias podem variar de acordo com objetivo do estudo.

este deveria ser marcado como 1M-EXPL. No exemplo 1, não se sabe o que é *Tepco*, e, no exemplo 2, falta definição do que é *Itaú*.

1. A **<e TYPE=1M-EXP>**Tepco**</e>** inicialmente declarou que o tremor não havia causado vazamentos, mas, mais tarde, revelou que 1.200 litros de água com materiais radioativos da usina haviam vazado para o mar.

2. Em comparação com a receita obtida nos seis primeiros meses de 2006, de R$ 2,958 bilhões, o lucro do **<e TYPE=1M-EXP>**Itaú**</e>** cresceu 36% neste ano.

O erro menções subsequentes com explicação (nM+EXP) acontece quando há menções de entidades nomeadas que aparecem com uma introdução explicativa inapropriada. No exemplo 3, explica-se novamente na segunda sentença quem é Leomar Quintanilha. O mesmo ocorre com o exemplo 4: explica-se novamente o que é CET na segunda sentença.

3. O presidente do Conselho de Ética do Senado, Leomar Quintanilha (PMDB-TO), disse hoje ser contrário à unificação dos processos contra o senador Renan Calheiros (PMDB-AL) que tramitam na Casa Legislativa.

   <e TYPE=nM+EXP SENT=S3 TEXT= "O presidente do Conselho de Ética do Senado, Leomar Quintanilha (PMDB-TO)" > O presidente do conselho, Leomar Quintanilha (PMDB-TO)**</e>**, disse que é contra a união das representações, mas que vai colocar a proposta em votação.

4. Às 9 horas, a cidade tinha 113 km de lentidão, sendo que a média para o horário é de 82 km, segundo a Companhia de Engenharia de Tráfego (CET).

   Naquele horário, segundo **<e TYPE=nM+EXP SENT=S4 TEXT= "a Companhia de Engenharia de Tráfego (CET)">** a CET (Companhia de Engenharia de Tráfego) **</e>**, havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76 km.

O campo **SENT** contém a identificação da sentença em que a primeira menção da entidade especificada no campo **TEXT** ocorre.

Sintagma nominal definido sem referência a menções anteriores (SNdef-REF) são sintagmas nominais definidos geralmente usados no texto para se referirem às entidades que já estão presentes no contexto do discurso. Assim, os anotadores marcaram os Sintagmas Nominais Definidos que violam esta regra. No exemplo 5, o erro está na última sentença, na qual "porta-voz" aparece como uma unidade definida que não faz referência a nenhuma entidade nas sentenças anteriores.

5. Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

   Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.

<e TYPE=SNdef-REF>O porta-voz**</e>** informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.

Sintagma nominal indefinido com referência a menções anteriores (SNind+REF) são sintagmas nominais indefinidos usados para introduzir novas entidades no discurso. Assim, os anotadores marcaram os Sintagmas Nominais Indefinidos com Referência a Menções Anteriores que violam esta regra. No exemplo 6, o erro está ao chamar um Airbus A320, pois se trata de uma entidade já definida.

6. O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13.

   O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, **<e TYPE=SNind+REF SENT=S6 TEXT= "O Airbus-A320">** um Airbus A320**</e>**, continuou voando, com o reverso direito desligado.

Pronome sem antecedente (PRO-ANT) ocorre quando um pronome não tem antecedente sintaticamente possível no sumário, ou seja, não há antecedente que combina em gênero e número. No exemplo 7, o pronome "ele" aparece na primeira sentença do sumário, sendo impossível saber de quem se fala.

7. Internado em um hospital em Buenos Aires, **<e TYPE=PRO-ANT>** ele **</e>** teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.

   "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".

Pronomes com antecedentes enganosos (PRO_ENG) ocorre quando uma expressão anafórica refere-se a um antecedente enganoso e seu antecedente correto não está presente no texto. No caso de sumários, pode ser necessário consultar o texto-fonte para identificação do antecedente correto. No exemplo 8, o pronome "ele" (segunda sentença) parece conectar-se à entidade "Kaká" ou à entidade "Ronaldinho" (primeira sentença), mas, no texto-fonte, o pronome refere-se ao jogador "Robinho", que não aparece no sumário. Além de identificar o tipo de erro, é importante deixar explícito o antecedente enganoso, usando o atributo ANT (antecedente) e colocando entre aspas o antecedente. Quando houver mais de um antecedente enganoso, eles devem aparecer separados por vírgula.

8. Aos 27, Kaká arriscou de muito longe e Ronaldinho colocou o desviou o chute.

   A 20cm da linha de fundo **<e TYPE=PRO_ENG ANT="Kaká, Ronaldinho">** ele **</e>** deu dois dribles humilhantes no zagueiro equatoriano e cruzou para Elano, que fez o quarto, aos 37.

Acrônimos sem explicação (ACR-EXP) são marcados quando os mesmos não foram explicados no sumário. Nos exemplos (9) e (10), consideram-se acrônimos sem explicação "Deic" e "PF".

Alguns acrônimos são de senso comum, tais como siglas de estados e partidos. Esses casos devem ser identificados com o atributo SC (Senso Comum) e colocar entre aspas o significado do acrônimo, conforme se vê na anotação do exemplo 10. Diferentemente do erro "Primeira menção sem explicação", a utilização do senso comum, que faz uso do conhecimento de mundo de cada anotador, apenas ajudou no acréscimo da informação de senso comum junto a marcação do erro "Acrônimos sem explicação". O acréscimo da informação de senso comum não interferiu na interpretação de uma possível marcação ou não do erro "Acrônimos sem explicação" por parte dos anotadores, como pode ocorrer no erro "Primeira menção sem explicação" caso o conhecimento de mundo fosse utilizado.

9. O outro suspeito tem 27 anos, é grafiteiro e, segundo o **<e TYPE=ACR-EXP>**Deic**</e>**, tem passagem por roubo, mas já cumpriu a pena.

10. A **<e TYPE=ACR-EXP SC="Polícia Federal">** PF **</e>** não soube informar se esse tipo de recompensa é paga para órgãos policiais.

6.1.2 Erros relacionados a Violações de Gramaticalidade e Redundância

Os erros relacionados a Violações de Gramaticalidade e Redundância são: Informação Redudante (RED), Contradição (CONTR), Sentenças Incompletas (SENT_INC), Sem Relacionamento Semântico (SEM_REL), Conectivo/Marcador Discursivo sem Contexto Apropriado (MD).

O erro de informação redundante (RED) é marcado quando há informações redundantes (total ou parcial) afetam negativamente os sumários. Os exemplos 12, 14, 16 e 17 mostram informações redundantes.

11. Na segunda etapa, a parte concluída será reaberta e a obra passará a ser feita na outra cabeceira.

12. **<e TYPE=RED SENT=S11>**Na segunda parte, a outra cabeceira será reformada e, na terceira etapa, o centro da pista será reformado.**</e>**

13. Uma bomba caseira foi jogada contra o prédio do Ministério Público, no centro da capital, mas não deixou feridos.

14. **<e TYPE=RED SENT=S13>**Uma bomba de fabricação caseira explodiu em frente ao prédio do Ministério Público Estadual e lojas vizinhas também foram atingidas por estilhaços.**</e>**

15. A Receita Federal intensificou a fiscalização sobre as declarações das pessoas físicas neste ano.

16. **<e TYPE=RED SENT=S15>**A Receita Federal intensificou a fiscalização e o resultado foi um aumento do número de contribuintes que caíram na malha fina. **</e>**

17. **<e TYPE=RED SENT=S15,S16>**Dobrou o número de pessoas físicas autuadas depois de cair na malha fina até julho, de acordo com a Receita Federal do Brasil.**</e>**

    A expectativa da Receita é que até o final do ano mais de 300 mil contribuintes sejam autuados pela malha fina.

A contradição (CONTR) ocorre quando duas sentenças apresentam informações conflitantes. O exemplo 16 mostra a contradição de informações sobre a quantidade de mortos e feridos.

16. O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.

    <e TYPE=CONTR SENT=S16> Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas nesta segunda-feira após a colisão de dois trens de passageiros no delta do Nilo, ao norte do Cairo, informaram fontes policiais e médicas.**</e>**

A sentença incompleta (SENT_INC) pode ocorrer na forma de sentenças incompletas, falta de sinais de pontuação ou espaços. Nesse caso, o erro se aplica a toda sentença. No exemplo 17, a última sentença está incompleta (terminando com uma vírgula).

17. Como esperado, a atleta Fabiana Murer conquistou a medalha de ouro no salto com vara nos Jogos Pan-Americanos do Rio, nesta segunda-feira, no Estádio João Havelange.

    <e TYPE=SENT_INC>Murer conquistou o lugar mais alto do pódio com a marca de 4m60, contra 4m40 da norte-americana April Steiner,**</e>**

O erro sem relacionamento semântico (SEM_REL) é marcado quando sentenças adjacentes que não possuem qualquer relacionamento semântico, i.e., casos em que o leitor imagina o que a sentença *x* tem a ver com a sentença *y*. No exemplo 18, não é possível entender porque o candidato Lula foi à pista de dança.

18. Após um fim de semana no Norte e Nordeste ao lado de caciques pefelistas adeptos de uma campanha mais ofensiva e com discursos duros contra o presidente Luiz Inácio Lula da Silva, o candidato do PSDB à Presidência, Geraldo Alckmin, deixou ontem a linha "paz e amor" e se curvou à temperatura alta do debate eleitoral.

    Alckmin acusou Lula de arrogante, de subestimar a inteligência dos brasileiros e relacionou o presidente aos escândalos do mensalão, sanguessuga e ao caso Waldomiro Diniz.

    No mesmo dia em que Lula se comprometeu a não atacar, o adversário tucano Geraldo Alckmin elevou o tom do seu discurso.

    Sem citar nominalmente o adversário, Alckmin criticou de novo Lula ao comentar especulações de que o petista, convicto na vitória no primeiro turno, já estaria fazendo planos sobre sua nova equipe ministerial.

<e TYPE=SEM_REL> Após comer, o candidato foi até a pista de dança e, ao som de Alcione, foi disputado pelas senhoras da velha guarda da escola.**</e>**

"Não sou nenhum expert, mas gosto de dançar. Conheci a Lu (sua mulher) assim, num baile em Pinda (Pindamonhangaba, sua cidade natal)".

Conectivo/Marcador discursivo sem contexto apropriado (MD) ocorre em sentenças que possuem marcadores discursivos explícitos ('mas', 'porque', 'porém') que não são mais apropriados no contexto do sumário. No exemplo 19, o marcador discursivo "Contudo" não possui ligação com a sentença anterior. O campo CONEC contém explicitamente o marcador discursivo que não foi utilizado apropriadamente.

19. Em meio ao tráfico de drogas constante, uma praça está quase pronta bem ao lado do fluxo de viciados na cracolândia, na Luz (centro de São Paulo).

    <e TYPE=MD CONEC = "Contudo"> Contudo, a secretária Municipal de Assistência Social não informou qual seria o prazo de entrega previsto. **</e>**

6.1.3 Outros tipos de erros

O erro OUTRO aconteça algum problema que não está listado em algum dos tipos acima, os anotadores deveriam anotar com OUTRO, com a explicação do erro no atributo EXPLANATION. A anotação inclui a etiqueta OUTRO para a sentença completa ou para um ponto específico da sentença, dependendo do problema. Os exemplos 20, 21, 22, 23, 24, 25 e 26 mostram a marcação desse tipo de erro.

20. Além de Rafael Nadal, o torneio contará com mais três atletas classificados entre os 20 melhores do ranking da ATP: o espanhol Nicolás Almagro (11º colocado e tricampeão do Brasil Open), o argentino Juan Mônaco (12º) e o suíço Stanilas Wawrinka (17º).

    A organização do **<e TYPE=Outros EXPLANATION="referência em português para termo introduzido em inglês">**Aberto do Brasil 2013**</e>** anunciou na manhã desta terça-feira que o torneio a ser disputado em fevereiro, no ginásio do Ibirapuera, em São Paulo, marcará a volta do espanhol Rafael Nadal às quadras.

21. Também **<e TYPE=Outros EXPLANATION="Falta sujeito">**disse**</e>** não saber em que momento **<e TYPE=Outros EXPLANATION="Falta sujeito">**foi filmado**</e>**.

22. A câmera da emissora teria registrado **<e TYPE=Outros EXPLANATION="Sintagma com referente ambíguo">**a cena**</e>** logo depois da transmissão da notícia, o que sugere uma comemoração festiva da matéria veiculada na TV.

23. Após vários desentendimentos com o então ministro da pasta, Roberto Brant, **<e TYPE=Outros EXPLANATION="Sem fonte">**foi demitida**</e>**.

24. **<e TYPE=Outros EXPLANATION="Inclusão de metadados">**RIO e NOVA YORK - **</e>**O presidente Luiz Inácio Lula da Silva abriu nesta terça-feira a Assembléia Geral da Organização das Nações Unidas (ONU) reforçando o discurso do secretário-geral Ban Ki-moon sobre a importância dos países desenvolvidos liderarem os esforços pela preservação do meio ambiente.

25. **<e TYPE=Outros EXPLANATION="Inclusão de metadado">**RIO - **</e>**A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos do Rio.

26. O tempo estabelecido pelos brasileiros ainda derrubou o recorde **<e TYPE=Outros EXPLANATION="Grafia diferente para entidade já mencionada">**Pan-Americano**</e>**, que pertencia aos Estados Unidos (7min18s93), em Santo Domingo (2003).

6.2 A Tarefa da Anotação de Erros Linguísticos

O objetivo dessa tarefa foi identificar os erros linguísticos, definidos nas Subseções 6.1.1, 6.1.2 e 6.1.3, nos sumários multidocumento gerados automaticamente.

O córpus de sumários automáticos multidocumento com anotação de erros linguísticos será útil para pesquisas que envolvem a produção de sumários informativos e coerentes, avaliação da coerência local (como o caso desta tese), tratamento dos principais problemas que afetam a QL dos sumários multidocumento, etc.

A tarefa foi realizada em grupo e de forma presencial. Além disso, tal tarefa foi realizada em 1 hora por dia em local e horário específico. Tal hora reservada a cada dia fez a tarefa menos exaustiva para os anotadores e isso pode ter influenciado positivamente a qualidade da anotação. Além disso, a tarefa pôde ser melhor gerenciada com todos os anotadores no mesmo lugar.

Inicialmente, 2 dias foram destinados para treinar os 6 anotadores (2 linguistas e 4 cientistas da computação) para esclarecer os procedimentos para a realização da tarefa. Esses anotadores foram escolhidos devido a experiência que cada um possui em PLN e em tarefa de anotação.

Devido a subjetividade da tarefa, os erros da QL foram somente marcados depois de um consenso entre os anotadores. Esta estratégia é interessante porque a mesma produz uma anotação mais concisa e correta. Entretanto, o tempo da tarefa foi mais longo em comparação com as estratégias tradicionais (em que cada anotador anota sumários diferentes por dia). Neste trabalho, a duração da tarefa de anotação foi de aproximadamente 150 dias.

A estratégia utilizada nesta anotação está relacionada a 2 questões em aberto dadas por Hovy & Lavid (2010), quando os autores analisam a concordância da tarefa de anotação. Para Hovy & Lavid,"Quanto de desacordo pode ser tolerado antes de refazer a anotação ou mudar a sua teoria ou sua instanciação?" e "Se o juiz (nesse caso, uma pessoa fora do grupo de anotadores) deveria ver a anotação feita pelos anotadores (a qual poderia influenciar a sua decisão) ou não?" Não há um entendimento sobre a melhor forma de realizar uma anotação, desde que ela reflita

o aspecto a ser considerado e siga um protocolo pré-estabelecido para a realização da tarefa. Dessa forma, todos os anotadores foram considerados juízes e a decisão de anotar um erro foi por meio de um senso comum entre os anotadores.

Erro como **Sem Relacionamento Semântico** necessitou de mais atenção e refinamento em sua interpretação. Em consequência disso, tal erro teve um alto grau de subjetividade, exigindo, assim, discussões entre os anotadores até chegar em um acordo entre todos ou pelo menos entre a maioria dos anotadores na marcação de um erro. De acordo com Hovy & Lavid (2010), esse processo de concordância pode ser realizado em tarefas de anotação consideradas complexas, como é o caso da tarefa de anotação de erros da QL proposta nesta tese.

O erro **Acrônimos sem Explicação** precisou do conhecimento de mundo de cada anotador para preencher o campo SC exigido na marcação desse erro. Para cada anotador, esse conhecimento de mundo pode ser diferente e isso pode causar a identificação inadequada do erro. Entretanto, a abordagem de anotação adotada neste trabalho evita esse tipo de problema.

Mesmo com todos os anotadores trabalhando juntos, a concordância entre eles foi verificada periodicamente. Nesses casos, cada anotador trabalhou separadamente com os mesmos sumários dos outros, e, depois disso, a concordância foi calculada por meio da medida Kappa. A medida Kappa foi usada para verificar a compreensão dos erros linguísticos pelos anotadores e, consequentemente, avaliar a dificuldade da tarefa.

6.3 Resultados e Análises da Anotação

Para os sumarizadores considerados nesta tese (GistSumm, MTRST-MLAD, RSumm e RC-4), 1359 erros linguísticos foram anotados. A Tabela 6.1 mostra a quantidade de erros encontrados nos sumários de cada sumarizador.

Tabela 6.1: Total de erros anotados nos sumários de cada sumarizador

Sumarizador	Erros Anotados	% de Erros
GistSumm	521	38,33
MTRST-MLAD	421	30,97
RC-4	220	16,20
RSumm	197	14,50

A Tabela 6.1 mostra que o sumarizador GistSumm apresenta mais erros do que os outros. Isso pode ser explicado pela grande quantidade de informações repetidas que o GistSumm produz em seus sumários. Além disso, Informação Redudante (RED) é o erro mais recorrente, com um total de 261 ocorrências (veja Tabela 6.2). Esses dados confirmam que um dos principais problemas relacionados a sumarização multidocumento é a informação reduntante.

A Tabela 6.3 mostra a quantidade de erros do tipo Informação Redundante (RED) anotada nos sumários de cada sumarizador.

Erros de redundância podem aumentar os problemas da categoria Menções de Entidades

Tabela 6.2: Quantidade de erros para cada tipo

Erro	Quantidade de Erros	% Erros
RED - Informação Redundante	261	19,20
ACR-EXP - Acrônimos sem Explicação	255	18,76
SNdef-REF - Sintagma Nominal Definido sem Referência a Menções Anteriores	182	13,39
nM+EXP - Menções Subsequentes com Explicação	152	11,18
SEM_REL - Sem Relacionamento Semântico	136	10,00
OUTRO	123	9,05
1M-EXP - Primeira Menção sem Explicação	103	7,57
CONTR - Contradição	41	3,01
MD - Conectivo/Marcador Discursivo sem Contexto Apropriado	37	2,72
PRO-ANT - Pronome sem Antecedente	30	2,20
SNind+REF - Sintagma Nominal Indefinido com Referência a Menções Anteriores	25	1,83
SENT_INC - Sentenças Incompletas	11	0,80
PRO_ENG: Pronomes com Antecedentes Enganosos	3	0,29

Tabela 6.3: Total de erros anotados do tipo Informação Redundante (RED)

Sumarizador	Quantidade de Erros	% Erros
GistSumm	160	61,30
MTRST-MLAD	55	21,08
RC-4	23	8,81
RSumm	23	8,81

porque os problemas dessa categoria podem estar embutidos nas sentenças redundantes. Por exemplo, a Figura 6.7 mostra parte de um sumário que ilustra tal situação.

(S1) Uma nova série de ataques criminosos foi registrada na madrugada desta segunda-feira, dia 7, em São Paulo e municípios do interior paulista.
(S2) Uma bomba caseira foi jogada contra o prédio do Ministério Público, na capital do estado.
(S3) As ações criminosas podem ter sido ordenadas pelos líderes do Primeiro Comando da Capital (PCC), que haviam prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo.
(S4) Na região do ABC Paulista pelo menos dez ônibus foram incendiados - sete em Mauá e três em Santo André.
(S5) **<e TYPE=RED SENT=S2><e TYPE=SNind+REF SENT=S2 TEXT= "Uma bomba caseira">**Uma bomba caseira**</e>** foi atirada contra a sede **<e TYPE=nM+EXP SENT=S2 TEXT="do Ministério Público">**do Ministério Público (MP)**</e>**.**</e>**
(S6) O prédio da secretaria da Fazenda, no centro, foi atingido por três bombas caseiras.
(S7) **<e TYPE=RED SENT=S3>**Os líderes da facção criminosa PCC haviam prometido **<e TYPE=SNind+REF SENT=S1 TEXT="Uma nova série de ataques criminosos">**uma nova onda de ataques**</e>** caso o Ministério Público de São Paulo negasse a saída temporária de presos em virtude do Dia dos Pais.**</e>**
...
(S17) **<e TYPE=RED SENT=S1,S3,S7,S8,S10>**Integrantes do PCC haviam prometido **<e TYPE=SNind+REF SENT=S1 TEXT="Uma nova série de ataques criminosos">**uma nova onda de ataques**</e>** caso o Ministério Público de São Paulo negasse a saída temporária de presos em virtude do Dia dos Pais.**</e>**
...

Figura 6.7: Parte de um sumário produzido pelo GistSumm

As sentenças com informação repetida no sumário da Figura 6.7 (como em S5, S7 e S17) apresentam outros erros da categoria Menções de Entidades. Nesse caso, para cada erro do tipo Informação Redundante (RED), havia um erro do tipo Sintagma Nominal Indefinido com Referência a Menções Anteriores (SNind+REF). Isso pode explicar a alta quantidade de erros anotados nos sumários produzidos pelo GistSumm, o qual, aparentemente, não gerencia corretamente as informações extraídas dos textos fonte para formar seus sumários.

Em relação a quantidade de erros por categoria, a Tabela 6.4 mostra os dados dessa quantidade.

Tabela 6.4: Quantidade de erros por categorias

Sumarizador	**Categorias**		
	Menções de Entidades	**Gramaticalidade e Redundância**	**Outros**
GistSumm	239	221	61
MTRST-MLAD	252	129	40
RC-4	123	83	14
RSumm	136	53	8
Total	**750**	**486**	**123**

De acordo com a Tabela 6.4, a categoria Menções de Entidades foi a categoria de erro mais anotada, 750 vezes. O fato que essa categoria teve a quantidade de erros mais alta foi esperado, já que há mais entidades do que sentenças em um sumário. Por exemplo, o sumário na Figura 6.8 foi gerado pelo sumarizador RSumm e esse não apresentou erros da categoria de Violações da Gramaticalidade e Redundância, mas 5 erros anotados foram relacionados a categoria Menções de Entidades e 1 erro da categoria Outro.

(S1) **<e TYPE=SNdef-REF>**No segundo turno**</e>**, as intenções de voto do presidente Lula caíram de 53% em junho para 50% em julho, enquanto o candidato Alckmin subiu de 29% para 36%.
(S2) A **<e TYPE=ACR-EXP>**CNI**</e>** explica que a pesquisa não traz a comparação com pesquisas anteriores para primeiro turno porque é a primeira vez que o **<e TYPE=ACR-EXP>**Ibope**</e>** utiliza a lista oficial de candidatos a presidente da República.
(S3) Embora não permita comparações, vale relembrar que na pesquisa de junho Lula tinha 48% das intenções de voto; Alckmin 18% e **<e TYPE=1M-EXP>**Heloísa Helena**</e>**, 5%.
(S4) A margem de erro é de dois pontos percentuais, para mais ou para menos.
(S5) **<e TYPE=Outros EXPLANATION="Sintagma com referente ambiguo">**A pesquisa**</e>** foi realizada entre os dias 29 e 31 de julho e foi registrada no **<e TYPE=ACR-EXP>**TSE**</e>** com o número 12.197/2006.

Figura 6.8: Parte de um sumário produzido pelo GistSumm

Segundo a Tabela 6.4, os sumarizadores RC-4 (profundo) e RSumm (superficial) apresentaram uma quantidade de erros mais baixa do que os outros. Em particular, o sumarizador RSumm teve a quantidade de erros anotada mais baixa nas 3 categorias. Tais dados mostram que os desenvolvedores dos sumarizadores RC-4 e RSumm tiveram uma preocupação maior

em selecionar um conteúdo único e mais ordenado, tal preocupação já elimina grande parte dos erros que afetam a qualidade linguística dos sumários.

Mesmo com a anotação feita em grupo, decidiu-se medir a concordância entre os anotadores para averiguar a compreensão dos tipos de erro. Para isso, a medida Kappa e a concordância da maioria para 4 coleções de sumários (C12, C22, C32 e C42) do córpus CSTNews foram calculadas. Note que cada coleção tem 1 sumário gerado por cada sumarizador (GistSumm, RSumm, RC-4 e MTRST-MLAD), ou seja, 4 sumários em cada coleção. A Tabela 6.5 mostra os valores Kappa de concordância entre os anotadores sobre a presença ou não de um erro em uma sentença, independente do tipo.

Tabela 6.5: Medida Kappa pela marcação de um erro ou não

Coleções	**Kappa**
C12	0,409
C22	0,641
C32	0,578
C42	0,324
Média	**0,488**

Segundo a Tabela 6.5, a medida Kappa na coleção 22 teve o melhor resultado. Entretanto, devido a dificuldade da tarefa, o resultado da medida Kappa para a coleção 42 não foi bom como o esperado, ficando abaixo do considerado satisfatório. A subjetividade causa diferentes interpretações principalmente quando os anotadores anotam de forma isolada. Esse comportamento é repetido pelos dados mostrados na Tabela 6.6, onde mediu-se a Kappa para as categorias de erros, ou seja, verificou-se a concordância entre os anotadores sobre a marcação de erro da mesma categoria.

Tabela 6.6: Medida Kappa por Categorias

Coleções	**Categorias**		
	Menções de Entidades	**Gramaticalidade e Redundância**	**Outros**
C12	0,356	0,560	-
C22	0,670	0,537	0,902
C32	0,552	0,616	0,627
C42	0,606	0,418	0,751
Média	**0,546**	**0,533**	**0,760**

Com os resultados da medida Kappa não tão expressivos, mas em grande parte satisfatórios, a concordância pela maioria também foi considerada. Dessa forma, a porcentagem das sentenças de todas as coleções (as quais participam do processo de concordância) em que a maioria dos anotadores concordam foi calculada. A Tabela 6.7 mostra os resultados da concordância pela maioria, considerando a marcação de um erro em uma certa sentença.

Tabela 6.7: Concordância pela maioria na identificação de um erro em uma sentença

Coleções	**% das Sentenças**
C12	100
C22	100
C32	91,89
C42	81,25

A Tabela 6.7 mostra que a maioria dos anotadores concordaram em identificar um erro em todas as sentenças das coleções C12 e C22. Nas coleções C32 e C42, os anotadores também tiveram uma boa porcentagem de concordância da maioria.

A concordância pela maioria também foi utilizada para as categorias de erro. Assim, a porcentagem das sentenças pelas quais a maioria identificou um erro de uma categoria específica foi calculada. A Tabela 6.8 mostra os resultados obtidos por essa medida de concordância.

Tabela 6.8: Concordância pela maioria em identificar um erro de uma categoria

Coleções	**% das Sentenças para Menções de Ent.**	**% das Sentenças para Viol. de Gramat. e Redun.**	**% das Sentenças para Outros**
C12	100	100	-
C22	100	100	100
C32	94,59	91,89	100
C42	90,62	71,87	93,75

De acordo com a Tabela 6.8, os anotadores concordaram entre eles em 100% das sentenças nas coleções C12 e C22 nas categorias Menções de Entidades e Violações de Gramaticalidade e Redundância. Para a categoria Outros, a maioria dos anotadores concordaram em 100% das sentenças nas coleções C22 e C32. Na coleção C42, a categoria Violações de Gramaticalidade e Redundância foi a única em que a maioria dos anotadores concordaram abaixo dos 90% das sentenças. Esses resultados mostraram que a maioria dos anotadores compreenderam todos os tipos de erros linguísticos identificados nos sumários. Para confirmar isso, a Tabela 6.9 mostra a porcentagem de sentenças para as quais todos os anotadores (100%) concordaram na identificação de um certo erro linguístico.

Segundo a Tabela 6.9, mais da metade das sentenças pertencentes às coleções tinham 100% de concordância entre os anotadores na identificação de certo erro linguístico. Todas as sentenças da coleção 12 que continham o erro ACR-EXP foram anotadas por todos os anotadores. Os erros PRO_ENG e SENT_INC não foram identificados nas coleções usadas na concordância.

Pelos resultados de concordância da medida Kappa e da concordância pela maioria, pode-se afirmar que a compreensão dos erros da QL foi muito boa. Segundo a literatura, os resultados da medida Kappa podem ser considerados satisfatórios e, na concordância pela maioria pode-se afirma a confiabilidade da anotação dos erros da QL, mesmo levando em consideração a subje-

Tabela 6.9: Concordância de 100% dos anotadores para cada erro

Erros	% Sent. em C12	% Sent. em C22	% Sent. em C32	% Sent. em C42
1M-EXP	54,54	90,00	91,89	81,25
nM+EXP	81,81	76,66	84,48	90,62
SNdef-REF	63,63	93,33	83,78	53,12
SNind+REF	-	-	89,18	-
PRO-ANT	-	-	-	96,87
ACR-EXP	100	96,66	94,59	93,75
SEM_REL	81,81	76,66	75,67	75,00
MD	-	-	91,89	96,87
RED	-	83,33	89,18	81,25
CONTR	-	86,66	94,59	-
OUTROS	-	96,66	81,08	90,62

tividade dos anotadores. Além disso, os resultados da anotação mostraram que tanto os sumarizadores baseados na informação discursiva (abordagem profunda) quanto os sumarizadores considerados superficiais (não profunda) apresentaram uma quantidade considerada de erros da QL, os quais afetam a coerência. Isso é devido a pouco ou nenhum tratamento adequado dos erros que afetam a QL por parte do sumarizadores. Acredita-se que os desenvolvedores de tais sumarizadores não possuíam conhecimentos de grande parte dos erros levantados nesta tese. Assim, esses sumarizadores e os futuros podem ser beneficiados com esse levantamento, para que tais erros sejam identificados e tratados em sua maioria e, consequentemente, haja sumários altamente informativos e coerentes. Além disso, acredita-se que a metodologia utilizada no processo de anotação de erros linguísticos foi a mais adequada para esse tipo de tarefa complexa.

Na próxima seção, o possível relacionamento dos modelos da coerência local com os erros da QL dos sumários multidocumento serão mostrados e analisados.

6.4 Experimentos e Resultados

Os experimentos sobre o córpus de sumários automáticos multidocumento anotados com erros da QL tiveram os propósitos de: verificar, na prática, se os modelos de coerência local desenvolvidos nesta tese são capazes de identificar os erros da QL nos sumários automáticos multidocumento e verificar o possível relacionamento entre a informatividade e a coerência local dos sumários gerados pelos sumarizadores considerados nesta tese.

Para medir a influência negativa de um erro de QL na coerência local de um sumário, foi considerado que a frequência de um determinado erro em um sumário deveria ser maior do que os outros erros presentes nesse sumário. Tal restrição foi denominada de FREQ1. De forma mais rigorosa, a frequência de um erro em um sumário deveria ser maior do que a soma da frequência dos outros erros no sumário. Tal restrição, considerada mais rigorosa, foi deno-

minada de FREQ2. Dessa forma, levantamentos da quantidade de sumários que possuem um determinado erro em FREQ1 e em FREQ2 foram realizados.

A Figura 6.9 mostra um sumário gerado pelo MTRST-MLAD para a coleção 7 do córpus CSTNews, em que o erro Sem Relacionamento Semântico (SEM_REL) tem a frequência igual a 3 e os erros Conectivo/Marcador Discursivo sem Contexto Apropriado (MD), Sintagma Nominal Definido sem Referência a Menções Anteriores (SNdef-REF), Primeira Menção sem Explicação (1M-EXP) e OUTROS tem 1 ocorrência. Assim, tal sumário foi usado no subcórpus de sumários que tiveram o erro SEM_REL com maior frequência do que os outros erros (FREQ1). Já a Figura 6.10 mostra um sumário, gerado pelo sumarizador RC-4 para a coleção 21 do córpus CSTNews, que possui 3 ocorrências do erro Informação Redundante (RED) e 1 ocorrência do erro Sintagma Nominal Indefinido com Referência a Menções Anteriores (SNind+REF). Dessa forma, esse sumário poderia ser usado no subcórpus de sumários em que o erro RED tem a frequência maior do que a soma dos outros erros (FREQ2), além de ser usado também no subcórpus de sumários com a restrição de FREQ1. A Tabela 6.10 mostra a quantidade de sumários para cada erro em FREQ1 e FREQ2.

(S1) <e TYPE=MD CONEC="Mas">Mas**</e>** "é provável que **<e TYPE=SNdef-REF>**estes planemos irmãos**</e>** se formaram juntos, a partir de uma nuvem de gás em contração que se partiu em duas, como uma estrela binária em miniatura", disse **<e TYPE=1M-EXP>**Jayawardhana**</e>**.
(S2) O mais recente, Oph 162225-240515, é o primeiro planemo duplo.
(S3) Nos últimos cinco anos, astrônomos identificaram algumas dezenas de objetos ainda menores que anãs marrons que não estão presos a nenhum sistema estelar, apelidados de objetos de massa planetária, ou planemos, localizados nos arredores de regiões de formação de estrelas.
(S4) <e TYPE=SEM_REL>Cerca de metade das estrelas do tamanho do Sol existem em pares.**</e>**
(S5) <e TYPE=SEM_REL><e TYPE=Outros EXPLANATION="Declaração sem Fonte">"A mera existência é uma surpresa, e a origem e o destino são um mistério"**</e>**.**</e>**
(S6) <e TYPE=SEM_REL>"É um par de irmãos admirável, cada um com cerca de 1% da massa do Sol", disse Jayawardhana.**</e>**

Figura 6.9: Sumário da coleção 7 do CSTNews gerado pelo MTRST-MLAD

(S1) A pista principal do Aeroporto Internacional de São Paulo (Cumbica), em Guarulhos, será totalmente reformada em março de 2008, segundo informações do Ministério da Defesa anunciadas nesta segunda-feira, 6.
(S2) <e TYPE=RED SENT=S1>O ministro da Defesa, Nelson Jobim, decidiu que será realizada **<e TYPE=SNind+REF SENT=S1 TEXT="A pista principal do Aeroporto Internacional de São Paulo (Cumbica), em Guarulhos, será totalmente reformada">**uma reforma definitiva na pista principal de Guarulhos**</e>**, o mais rápido possível, de acordo com a assessoria do ministério da Defesa.**</e>**
(S3) De acordo com informações da Defesa, a primeira etapa da reforma será feita com a reforma de um terço da pista, em uma das cabeceiras.
(S4) Na segunda etapa, a parte concluída será reaberta e a obra passará a ser feita na outra cabeceira.
(S5) <e TYPE=RED SENT=S4>Na segunda parte, a outra cabeceira será reformada e, na terceira etapa, o centro da pista será reformado.**</e>**
(S6) <e TYPE=RED SENT=S3>De acordo com o Ministério da Defesa, na primeira etapa das obras, será reformado um terço da pista em uma das cabeceiras - o restante dela ficará disponível para pousos e decolagens.**</e>**

Figura 6.10: Sumário da coleção 21 do CSTNews gerado pelo RC-4

Tabela 6.10: Quantidade de sumários para cada erro em FREQ1 e FREQ2

Erros	Total de Sumários em FREQ1	Total de Sumários em FREQ2
1M-EXP	7	3
nM+EXP	8	5
SNdef-REF	21	8
SNind+REF	0	0
PRO-ANT	0	0
PRO_ENG	0	0
ACR-EXP	39	18
SEM_REL	13	3
MD	0	0
RED	37	13
CONTR	0	0
SENT_INC	1	1
OUTROS	5	3

Observando os dados da Tabela 6.10, os subcorpora em FREQ1 tiveram mais sumários em comparação aos de FREQ2. Mesmo a quantidade de sumários em FREQ1 não sendo a ideal para uma conclusão definitiva e abrangente da influência dos erros que fazem parte de FREQ1 nos sumários, os subcorpora em FREQ1 foram utilizados nos experimentos de relacionamento entre Erros Linguísticos e Sumarizadores Multidocumento, entre Erros Linguísticos e Modelos de Coerência e, por fim, entre Modelos de Coerência e Sumarizadores Multidocumento. Tais experimentos com seus respectivos resultados serão relatados nas Seções 6.4.1, 6.4.2 e 6.4.3.

6.4.1 Relacionamento entre Erros Linguísticos e Sumarizadores Multidocumento

Para ilustrar o relacionamento dos erros linguísticos com os sumarizadores automáticos multidocumento é necessário observar, na Tabela 6.11, os dados de porcentagem de ocorrência dos erros linguísticos nos sumários produzidos pelos 4 sumarizadores.

De acordo com a Tabela 6.11, o erro Informação Redundante (RED) é o principal problema em 2 dos 4 sumarizadores de abordagens diferentes, isto é, o sumarizador GistSumm de abordagem superficial e o sumarizador RC-4 de abordagem profunda. Esses dois sumarizadores deveriam verificar semanticamente as sentenças que já compõem o sumário com as próximas a serem extraídas dos textos fonte para evitar, ou pelo menos minimizar, o problema de informações redundantes.

O erro Acrônimos sem Explicação (ACR-EXP) teve a maior ocorrência no sumarizador superficial RSumm. Isso pode ter acontecido pelo fato de que as sentenças mais importantes de cada tópico que o RSumm seleciona para o sumário não continham a explicação do acrônimo presente em um dos tópicos. Caso um acrônimo com explicação fosse considerado uma restrição na escolha das sentenças, tal erro poderia ser evitado.

Tabela 6.11: Porcentagem de ocorrência de cada erro nos sumários produzidos pelos sumarizadores

Erros	Sumarizadores Multidocumento			
	MTRST-MLAD	**GistSumm**	**Rsumm**	**RC-4**
1M-EXP	10,69%	4,03%	12,18%	5,91%
nM+EXP	5,23%	16,51%	6,60%	14,09%
SNdef-REF	**25,42%**	3,45%	18,78%	9,09%
SNind+REF	0,95%	2,30%	2,03%	2,27%
PRO-REF	4,75%	0,77%	1,52%	1,36%
PRO_ENG	0,00%	0,19%	0,00%	0,91%
ACR-EXP	12,83%	18,62%	**27,92%**	22,27%
RED	5,46%	**30,71%**	11,68%	**25,00%**
CONTR	0,95%	4,80%	1,02%	4,55%
SENT_INC	0,71%	0,96%	1,02%	0,45%
SEM_REL	19,95%	4,22%	8,62%	5,91%
MD	3,56%	1,73%	4,57%	1,83%
OUTRO	9,50%	11,71%	4,06%	6,36%

No sumarizador MTRST-MLAD, 25,42% dos erros identificados foram relacionados ao erro Sintagma Nominal Definido sem Referência a Menções Anteriores (SNdef-REF), ou seja, erro fortemente relacionado com a quebra na ordem de sentenças vindas do mesmo texto fonte e possivelmente adjacentes. Assim, uma restrição de não mudar a ordem das sentenças vindas do mesmo texto seria conveniente para evitar esse tipo de erro. Além disso, o erro Sem relacionamento semântico (SEM_REL) tem uma alta frequência nos sumários gerados pelo sumarizador MTRST-MLAD em relação aos sumários gerados pelos outros sumarizadores, ou seja, o sumarizador MTRST-MLAD pode estar apresentando problemas na ordenação de sentenças ou na seleção de conteúdo. Em se tratando da seleção de conteúdo proporcionada pelo sumarizador MTRST-MLAD, a mesma assegurou que a frequência do erro de redundância foi a mais baixa entre os sumarizadores avaliados.

Exceto o erro de Pronomes com Antecedentes Enganosos (PRO_ENG), o qual não foi identificado nos sumários gerados por MTRST-MLAD e RSumm, todos os outros erros aconteceram pelo menos em 1 sumário de cada sumarizador.

Observando somente a Tabela 6.11 pode-se concluir que os sumarizadores MTRST-MLAD e GistSumm são os menos problemáticos em relação aos erros que afetam a QL, pois os mesmos apresentam as menores porcentagens de ocorrência em 11 dos 13 tipos de erros, ou seja, 6 tipos de erros (nM+EXP, SNind+REF, PRO_ENG, ACR-EXP, RED e CONTR) com as menores porcentagens ocorreram nos sumários gerados pelo sumarizador MTRST-MLAD e 5 tipos de erros (1M-EXP, SNdef-REF, PRO-REF, SEM_REL e MD) com as menores porcentagens ocorreram nos sumários gerados pelo sumarizador GistSumm. Entretanto, tais porcentagens são relacionadas ao número total de erros que os sumários de cada sumarizador apresentaram. Como pode ser observado na Tabela 6.1, os sumários de cada sumarizador apresentaram uma quantidade total de erros diferentes entre si, assim, a Tabela 6.11 mostra quais os erros que mais

afetam os sumários de cada sumarizado e que os sumarizadores RSumm e RC-4 apresentam um distribuição maior dos erros em seus sumários do que os sumarizadores MTRST-MLAD e GistSumm.

Segundo os resultados mostrados na Tabela 6.11, os sumarizadores não tratam ou tratam inadequadamente os erros que afetam a QL dos sumários. Assim, a identificação e um possível tratamento para os erros linguísticos expostos nesta tese fariam os sumários multidocumento mais coerentes. Utilizando os modelos de coerência desenvolvidos nesta tese, uma tentativa de identificação dos erros linguísticos foi proposta e descrita na Seção 6.4.2.

6.4.2 Relacionamento entre Erros Linguísticos e Modelos de Coerência

O relacionamento entre os erros linguísticos e os modelos de coerência é fundamentada na possibilidade de identificar os erros da QL dos sumários pelos modelos. O modelo de coerência local que melhor avaliar um sumário como incoerente que possuir um erro da QL de maior frequência do que outros erros será o mais propenso a identificar esse erro mais frequente. Para isso, a diferença entre o valor de ranque ou de coerência do sumário de referência (sumário humano considerado coerente) e o valor de ranque ou de coerência do sumário de teste (sumário automático que possui um determinado erro linguístico mais frequente) é calculada. Quanto maior for essa diferença, melhor o modelo conseguiu distinguir um sumário humano coerente do sumário automático com o erro linguístico. Para exemplificar tal abordagem, considere a necessidade de descobrir qual dos modelos de coerência adaptados ou com discurso desenvolvidos nesta tese é o mais adequado para avaliar um sumário que possui o erro Acrônimos sem explicação (ACR-EXP) de maior frequência. Para o erro ACR-EXP, tem-se 39 sumários automáticos de diferentes coleções do CSTNews em que tal erro é o mais frequente. Para cada um desses sumários, os modelos de coerência geram um valor de ranque ou de coerência (dependendo do tipo de modelo) para cada um dos sumários que possuem o erro ACR-EXP mais frequente, ou seja, para cada um dos 39 sumários. Além disso, os modelos de coerência também geram o valor de ranque ou de coerência para o sumário de referência (humano e coerente) da mesma coleção do CSTNews do seu respectivo sumário automático. Para verificar se um determinado modelo é mais adequado para avaliar um determinado erro, calcula-se para todos os sumários que possuem tal erro mais frequente, a diferença entre os valores de ranque ou de coerência (dependendo do tipo de modelo) do sumário humano e dos automáticos que pertencem a mesma coleção do CSTNews. O modelo que, em média, tiver a maior diferença, este será o mais adequado para avaliar o erro. No caso do erro ACR-EXP, houveram 39 valores das diferenças entre os valores de ranque ou de coerência calculadas em cada modelo de coerência; assim, a média das diferenças foi realizada para cada modelo de coerência. O modelo com a maior média é o mais adequado para avaliar o erro ACR-EXP.

A média das diferenças entre os valores de ranque ou de coerência entre os sumários humanos e automáticos só foi calculada quando necessária, pois, segundo a Tabela 6.10, os erros SNind+REF, PRO-ANT, PRO_ENG, MD e CONTR não puderam ser analisados, já que os mesmos não se enquadraram tanto na FREQ1 quanto na FREQ2. Além disso, o erro SENT_INC só

teve 1 ocorrência tanto na FREQ1 quanto na FREQ2.

O experimento de verificar qual o modelo de coerência que melhor avalia um erro linguístico foi realizado somente na FREQ1, ou seja, um erro que tinha a maior frequência em relação aos outros possíveis. Tal escolha é devida a quantidade de dados que já é pouca em FREQ1 e é ainda menor em FREQ2. A Tabela 6.12 mostra a diferença média dos valores de ranque ou de coerência para cada modelo de coerência sem informação discursiva em relação aos erros linguísticos.

Os valores negativos da Tabela 6.12 indicam que, em média, os valores de ranque ou de coerência para sumários automáticos foram maiores do que os de sumários humanos. Assim, a maioria dos modelos de coerência que não utilizaram discurso não conseguiram avaliar os sumários automáticos com erros linguísticos como incoerentes. Isso pode ter acontecido por causa da pouca quantidade de erros frequentes nos sumários automáticos, o que torna os sumários automáticos não tão incoerentes. Além disso, a própria quantidade de sumários automáticos não foi suficiente para verificar o desempenho dos modelos em uma variabilidade maior de sumários com erros linguísticos.

A expectativa é que haja indícios de que os modelos de coerência desta tese possam ser utilizados na avaliação dos erros da QL, pois de acordo com os dados em FREQ1 não houve muitos exemplos para cada tipo de erros de forma que se chegue a uma conclusão definitiva.

Observando a Tabela 6.12 o modelo SINTÁTICO+SALIÊNCIA- de Grade de Entidades é o mais recomendável para avaliar sumários automáticos que contenham erros mais frequentes como Informação Redudante (RED), Sem Relacionamento Semântico (SEM_REL), Acrônimos sem Explicação (ACR-EXP), Sintagma Nominal Definido sem Referência a Menções Anteriores (SNdef-REF) e Menções Subsequentes com Explicação (nM+EXP). Além disso, o modelo SINTÁTICO-SALIÊNCIA- foi outro modelo de Grade de Entidades que conseguiu distinguir melhor sumários humanos dos sumários automáticos com o erro Primeira Menção sem Explicação (1M-EXP). Esses resultados eram esperados, principalmente para os erros relacionados a Menções de Entidades, já que o modelo de Grade de Entidades é baseado na distribuição de entidades ao longo do sumário e isso pode ter influenciado na melhor avaliação da coerência dos sumários automáticos com os erros respectivos erros linguísticos.

Com relação aos erros Sentença Incompleta (SENT_INC) e OUTRO os modelos de Padrões Sintáticos com a opção *d-sequence* ($d = 1$ e $d = 2$, respectivamente), suavização de 0,001 e utilizando todas as expressões sintáticas das sentenças dos sumários foram os que tiveram as maiores diferenças. Entretanto, a quantidade de sumários automáticos com esses erros (SENT_INC = 1 e OUTRO = 3) é insuficiente para se chegar a uma conclusão sobre o modelo mais adequado para avaliar sumários com esses tipos de erros.

Os modelos com informação discursiva desenvolvidos nesta tese também foram utilizados no experimento que verifica se tais modelos podem ser utilizados na possível identificação de algum erro linguístico mais frequente em um sumário. A Tabela 6.13 mostra as médias das diferenças dos valores de ranque ou de coerência (dependendo do modelo de coerência) entre os sumários humanos e sumários automáticos com os respectivos erros linguísticos.

Tabela 6.12: Resultados das diferenças dos valores de ranque ou de coerência para modelos sem discurso

MODELOS	ERROS LINGUÍSTICOS							
	RED	SENT_INC	SEM_REL	ACR-EXP	SNdef-REF	nM+EXP	1M-EXP	OUTROS
LSA	-0,02210556	0,012335255	0,074376875	0,016223209	0,0616106	-0,086047191	0,007924349	0,008343898
Grade Entidades (Sintático-Saliencia-)	1,46895727	6,52439191	-0,54476912	2,69566975	1,03158216	4,40351480	**4,40351480**	2,46112054
Grade Entidades (Sintático-Saliencia+)	-1,146375602	-0,31332799	-1,501376504	-0,978492912	-1,087533028	-0,032354536	-0,606149259	-1,254741858
Grade Entidades (Sintático+Saliencia-)	**2,896848271**	8,4466062	**0,339365846**	**4,190288843**	**3,538775993**	**6,89025039**	0,2076339	1,13538198
Grade Entidades (Sintático+Saliencia+)	0,044480516	2,039369	-0,541271823	0,443521824	0,703555268	2,105371928	-0,436886394	-0,69850978
Grafo com Inf. Distância (P_U)	-0,14261518	0,625	0,042109125	0,043453043	0,110161708	-0,143183697	-0,097076204	0,017507937
Grafo com Inf. Distância (P_W)	-0,30198536	0,625	0,06610444	-0,048789218	0,099135631	-0,344804067	-0,168504776	-0,080269841
Grafo com Inf. Distância (P_{Acc})	-0,365121381	3,75	0,109435081	0,247596925	1,086612726	-0,410582011	-0,354505944	0,702412698
Grafo sem Inf. Distância (P_U)	-0,418747173	0,75	0,060763168	-0,047681923	0,142460317	-0,279166667	-0,036636776	-0,045555556
Grafo sem Inf. Distância (P_W)	-0,789991	0,75	0,074890387	-0,142377469	0,15175737	-0,695833333	-0,108065347	-0,192222222
Grafo sem Inf. Distância (P_{Acc})	-1,783441198	4,25	0,161624487	-0,372986553	1,353312032	-0,810416667	0,019475366	0,79
Padrões Sintáticos -1- d1-Suavização_0,1	-2,35E-18	1,28E-31	1,43E-48	-4,87E-30	-2,20E-13	-4,87E-24	-1,90E-27	2,79E-42
Padrões Sintáticos -25- d1-Suavização_0,1	-8,55E-14	2,48E-05	1,83E-07	-4,90E-15	2,04E-04	-5,33E-11	-1,44E-10	5,60E-17
Padrões Sintáticos -400- d1-Suavização_0,1	-2,46E-23	-1,14E-21	2,95E-33	-1,11E-35	-3,25E-15	-1,17E-22	-6,72E-23	4,71E-44
Padrões Sintáticos -1- d1-Suavização_0,01	-7,09E+04	8,83E-06	5,59E-09	-1,05E-02	-1,28E+01	-1,78E+00	-1,92E+01	2,34E-06
Padrões Sintáticos -25- d1-Suavização_0,01	-2,80E+00	-6,27E-02	1,02E-06	-3,15E-01	-1,49E-03	-1,27E+03	-1,31E+03	2,99E-05
Padrões Sintáticos -400- d1-Suavização_0,01	-1,45E-05	-3,79E-12	1,63E-22	-3,96E-15	-8,07E-08	-1,28E-04	-7,40E-06	2,81E-21
Padrões Sintáticos -1- d1-Suavização_0,001	-8,31E-04	9,14E+00	5,52E-02	-5,41E+01	-1,49E+04	-1,25E+03	-2,88E+05	**3,45E+00**
Padrões Sintáticos -25- d1-Suavização_0,001	-9,10E+00	-3,63E-01	1,18E-06	-8,85E-01	-7,91E-02	-3,72E+03	-3,14E+03	3,63E-05
Padrões Sintáticos -400- d1-Suavização_0,001	-6,56E-04	-3,58E-09	1,12E-20	-1,67E-13	-1,73E-06	-3,84E-04	-1,25E-04	4,90E-19
Padrões Sintáticos -1- d2-Suavização_0,1	-5,41E-02	1,28E-31	7,69E-02	2,63E-02	-1,43E-01	-3,04E-13	-2,86E-01	2,79E-31
Padrões Sintáticos -25- d2-Suavização_0,1	-5,41E-02	2,48E-16	7,69E-02	2,63E-02	-1,43E-01	-7,38E+00	-2,86E-01	5,60E-06
Padrões Sintáticos -400- d2-Suavização_0,1	-5,41E-02	-1,14E-21	7,69E-02	2,63E-02	-1,43E-01	-3,83E-12	-2,86E-01	4,71E-33
Padrões Sintáticos -1- d2-Suavização_0,01	-5,41E-02	8,83E-06	7,69E-02	2,63E-02	-1,43E-01	-2,05E+00	-2,86E-01	2,34E-06
Padrões Sintáticos -25- d2-Suavização_0,01	-5,41E-02	-6,27E-02	7,69E-02	2,63E-02	-1,43E-01	-1,76E+03	-2,86E-01	2,99E-05
Padrões Sintáticos -400- d2-Suavização_0,01	-5,41E-02	-3,79E-12	7,69E-02	2,63E-02	-1,43E-01	-4,60E-05	-2,86E-01	2,81E-21
Padrões Sintáticos -1- d2-Suavização_0,001	-5,49E-02	**9,14E+00**	7,69E-02	2,63E-02	-1,43E-01	-2,04E+03	-2,86E-01	3,45E+00
Padrões Sintáticos -25- d2-Suavização_0,001	-5,41E-02	-3,63E-01	7,69E-02	2,63E-02	-1,43E-01	-5,16E+03	-2,86E-01	3,63E-05
Padrões Sintáticos -400- d2-Suavização_0,001	-5,41E-02	-3,58E-09	7,69E-02	2,63E-02	-1,43E-01	-1,44E-03	-2,86E-01	4,90E-19
Padrões Sint.-Production-1-Suavização_0,1	-3,37E-08	7,96E-32	9,37E-40	-1,12E-19	-1,17E-14	-5,39E-14	-8,32E-17	1,81E-32
Padrões Sint.-Production-25-Suavização_0,1	-3,82E-03	2,42E-05	1,02E-07	-2,09E-05	2,12E-05	-1,01E+00	-1,34E+02	5,09E-06
Padrões Sint.-Production-400-Suavização_0,1	-2,00E-13	-1,06E-22	1,18E-33	-5,78E-25	-9,03E-18	-3,83E-13	-1,44E-10	2,79E-33
Padrões Sint.-Production-1-Suavização_0,01	-2,76E+04	2,87E-06	2,04E-09	-8,46E-03	-9,87E+00	-5,96E+00	-7,19E+00	6,57E-07
Padrões Sint.-Production-25-Suavização_0,01	-1,76E+00	-3,47E-02	5,59E-07	-2,84E-02	-8,54E-03	-2,91E+02	-1,28E+04	2,65E-05
Padrões Sint.-Production-400-Suavização_0,01	-2,13E-06	-5,96E-12	8,01E-23	-3,67E-15	-8,65E-09	-6,67E-06	-1,75E-04	8,65E-22
Padrões Sint.-Production-1-Suavização_0,001	-5,91E-04	7,47E+00	1,87E-01	-5,66E+00	-3,18E+04	-7,51E+03	-1,04E+05	1,84E+00
Padrões Sint.-Production-25-Suavização_0,001	-6,46E+00	-2,19E-01	6,48E-07	-9,19E-03	-4,64E-02	-9,08E+02	-3,14E+04	3,19E-05
Padrões Sint.-Production-400-Suavização_0,001	-1,28E-05	-6,43E-10	5,33E-21	-1,78E-13	-2,70E-07	-2,34E-04	-3,02E-03	1,19E-19

Tabela 6.13: Resultados das diferenças dos valores de ranque ou de coerência para modelos com discurso

MODELOS	ERROS LINGUÍSTICOS							
	RED	SENT_INC	SEM_REL	ACR-EXP	SNdef-REF	nM+EXP	1M-EXP	OUTROS
SINTÁTICA+SALIÊNCIA- com CST	0,485823779	0	0,094910258	0,141957131	0	0,576511721	0,569431258	-0,059033472
SINTÁTICA+SALIÊNCIA- CST e RST	0,14189286	0	-0,38096884	0,34004708	0,01533741	0,53711052	0,56471631	-0,24781098
SINTÁTICA+SALIÊNCIA- com RST	-0,133861686	0	-0,469550345	0,218021994	0,02900534	0,002542983	0,166134073	-0,24642743
SINTÁTICA-SALIÊNCIA- com CST	0,236156032	0	-0,111338938	-0,011709114	0	0,332716235	0,54449463	-0,13258327
SINTÁTICA-SALIÊNCIA- com CST e RST	0,03504862	0	-0,592592237	0,483914925	0,00983899	0,726435534	0,52958792	-0,55678441
SINTÁTICA-SALIÊNCIA-RST+	-0,246364104	0	-0,605856427	0,316487565	0,020092774	0,206209336	0,019175821	-0,447177333
Grafo com Discurso com Inf. Distância (PU)	-0,483674177	0,291666667	-0,03335854	-0,202509221	0,052947846	-0,249520503	-0,019217687	-0,221450617
Grafo com Discurso com Inf. Distância (PW)	-0,919299049	0,291666667	-0,026592158	-0,439784292	0,040235261	-0,789045966	-0,090646259	-0,420524691
Grafo com Discurso sem Inf. Distância (PU)	-1,015905013	**0,416666667**	-0,041312854	-0,361203295	0,077380952	-0,427083333	0,016439909	-0,390740741
Grafo com Discurso sem Inf. Distância (PW)	-1,762822262	**0,416666667**	-0,029560717	-0,693052257	0,058503401	-1,302083333	-0,066893424	-0,784259259
Entidades com RST Local	-0,04501219	-1,88153533	-0,339509233	0,634136594	0,511774034	0,24247131	**2,05920525**	0,458805418
Termo com RST	0,54358876	-1,60828196	**3,437823057**	**1,923075371**	1,150196165	**5,280541534**	-8,877487433	-3,032158567
SINTÁTICA+SALIÊNCIA- Booleana CST	**1,743037081**	-5,6028159	1,503536077	0,682458605	0,224109381	-0,843944325	-3,175698143	3,864588667
SINTÁTICA+SALIÊNCIA- Booleana CST e RST	0,45381397	-4,4700207	0,752688015	0,247753689	0,038470414	-0,843745463	-2,901702014	2,264796667
Relações Discursivas	0,227743232	0,013827	3,076109027	1,893699727	**3,549767548**	1,470658613	-2,6160679	**4,820774667**

Segundo a Tabela 6.13, o modelo de Termo com RST teve as maiores diferenças para os erros linguísticos: Sem Relacionamento Semântico (SEM_REL), Acrônimos sem Explicação (ACR-EXP) e Menções Subsequentes com Explicação (nM+EXP). Como esse modelo usa termo em vez de entidade, e também usa relações RST, que nos sumários multidocumento ocorrem entre sentenças adjacentes vindas do mesmo texto fonte, o modelo de Termo com RST conseguiu capturar padrões distintos entre os sumários humanos e os sumários automáticos que possuem tais erros linguísticos frequentes.

O modelo de Relações Discursivas pode ser usado para avaliar sumários que possuem os erros Sintagma Nominal Definido sem Menções Anteriores (SNdef-REF) e OUTROS como os

mais frequentes. Entretanto, o mesmo modelo também teve um bom desempenho para os erros Sem Relacionamento Semântico (SEM_REL) e Acrônimos sem Explicação (ACR-EXP), que tiveram as médias das diferenças bem próximas das melhores médias.

Para o erro Sentença Incompleta (SENT_INC), os modelos de Grafo com Discurso sem Informação de Distância para as projeções *one mode* P_U e P_W tiveram as mesmas e maiores médias das diferenças dos valores de coerência. Esses mesmos valores iguais devem-se ao fato de que houve apenas 1 sumário automático que contém tal problema mais frequente, o que gerou a mesma matriz de projeção *one mode* para P_U e P_W. O mesmo comportamento foi repetido quando a informação de distância foi utilizada nos modelos de Grafo com Discurso.

A Tabela 6.14 resume de forma geral, os modelos de coerência que podem ser os mais recomendados para uma possível identificação de erros da QL dos sumários automáticos multidocumento.

Tabela 6.14: Melhores médias das diferenças dos valores de ranque ou de coerência para cada erro linguístico

MODELOS	ERROS LINGUÍSTICOS							
	RED	SENT_INC	SEM_REL	ACR-EXP	SNdef-REF	nM+EXP	1M-EXP	OUTROS
Grade Entidades (Sintático+Saliência-)	**2,896848271**	8,4466062	0,339365846	**4,190288843**	3,538775993	**6,89025039**	0,2076339	1,13538198
Relações Discursivas	0,227743232	0,013827	3,076109027	1,893699727	**3,549767548**	1,470658613	-2,6160679	**4,820774667**
Termo com RST	0,54358876	-1,60828196	**3,437823057**	1,923075371	1,150196165	5,280541534	-8,877487433	-3,032158567
Grade Entidades (Sintático-Saliência-)	1,46895727	6,52439191	-0,54476912	2,69566975	1,03158216	4,40351480	**4,40351480**	2,46112054
Padrões Sintáticos-1-d2-Suavização_0,001	-5,49E-02	**9,14E+00**	7,69E-02	2,63E-02	-1,43E-01	-2,04E+03	-2,86E-01	3,45E+00

Segundo a Tabela 6.14, 5 erros linguísticos dos 8 analisados foram melhores avaliados por modelos de coerência que não utilizam informação discursiva. Tal resultado mostra que, em princípio, os modelos não discursivos podem ser os mais adequados na identificação de erros da QL e que devem ser mais explorados para tal fim. Entretanto, uma conclusão mais exata dos resultados apresentados nessa seção só será possível quando mais exemplos de sumários com uma variabilidade maior de erros dominantes estiverem disponíveis, para que uma possível padronização na identificação dos erros da QL pelos modelos de coerência possa ser mais precisa.

6.4.3 Relacionamento entre Modelos de Coerência e Sumarizadores Multidocumento

O relacionamento entre os modelos de coerência e os sumarizadores automáticos multidocumento utilizados nesta tese foi estudado considerando dois aspectos: quantidade de erros linguísticos e a informatividade dos sumários gerados por cada sumarizador.

Baseado na quantidade de erros linguísticos presentes nos sumários de cada sumarizador (ver Tabela 6.1), um ranque de sumarizadores automáticos multidocumento foi criado: **RSumm > RC-4 > MTRST-MLAD > GistSumm**. Esse ranque indica que, em média, os sumários do RSumm têm menos erros linguísticos que afetam a coerência local do que os sumários do RC-4, ou seja, sumários de RSumm são mais coerentes do que os sumários de RC-4; o mesmo acontece tanto nos sumários de RC-4 em comparação aos sumários do MTRST-MLAD quanto

nos sumários do MTRST-MLAD em comparação com os sumários do GistSumm.

Já em relação à informatividade, o trabalho de Cardoso et al. (2015) usa uma medida clássica denominada ROUGE (Lin, 2004) para medir a informatividade dos sumários gerados pelos sumarizadores utilizados nesta tese, ou seja, GistSumm, MTRST-MLAD, RC-4 e RSumm. Baseado no trabalho de Cardoso et al. (2015), um ranque de informatividade dos sumarizadores foi formado, ou seja, **RC-4 > RSumm > MTRST-MLAD > GistSumm**. Tal ranque informa que, em média, os sumários do sumarizador RC-4 são mais informativos do que os sumários do RSumm; já os sumários do RSumm são mais informativos do que os sumários do MTRST-MLAD; e, por fim, os sumários do sumarizador MTRST-MLAD são mais informativos do que os sumários do sumarizador GistSumm.

O objetivo desse experimento é verificar se os modelos são realmente sensíveis aos erros da QL quando se comparam dois sumários automáticos de diferentes sumarizadores e se tais modelos podem ser úteis na verificação da informatividade do sumário. Para isso, o valor de ranque ou de coerência (dado por cada modelo de coerência local) de cada sumário automático foi utilizado de forma que pudessem ser comparados entre si e, assim, verificar qual sumário teve o maior valor de ranque ou de coerência. Dessa forma, contabiliza-se a porcentagem dos casos em que os sumários do sumarizador RC-4, por exemplo, tiveram os valores de ranque ou de coerência maiores do que os dos sumários do sumarizador RSumm, quando, por exemplo, o modelo SINTÁTICO+SALIÊNCIA- de Grade de Entidades gerou os respectivos valores de ranque ou de coerência. Todos os valores de ranque ou de coerência (gerados pelos modelos de coerência) de cada sumário automático de diferentes sumarizadores das coleções do córpus CSTNews foram comparados segundo os ranques dados pela quantidade de erros da QL e pela informatividade, com isso, pode-se verificar se os modelos conseguem replicar os mesmos ranques.

A Tabela 6.15 mostra a porcentagem de casos em que os valores de ranque ou de coerência dos sumários gerados pelo sumarizador GistSumm, por exemplo, são menores do que os dos sumários do MTRST-MLAD, ou seja, os sumários do GistSumm são mais incoerentes do que os do MTRST-MLAD (GistSumm <MTRST-MLAD), e os sumários do GistSumm possuem mais erros em relação aos sumários do sumarizador MTRST-MLAD. Tais valores de ranque ou de coerência foram produzidos por cada modelo de coerência adaptados da literatura, ou seja, modelos que não foram incrementados com informações discursivas. A comparação entre os valores de ranque ou de coerência dos sumários de cada sumarizador foi realizada em pares, seguindo a mesma metodologia utilizada na tarefa de ordenação de sentenças de Barzilay & Lapata.

Segundo a Tabela 6.15, os modelos baseados em Padrões Sintáticos foram os que tiveram as maiores porcentagens de casos em que o ranque de valores de coerência foram os mesmos dos gerados pela quantidade de erros. Por exemplo, o modelo de Padrões Sintáticos com d=1, com as 400 expressões sintáticas mais frequentes e Suavização igual a 0,1, tiveram 91,49% dos casos em que os valores de coerência dos sumários do GistSumm foram menores do que os dos sumários do MTRST-MLAD, 91,49% dos casos em que os valores de coerência dos sumários

Tabela 6.15: Porcentagem dos casos em que o modelo segue o mesmo ranque dado por Erros Linguísticos

MODELOS	Ranques Baseados na Quantidade de Erros Linguísticos		
	GistSumm <MTRST-MLAD (%)	MTRST-MLAD <RC-4 (%)	RC-4 <RSumm (%)
LSA	19,15	87,23	75,59
Grade Entidades (Sintático-Saliência-)	70,21	2,13	6,38
Grade Entidades (Sintático-Saliência+)	27,66	14,89	19,15
Grade Entidades (Sintático+Saliência-)	44,68	10,64	23,4
Grade Entidades (Sintático+Saliência+)	44,68	12,77	19,15
Grafo com Inf. Distância (P_U)	19,15	85,1	74,45
Grafo com Inf. Distância (P_W)	19,15	85,1	76,6
Grafo com Inf. Distância (P_{Acc})	21,28	72,34	68,09
Grafo sem Inf. Distância (P_U)	19,15	85,1	74,47
Grafo sem Inf. Distância (P_W)	19,15	85,1	76,6
Grafo sem Inf. Distância (P_{Acc})	21,28	72,34	68,09
Padrões Sintáticos-1- d1-Suavização_0,1	**91,49**	89,36	87,23
Padrões Sintáticos-25- d1-Suavização_0,1	**91,49**	**91,49**	91,49
Padrões Sintáticos-400- d1-Suavização_0,1	**91,49**	**91,49**	**95,74**
Padrões Sintáticos-1- d1-Suavização_0,01	**91,49**	89,36	89,36
Padrões Sintáticos-25- d1-Suavização_0,01	**91,49**	**91,49**	91,49
Padrões Sintáticos-400- d1-Suavização_0,01	**91,49**	89,36	**95,74**
Padrões Sintáticos-1- d1-Suavização_0,001	**91,49**	**91,49**	89,36
Padrões Sintáticos-25- d1-Suavização_0,001	**91,49**	**91,49**	91,49
Padrões Sintáticos-400- d1-Suavização_0,001	**91,49**	89,36	**95,74**
Padrões Sintáticos-1- d2-Suavização_0,1	85,1	76,6	74,47
Padrões Sintáticos-25- d2-Suavização_0,1	85,1	78,72	78,72
Padrões Sintáticos-400- d2-Suavização_0,1	85,1	78,72	80,85
Padrões Sintáticos-1- d2-Suavização_0,01	85,1	76,6	76,6
Padrões Sintáticos-25- d2-Suavização_0,01	85,1	78,72	78,72
Padrões Sintáticos-400- d2-Suavização_0,01	85,1	76,6	80,85
Padrões Sintáticos-1- d2-Suavização_0,001	85,1	78,72	78,72
Padrões Sintáticos-25- d2-Suavização_0,001	85,1	78,72	78,72
Padrões Sintáticos-400- d2-Suavização_0,001	85,1	76,6	80,85
Padrões Sint.-Productions-1-Suavização_0,1	**91,49**	**91,49**	87,23
Padrões Sint.-Productions-25-Suavização_0,1	**91,49**	**91,49**	93,62
Padrões Sint.-Productions-400-Suavização_0,1	**91,49**	**91,49**	93,62
Padrões Sint.-Productions-1-Suavização_0,01	**91,49**	89,36	89,36
Padrões Sint.-Productions-25-Suavização_0,01	**91,49**	**91,49**	93,62
Padrões Sint.-Productions-400-Suavização_0,01	**91,49**	89,36	**95,74**
Padrões Sint.-Productions-1-Suavização_0,001	**91,49**	89,36	89,36
Padrões Sint.-Productions-25-Suavização_0,001	**91,49**	**91,49**	93,62
Padrões Sint.-Productions-400-Suavização_0,001	**91,49**	89,36	**95,74**

do MTRST-MLAD foram menores do que os dos sumários do RC-4, e 95,74% dos casos em que os valores de coerência dos sumários do RC-4 foram menores do que os dos sumários do RSumm. Assim, dentre os modelos de coerência adaptados e que não usam informações discursivas, o modelo baseado em Padrões Sintáticos foi o mais sensível à quantidade de erros linguísticos presente nos sumários dos sumarizadores envolvidos na comparação. Tal comportamento não foi repetido na distinção dos sumários de referência (coerentes) dos incoerentes (versão permutada), isso pode ser explicado pela pouca ou nenhuma diferença dos possíveis padrões sintáticos dos sumários de referência e de suas versões permutadas entendidos pelo modelo.

Considerando a informatividade, os sumários do sumarizador MTRST-MLAD são mais informativos do que os sumários do GistSumm, isto é, GistSumm < MTRST-MLAD. Tal resultado pode ser explicado pela grande quantidade do erro Informação Redundante (RED) que afeta diretamente a informatividade (ver Tabela 6.11). Como o ranque GistSumm < MTRST-MLAD é o mesmo tanto para a quantidade de erros quanto para a informatividade, os modelos baseados em Padrões Sintáticos também tiveram um bom desempenho na questão da informatividade entre os sumários dos sumarizadores GistSumm e MTRST-MLAD.

Segundo a Tabela 6.16, o modelo baseado em Padrões Sintáticos que utiliza as 400 expressões sintáticas mais frequentes das sentenças dos sumários e com os diferentes valores de suavização também foram os modelos de coerência sem discurso que obtiveram os maiores va-

Tabela 6.16: Porcentagem dos casos em que o modelo segue o mesmo ranque dado pela Informatividade

MODELOS	Ranques Baseados na Informatividade		
	GistSumm <MTRST-MLAD (%)	MTRST-MLAD <RSumm (%)	RSumm <RC-4 (%)
LSA	19,15	76,59	**74,47**
Grade Entidades (Sintático-Saliência-)	70,21	6,38	42,55
Grade Entidades (Sintático-Saliência+)	27,66	19,15	48,94
Grade Entidades (Sintático+Saliência-)	44,68	23,4	36,17
Grade Entidades (Sintático+Saliência+)	44,68	19,15	53,19
Grafo com Inf. Distância (P_U)	19,15	74,47	63,83
Grafo com Inf. Distância (P_W)	19,15	76,6	68,09
Grafo com Inf. Distância (P_{Acc})	21,28	68,09	51,06
Grafo sem Inf. Distância (P_U)	19,15	74,47	63,83
Grafo sem Inf. Distância (P_W)	19,15	76,6	68,09
Grafo sem Inf. Distância (P_{Acc})	21,28	68,09	51,06
Padrões Sintáticos-1- d1-Suavização_0,1	**91,49**	87,23	53,19
Padrões Sintáticos-25- d1-Suavização_0,1	**91,49**	91,49	61,7
Padrões Sintáticos-400- d1-Suavização_0,1	**91,49**	**95,74**	55,32
Padrões Sintáticos-1- d1-Suavização_0,01	**91,49**	89,36	51,06
Padrões Sintáticos-25- d1-Suavização_0,01	**91,49**	91,49	61,7
Padrões Sintáticos-400- d1-Suavização_0,01	**91,49**	**95,74**	53,19
Padrões Sintáticos-1- d1-Suavização_0,001	**91,49**	89,36	57,45
Padrões Sintáticos-25- d1-Suavização_0,001	**91,49**	91,49	55,32
Padrões Sintáticos-400- d1-Suavização_0,001	**91,49**	**95,74**	53,19
Padrões Sintáticos-1- d2-Suavização_0,1	85,1	74,47	51,06
Padrões Sintáticos-25- d2-Suavização_0,1	85,1	78,72	63,83
Padrões Sintáticos-400- d2-Suavização_0,1	85,1	80,85	57,45
Padrões Sintáticos-1- d2-Suavização_0,01	85,1	76,6	53,19
Padrões Sintáticos-25- d2-Suavização_0,01	85,1	78,72	63,83
Padrões Sintáticos-400- d2-Suavização_0,01	85,1	80,85	53,19
Padrões Sintáticos-1- d2-Suavização_0,001	85,1	78,72	55,32
Padrões Sintáticos-25- d2-Suavização_0,001	85,1	78,72	59,57
Padrões Sintáticos-400- d2-Suavização_0,001	85,1	80,85	53,19
Padrões Sint.-Productions-1-Suavização_0,1	**91,49**	87,23	55,32
Padrões Sint.-Productions-25-Suavização_0,1	**91,49**	93,62	57,45
Padrões Sint.-Productions-400-Suavização_0,1	**91,49**	93,62	55,32
Padrões Sint.-Productions-1-Suavização_0,01	**91,49**	89,36	55,32
Padrões Sint.-Productions-25-Suavização_0,01	**91,49**	93,62	57,45
Padrões Sint.-Productions-400-Suavização_0,01	**91,49**	**95,74**	53,19
Padrões Sint.-Productions-1-Suavização_0,001	**91,49**	89,36	59,57
Padrões Sint.-Productions-25-Suavização_0,001	**91,49**	93,62	59,57
Padrões Sint.-Productions-400-Suavização_0,001	**91,49**	**95,74**	53,19

lores de porcentagem dos casos em que os valores de coerência dos sumários do sumarizador MTRST-MLAD foram menores do que os do sumarizador RSumm, repetindo a mesma relação dada pela informatividade dos sumários desses dois sumarizadores.

O ranque RSumm < RC-4 dado pela informatividade foi contemplado mais vezes pelo modelo de coerência LSA: 74,47% dos casos. Tal modelo usa a similaridade das sentenças para contabilizar a coerência local. Isso quer dizer que o sumarizador RC-4 utiliza mais sentenças adjacentes do mesmo texto fonte para compor um sumário multidocumento do que o sumarizador RSumm.

Em princípio, os dois modelos mais pobres em informação linguística dentre os modelos de coerência sem informação discursiva foram os que obtiveram os melhores resultados na comparação entre os sumários automáticos, levando em conta a quantidade de erros e a informatividade desses sumários.

Os modelos de coerência com informação discursiva também foram submetidos aos mesmos experimentos dos modelos de coerência sem informação discursiva. A Tabela 6.17 mostra a porcentagem dos casos em que os valores de ranque ou de coerência dos sumários de cada sumarizador, os quais foram gerados pelos modelos de coerência discursivos, são comparados e que seguem com o ranque desses sumarizadores dado pela quantidade de erros linguísticos.

De acordo com a Tabela 6.17, os modelos SINTÁTICO+SALIÊNCIA- Booleana CST e SINTÁTICO+SALIÊNCIA- Booleana CST e RST foram os que tiveram a maior porcentagem

Tabela 6.17: Porcentagem dos casos em que o modelo discursivo segue o mesmo ranque dado por Erros Linguísticos

MODELOS	Ranques Baseados na Quantidade de Erros Linguísticos		
	GistSumm <MTRST-MLAD (%)	MTRST-MLAD <RC-4 (%)	RC-4 <RSumm (%)
SINTÁTICO+SALIÊNCIA- com CST	21,28	80,85	85,1
SINTÁTICO+SALIÊNCIA- CST e RST	42,55	55,32	63,83
SINTÁTICO+SALIÊNCIA- com RST	27,66	57,45	63,83
SINTÁTICO-SALIÊNCIA- com CST	23,4	78,72	87,23
SINTÁTICO-SALIÊNCIA- com CST e RST	40,42	53,19	55,32
SINTÁTICO-SALIÊNCIA-RST+	34,04	51,06	57,45
Grafo com Discurso com Inf. Distância (P_U)	6,38	**95,74**	87,23
Grafo com Discurso com Inf. Distância (P_W)	8,51	93,62	**91,49**
Grafo com Discurso sem Inf. Distância (P_U)	2,13	89,36	87,23
Grafo com Discurso sem Inf. Distância (P_W)	2,13	93,62	87,23
Entidades com RST Local	55,32	31,91	51,06
Termo com RST	42,55	61,7	72,34
SINTÁTICO+SALIÊNCIA- Booleana CST	**85,11**	91,49	14,89
SINTÁTICO+SALIÊNCIA- Booleana CST e RST	**85,11**	91,49	14,89
Relações Discursivas	17,02	78,72	78,72

dos casos em que os valores de ranque dos sumários do GistSumm foram menores do que os dos sumários do sumarizador MTRST-MLAD, seguindo assim, o ranque GistSumm < MTRST-MLAD. Já para os ranques MTRST-MLAD < RC-4 e RC-4 < RSumm, os modelos baseados em Grafo com Discurso com Inf. Distância (P_U) e Grafo com Discurso com Inf. Distância (P_W), respectivamente, foram os modelos que tiveram os maiores valores de porcentagem dos casos em que os valores de coerência dos sumários pertencentes aos respectivos sumarizadores acompanharam os respectivos ranques dados pela quantidade de erros.

Segundo os dados das Tabelas 6.15 e 6.17, os modelos que não fazem uso de aprendizado de máquina foram mais sensíveis aos erros quando os sumários dos sumarizadores foram comparados entre si. Isso confirma a necessidade de mais sumários para verificar a capacidade dos modelos de coerência na identificação dos erros da QL, pois os modelos de coerência que tiveram bons resultados no experimento de Ranques Baseados na Quantidade de Erros Linguístico dessa seção não foram os mesmos que possivelmente podem identificar os erros linguísticos mostrados na Seção 6.4.3, salvo a exceção do modelo baseado em Padrões Sintáticos que utilizou todas as expressões sintáticas do nível 2 da árvore sintática das sentenças dos sumários e com o parâmetro de suavização igual a 0,001, o qual teve bons resultados em ambos os experimentos, apesar de que apenas em 1 sumário ocorreu o erro SENT_INC de maneira mais frequente (ver Tabela 6.10).

Para a informatividade, os modelos de coerência com discurso SINTÁTICO+SALIÊNCIA- Booleana CST e SINTÁTICO+SALIÊNCIA- Booleana CST e RST foram os que tiveram as maiores porcentagens de casos em que os valores de ranque dos sumarizadores comparados entre si seguiram os mesmos ranques GistSumm < MTRST-MLAD e RSumm < RC-4 dados pela informatividade (ver Tabela 6.18). Já o modelo baseado em Grafo com Discurso com Inf. Distância (P_W) foi o que melhor relacionou os sumários criados pelos sumarizadores MTRST-MLAD e RSumm, por meio da comparação entre os seus respectivos valores de coerência, segundo o ranque MTRST-MLAD < RSumm.

Observando os resultados mostrados nas Tabelas 6.16 e 6.18, há modelos de coerência que podem ser usados como um mecanismo adicional na avaliação da informatividade de um sumário. Tal afirmação necessita ser melhor comprovada com novos sumários gerados pelos mesmos

Tabela 6.18: Porcentagem dos casos em que o modelo discursivo segue o mesmo ranque dado pela Informatividade

MODELOS	Ranques Baseados na Informatividade		
	GistSumm <MTRST-MLAD (%)	**MTRST-MLAD <RSumm (%)**	**RSumm <RC-4 (%)**
SINTÁTICO+SALIÊNCIA- com CST	21,28	85,1	72,34
SINTÁTICO+SALIÊNCIA- CST e RST	42,55	63,83	68,09
SINTÁTICO+SALIÊNCIA- com RST	27,66	63,83	80,85
SINTÁTICO-SALIÊNCIA- com CST	23,4	87,23	74,47
SINTÁTICO-SALIÊNCIA- com CST e RST	40,42	55,32	72,34
SINTÁTICO-SALIÊNCIA-RST+	34,04	57,45	82,98
Grafo com Discurso com Inf. Distância (P_U)	6,38	87,23	63,83
Grafo com Discurso com Inf. Distância (P_W)	8,51	**91,49**	68,09
Grafo com Discurso sem Inf. Distância (P_U)	2,13	87,23	63,83
Grafo com Discurso sem Inf. Distância (P_W)	2,13	87,23	70,21
Entidades com RST Local	55,32	51,06	36,17
Termo com RST	42,55	72,34	36,17
SINTÁTICO+SALIÊNCIA- Booleana CST	**85,11**	14,89	**93,62**
SINTÁTICO+SALIÊNCIA- Booleana CST e RST	**85,11**	14,89	**93,62**
Relações Discursivas	17,02	78,72	48,94

sumarizadores utilizados nesta tese e possivelmente sumários de novos sumarizadores. Entretanto, há resultados que se mostraram promissores para um possível relacionamento entre os modelos de coerência e a informatividade dada pelos sumarizadores automáticos multidocumento que merecem serem mais aprofundados.

Nesse capítulo, os erros de QL foram mostrados e definidos. Além disso, experimentos foram realizados com o intuito de verificar alguns possíveis relacionamentos envolvendo os modelos de coerência, os sumarizadores automáticos multidocumento, os erros de QL e a informatividade dos sumários. Com tais relacionamentos, verificou-se que os sumarizadores necessitam tratar os erros de QL, pois não há ou há pouca preocupação, por parte dos sumarizadores, em tratar tais erros; alguns modelos de coerência se mostraram propícios em avaliar determinados erros de QL, e que os mesmos poderão ser úteis a sumarizadores no processo de geração de sumários, quando tais modelos indicarão o sumário mais coerente dentre os possíveis gerados; e que algus modelos de coerência podem colaborar, como um possível recurso, na avaliação da informatividade dos sumários automáticos multidocumento.

Capítulo

7

Considerações Finais

Como foi mostrado nesta tese, os sumarizadores não tratam adequadamente os erros que afetam a Qualidade Linguística (QL) e isso pode afetar a coerência de seus sumários. Esse não tratamento dos erros da QL pode ser explicado pela própria dificuldade em tratar erros como: i) o tratamento dos fenômenos multidocumento de redundância, complementaridade e contradição de informações, ii) uniformização de estilos de escrita, iii) tratamento de expressões referenciais, iv) manutenção de focos e perspectivas diferentes nos textos e v) ordenação temporal das informações no sumário. Assim, este trabalho investigou e desenvolveu modelos que são capazes de avaliar a coerência local em sumários multidocumento para o Português do Brasil. Não há conhecimento de outro trabalho que se propôs a utilizar informações discursivas dos modelos RST (Mann & Thompson, 1987) e CST (Radev, 2000) na avaliação da coerência em sumários multidocumento. Além disso, um estudo sobre os erros que afetam a QL dos sumários multidocumento também foi realizado, possibilitando, assim, uma análise inicial dos relacionamentos entre erros da QL e sumarizadores automáticos multidocumento, modelos de coerência e erros da QL e, por fim, modelos de coerência e sumarizadores automáticos multidocumento, algo também inédito.

A tese deste trabalho é que conhecimento discursivo pode ser usado de forma satisfatória na avaliação da coerência local em sumários multidocumento, tanto no enriquecimento de modelos já existentes quanto na criação de modelos puramente discursivos. Além disso, as hipóteses consideradas neste trabalho foram: i) as informações das teorias discursivas escolhidas são úteis para a avaliação de coerência local, ii) item Os sumários coerentes possuem uma organização textual padrão baseado em relações discursivas que os distinguem dos sumários incoerentes, iii) A utilização de técnicas de Aprendizado de Máquina proporcionará maior eficiência se comparada a métodos heurísticos e iv) Os modelos de coerência local tem poder variado de discriminação de certos tipos de erros linguísticos.

Para comprovar a tese deste trabalho, aproximadamente 60 modelos entre os adaptados da

literatura, enriquecidos com informação discursiva, novos modelos e variações foram desenvolvidos. Pelos resultados apresentados nos Capítulos 4 e 5, o principal modelo da área, o de Grade de Entidades de Barzilay & Lapata (2008), teve um ganho máximo de 52% na acurácia com o uso de informações discursivas. Outro modelo importante na área é o modelo Baseado em Grafo, o qual teve um ganho de 39,05% a 53,15% com a utilização de informações discursivas. Já outros modelos da literatura como o de Lin et al. (2011) e Feng et al. (2014), que originalmente utilizam informações discursivas e foram adaptados nesta tese, não se mostraram tão competitivos na avaliação da coerência no cenário da sumarização multidocumento.

O modelo que obteve a maior acurácia foi o modelo de Relações Discursivas, modelo novo que foi proposto nesta tese. Tal modelo obteve 92,69% de acurácia na avaliação da coerência dos sumários multidocumento. Todos os outros modelos adaptados e que não utilizam informação discursiva tiveram desempenho abaixo dos modelos que utilizam informação discursiva, lembrando que tais resultados seguiram os mesmos procedimentos difundidos na literatura. Assim, as teorias discursivas (CST e RST) puderam ser usadas de forma satisfatória na distinção de sumários multidocumentos coerentes dos incoerentes, tanto no enriquecimento de modelos já existentes quanto na criação de modelos puramente discursivos, ou seja, a tese desse trabalho foi validade e comprovada.

A partir da comprovação da tese, quase todas as hipóteses deste trabalho foram validadas. Entretanto, a hipótese de que os modelos de coerência local podem avaliar certos tipos de erros da QL necessita de um cuidado maior, devido a baixa quantidade de dados relacionados aos erros da QL utilizada neste trabalho. No entanto, pelos resultados preliminares alcançados há um bom indício de que alguns modelos podem ser utilizados na avaliação de erros específicos da QL. Outra questão importante foi o relacionamento entre os modelos de coerência local e a informatividade dos sumários automáticos multidocumento que se mostrou bem promissor.

7.1 Contribuições

7.1.1 Teóricas

Uma contribuição teórica importante são os próprios modelos de coerência desenvolvidos nesta tese, os quais são voltados para a sumarização multidocumento. Tais modelos podem ser utilizados como uma etapa na geração de sumários dos sumarizadores automáticos, ou seja, um sumarizador pode criar sumários que serão analisados por um dos modelos de coerência que irá gerar um valor de ranque ou de coerência para os sumários: o sumário que possuir o maior desses valores será considerado o mais coerente, e este será o sumário final do sumarizador. Tal uso foi realizado em experimentos do trabalho de Castro Jorge (2015). Os modelos também são independentes de língua, ou seja, podem ser utilizados em qualquer idioma deste que os recursos (*parser* sintático e córpus multidocumento anotado com relações CST e RST ou um *parser* discursivo), necessários para o funcionamento dos modelos, tenham a sua versão para o idioma alvo.

Outra contribuição é a modelagem das informações discursivas CST e RST, principalmente

no enriquecimentos dos modelos da literatura e nos novos modelos puramente discursivos. O ganho de acurácia evidenciado nos resultados dos modelos de coerência que utilizam as informações discursivas CST e RST é uma contribuição significativa.

Os estudos iniciais sobre os relacionamentos entre os erros da QL e sumarizadores automáticos multidocumento, entre os erros da QL e modelos de coerência, e entre os modelos de coerência e sumarizadores automáticos multidocumento abrirão novas pespectivas de pesquisa.

7.1.2 Práticas

Uma das contribuições práticas deste trabalho diz respeito a criação de novos sumários multidocumento por humanos, que, no caso deste trabalho, reuniu pesquisadores da área de Ciências da Computação e de Linguística para tal procedimento. A criação de novos sumários foi importante para este trabalho, pois, até então, o córpus CSTNews contava com apenas 1 sumário humano multidocumento, totalizando 50 sumários. Essa quantidade poderia prejudicar o desempenho de alguns modelos de coerência, principalmente os que fazem uso de Aprendizado de Máquina.

O levantamento dos erros da Qualidade Linguística, a formação de um córpus de sumários automáticos e a metodologia da anotação dos erros da QL em tais sumários são contribuições práticas ricas que serão de grande valia para trabalhos futuros que necessitam de tais recursos.

7.2 Limitações

Apesar das contribuições oferecidas por esta tese, algumas limitações também foram identificadas, como a necessidade de um córpus anotado, pois, apesar de existirem *parsers* discursivos tanto para CST e RST, as anotações automáticas ainda estão sujeitas a erros. Além disso, o tempo de execução dos *parsers* é longo. Para não comprometer a investigação, ainda se utiliza córpus anotado manualmente, o que resulta em um processo trabalhoso e subjetivo.

Outra limitação foi a quantidade de sumários automáticos de exemplos com erros da QL para realizar uma análise mais precisa sobre os relacionamentos propostos no Capítulo 6, pois a quantidade de sumários com erros da QL mais frequentes utilizada nesta tese foi insuficiente para se chegar a uma conclusão, apesar dos bons indícios verificados.

Os modelos de coerência apenas distinguem sumários coerentes dos incoerentes e não avaliam em uma escala a coerência de um sumário. Além disso, a distinção entre sumários coerentes e sumários com poucos problemas que afetam a coerência (quase coerentes) pode não ser precisa, principalmente nos modelos que não utilizam Aprendizado de Máquina, os quais não demonstraram ser tão sensíveis a erros de QL da categoria Menções de Entidades.

7.3 Trabalhos Futuros

Ao longo deste trabalho, foram tomadas decisões que determinaram um caminho a seguir. Outras decisões de projeto poderiam ter sido tomadas e certamente outros resultados seriam encontrados e novas perspectivas se abririam. Por conta disso, serão elencadas algumas sugestões para trabalhos a desenvolver que visam não só complementar o trabalho realizado, como também abrir novos percursos de investigação.

O córpus utilizado nesta tese já estava anotado com relações CST e RST. Entretanto, *parsers* discursivos disponíveis para o português do Brasil, como o DiZer (Pardo et al., 2004; Pardo & Nunes, 2006; Maziero & Pardo, 2008) de relações RST, e o CSTParser (Maziero & Pardo, 2012), de relações CST, poderiam re-anotar o córpus CSTNews para que novos experimentos com os modelos de coerência pudessem ser realizados. Com isso, comparar e analisar os resultados provenientes dos modelos de coerência da utilização do córpus anotado manualmente e automaticamente em função do desempenho desses modelos poderá contemplar a automatização de todo o processo do modelo de coerência.

Os modelos discursivos baseados em Grade de Entidades, desenvolvidos nesta tese, não utilizaram da informação de saliência (SALIÊNCIA-), e mesmo assim , tais modelos apresentaram resultados superiores em comparação aos modelos adaptados da literatura. Entretanto, a saliência no modelo de Grade de Entidades original produziu resultados interessantes juntamente com a informação sintática. Desse forma, a informação de saliência poderia ser explorada com as relações CST e RST, para aumentar o poder preditivo dos modelos discursivos.

Utilizar outros conhecimentos discursivos, como os Aspectos Informativos (Dias et al., 2012; Rassi et al., 2013; DiFelippo et al., 2014), na criação de novos modelos de coerência local.

Os trabalhos baseados em aprendizado de máquina nesta tese utilizaram o pacote de aprendizado chamado SVM^{light} (Joachims, 2002) com a opção de ranque, já que todos os trabalhos da literatura utilizam esse mesmo pacote. Outros algoritmos de aprendizado de máquina de ranqueamento não foram utilizados em experimentos nos moldes dos realizados neste trabalho. Desse forma, novos experimentos poderiam ser feitos com o intuito de verificar o desempenho dos modelos de coerência com os novos algoritmos de aprendizado de máquina. Além disso, um estudo sobre a viabilidade da utilização dos algoritmos de aprendizado de máquina voltados para a classificação nos modelos de coerência local seria uma frente interessante a ser considerada.

Os modelos de coerência apresentados nesta tese definem uma relação de ordem entre um par formado por um sumário de referência e um sumário alvo, de tal modo que o sumário possuidor de maior qualidade (quanto à coerência) é o preferido. É importante ressaltar que os modelos propostos neste trabalho não podem ser aplicados a fim de classificar a coerência de um sumário em função de um conjunto de categorias pré-definidas. Por exemplo, uma estratégia possível de estudo seria modelar o problema de inferência de coerência como um problema de classificação multiclasse, no qual se deseja determinar se um texto é "muito coerente", "coerente", "razoavelmente coerente" ou "incoerente". Tal cenário se apresenta como

uma oportunidade de pesquisa tendo em vista aplicações (em tempo real) de análise automática de coerência em textos .

Um estudo de córpus para os erros de Qualidade Linguística com o intuito de aumentar a quantidade de exemplos dos erros para verificar se os modelos de coerência podem realmente ser utilizados na identificação de erros que afetam a Qualidade Linguística dos Sumários.

Por fim, um estudo do possível relacionamento entre os modelos de coerência local e a informatividade dos sumários multidocumento poderia ser realizado com uma profundidade maior. Para isso, é necessário mais sumários multidocumento e uma variedade maior de sumarizadores multidocumento.

7.4 Publicações Geradas

Nesta seção são apresentados os resultados do doutorado em termos de publicações. A seguir estão listados as publicações que direta ou indiretamente estão relacionados a esta tese.

- Dias, M.S. and Pardo, T.A.S. (2015). Enriching entity grids and graphs with discourse relations: the impact in measuring local coherence in multi-document summaries. In the Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology - STIL, pp. 151-160. November 4-7. Natal/Brazil.

- Dias, M.S. and Pardo, T.A.S. (2015). A Discursive Grid Approach to Model Local Coherence in Multi- document Summaries. In the Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue - SIGDial, pp. 60-67. September 2-4. Prague/Czech Republic.

- Sobrevilla Cabezudo, M.A.; Maziero, E.G.; Souza, J.W.C.; Dias, M.S.; Cardoso, P.C.F.; Balage Filho, P.P.; Agostini, V.; Nóbrega, F.A.A.; Barros, C.D.; Di Felippo, A.; Pardo, T.A.S. (2015). Anotação de Sentidos de Verbos em Textos Jornalísticos do Corpus CST-News. Revista de Estudos da Linguagem - RELIN, Vol. 23, N. 3, pp. 797-832.

- Sobrevilla Cabezudo, M.A.; Maziero, E.G.; Souza, J.W.C.; Dias, M.S.; Cardoso, P.C.F.; Balage Filho, P.P.; Agostini, V.; Nóbrega, F.A.A.; Barros, C.D.; Di Felippo, A.; Pardo, T.A.S. (2014). Anotação de Sentidos de Verbos em Notícias Jornalísticas em Português do Brasil. In the Proceedings of the XII Encontro de Linguística de Corpus - ELC. November 6-7. Uberlândia-MB/Brazil.

- Dias, M.S.; Castro Jorge, M.L.R.; Pardo, T.A.S. (2014). Building a Language Model for Local Coherence in Multi-document Summaries using a Discourse-enriched Entity-based Model. In the Proceedings of the Brazilian Conference on Intelligent Systems - BRACIS, pp. 44-49. October 18-23. São Carlos-SP/Brazil.

- Dias, M.S.; Feltrim, V.D.; Pardo, T.A.S. (2014). Using Rhetorical Structure Theory and Entity Grids to Automatically Evaluate Local Coherence in Texts. In the Proceedings of

the 11st International Conference on Computational Processing of Portuguese - PROPOR (LNAI 8775), pp. 232-243. October 6-9. São Carlos- SP/Brazil.

- Dias, M.S.; Bokan Garay, A.Y.; Chuman, C.; Barros, C.D.; Maziero, E.G.; Nobrega, F.A.A.; Souza, J.W.C.; Sobrevilla Cabezudo, M.A.; Delege, M.; Castro Jorge, M.L.R.; Silva, N.L.; Cardoso, P.C.F.; Balage Filho, P.P.; Lopez Condori, R.E.; Marcasso, V.; Di Felippo, A.; Nunes, M.G.V.; Pardo, T.A.S. (2014). Enriquecendo o Corpus CSTNews - a Criação de Novos Sumários Multidocumento. In the (on-line) Proceedings of the I Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish - ToRPorEsp, pp. 1-8. October 9. São Carlos-SP/Brazil.

- Di Felippo, A.; Rino, L.H.M.; Pardo, T.A.S.; Cardoso, P.C.F.; Seno, E.R.M.; Balage Filho, P.P.; Rassi, A.P.; Dias, M.S.; Castro Jorge, M.L.R.; Maziero, E.G.; Zacarias, A.C.I.; Souza, J.W.C.; Camargo, R.T.; Agostini, V. (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, pp. 171-192. Cambridge Scholars Publishing.

- Rassi, A.P.; Zacarias, A.C.I.; Maziero, E.G.; Souza, J.W.C.; Dias, M.S.; Castro Jorge, M.L.R.; Cardoso, P.C.F.; Balage Filho, P.P.; Camargo, R.T.; Agostini, V.; Di Felippo, A.; Seno, E.R.M.; Rino, L.H.M.; Pardo, T.A.S. (2013). Anotação de Aspectos Textuais em Sumários do Córpus CSTNews. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 394. NILC- TR-13-01. São Carlos-SP, Outubro, 59p.

- Cardoso, P.C.F.; Rassi, A.P.; Maziero, E.G.; Nóbrega, F.A.A.; Souza, J.W.C.; Dias, M.S.; Castro Jorge, M.L.R.; Balage Filho, P.P.; Camargo, R.T.; Agostini, V.; Di Felippo, A.; Rino, L.H.M.; Pardo, T.A.S. (2012). Anotação de Subtópicos do Córpus Multidocumento CSTNews. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 389. NILC-TR-12-07. São Carlos- SP, Junho, 18p.

- Dias, M. S.; Rassi, A. P. ; Rino, L. H. M. (2012). Preliminary Aspects Distribution in Political Texts. In: XI Encontro de Linguística de Corpus (ELC), São Carlos - SP. Anais do XI Encontro de Linguística de Córpus (ELC).

Além das publicações listadas acimas, há um artigo sobre a tarefa de anotação dos erros da Qualidade Linguística em fase final de produção para a submissão a um periódico internacional e um artigo final, relatando os principais avanços deste doutorado, que está sendo preparado para submisão a um periódico internacional.

Referências Bibliográficas

Aktas, R. N. & V. Cortes (2008). Shell nouns as cohesive devices in published and esl student writing. *Journal of English for Academic Purposes 7*(1), 3 – 14.

Aleixo, P. & T. A. S. Pardo (2008). Cstnews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structuretheory). Technical report, NILC - ICMC - USP. 12p.

Barzilay, R. & M. Lapata (2005). Modeling local coherence: an entity-based approach. Em *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Stroudsburg, PA, USA, pp. 141–148. Association for Computational Linguistics.

Barzilay, R. & M. Lapata (2008). Modeling local coherence: An entity-based approach. *Comput. Linguist. 34*(1), 1–34.

Bick, E. (2000). *The Parsing System PALAVRAS - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Tese de Doutorado, Department of Linguistics, University of Arhus, DK.

Brants, S. & S. Hansen (2002). Developments in the tiger annotation scheme and their realization in the corpus. Em *Third Conference on Language Resources and Evaluation LREC-02. Las Palmas de Gran Canaria*.

Brennan, S. E., M. W. Friedman, & C. J. Pollard (1987). A centering approach to pronouns. Em *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, ACL '87, Stroudsburg, PA, USA, pp. 155–162. Association for Computational Linguistics.

Burstein, J., J. Tetreault, & S. Andreyev (2010). Using entity-based features to model coherence in student essays. Em *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, Stroudsburg, PA, USA, pp. 681–684. Association for Computational Linguistics.

Cardoso, P., M. Castro Jorge, & T. Pardo (2015). Exploring the rhetorical structure theory for multi-document summarization. Em *Proceedings of the 5th Workshop RST and Discourse Studies*, pp. 1 – 10.

Cardoso, P., E. Maziero, M. Jorge, E. Seno, A. di Felippo, L. Rino, M. Nunes, & T. Pardo (2011). Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. Em *3rd RST Brazilian Meeting. 88-105 p.*

Cardoso, P. C. F. (2014). *Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - ICMC/USP.

Carletta, J. (1996, June). Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist. 22*(2), 249–254.

Carlson, L., D. Marcu, & M. E. Okurowski (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. Em *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, Stroudsburg, PA, USA, pp. 1–10. Association for Computational Linguistics.

Castro Jorge, M. (2010). Sumarização automática multidocumento: seleção de conteúdo com base no modelo cst (cross-document structure theory). Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP.

Castro Jorge, M. & T. Pardo (2012). Multi-document summarization: Content selection based on cst model (cross-document structure theory). Em *PROPOR 2012 PhD and MSc/MA Dissertation Contest*, Coimbra, Portugal, pp. 1–8.

Castro Jorge, M. L. R. (2015). *Modelagem gerativa para sumarização automática multidocumento*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - ICMC/USP.

Chawla, N. V., N. V. Chawla, K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer (2002). Smote: Synthetic minority over-sampling technique. *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH 16*, 321–357.

Chen, L. & R. Ng (2004). On the marriage of lp-norms and edit distance. Em *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pp. 792–803. VLDB Endowment.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. Em *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, EACL '97, Stroudsburg, PA, USA, pp. 16–23. Association for Computational Linguistics.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, & P. Kuksa (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res. 12*, 2493–2537.

Collovoni, S., T. Carbonel, J. T. Fuchs, J. C. Coelho, L. Rino, & R. Vieira (2007). Summit: Um corpus anotado com informações discursivas visando à sumarização automática. Em *5o Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007)*.

Cortes, C. & V. Vapnik (1995a). Support-vector networks. *Mach. Learn. 20*(3), 273–297.

Cortes, C. & V. N. Vapnik (1995b). Support-vector networks. Em *Machine Learning*. Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Cunha, I., E. SanJuan, J. Torres-Moreno, M. Lloberes, & I. Castellón (2010). Discourse segmentation for spanish based on shallow parsing. Em *Advances in Artificial Intelligence - 9th Mexican International Conference on Artificial Intelligence, MICAI 2010, Pachuca, Mexico, November 8-13, 2010, Proceedings, Part I*, pp. 13–23.

da Cunha Fanego, I. (2008). *Hacia un modelo linguÌiÌstico de resumen automaÌtico de artiÌculos meÌdicos en espanÌol*. Tese de Doutorado, Universitat Pompeu Fabra.

de Beaugrande, R.-A. & W. Dressler (1981). *Introduction to Textlinguisticas*. Longamn.

Demăr, J. (2008). On the appropriateness of statistical tests in machine learning.

Dias, M. S., A. P. Rassi, & L. H. M. Rino (2012). Preliminary aspects distribution in political texts. Em *XI Encontro de Linguística de Córpus (ELC)*.

DiFelippo, A., L. Rino, T. Pardo, P. Cardoso, E. Seno, P. Balage Filho, A. Rassi, M. Dias, M. Castro Jorge, E. Maziero, A. Zacarias, J. Souza, R. Camargo, & V. Agostini (2014). *Corpus Annotation of Textual Aspects in Multi-document Summaries*, pp. 171–192. Cambridge Scholars Publishing.

Dijk, T. V. & W. Kintsch (1983). *Strategics in Discourse Comprehension*. New York, N.Y.: Academic Press.

Eisner, M. & E. Charniak (2011). Extending the entity grid with entity-specific features. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, Stroudsburg, PA, USA, pp. 125–129. Association for Computational Linguistics.

Feltrim, V. D., S. Teufel, M. G. V. Nunes, & S. M. Aluísio (2006). Argumentative zoning applied to criquing novices scientific abstracts. Em J. W. James G. Shanahan, Yan Qu (Ed.), *Computing Attitude and Affect in Text: Theory and Applications*, Volume 20. Springer Netherlands.

Feng, V. W. & G. Hirst (2012). Extending the entity-based coherence model with multiple ranks. Em *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, Stroudsburg, PA, USA, pp. 315–324. Association for Computational Linguistics.

Feng, V. W., Z. Lin, & G. Hirst (2014). The impact of deep hierarchical discourse structures in the evaluation of text coherence. Em *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 940–949.

Filippova, K. & M. Strube (2007). Extending the entity-grid coherence model to semantically related entities. Em *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, Stroudsburg, PA, USA, pp. 139–142. Association for Computational Linguistics.

Foltz, P. W., P. W. Foltz, W. Kintsch, & T. K. Landauer (1998). The measurement of textual coherence with latent semantic analysis.

Freitas, A. R. P. (2013). Análise automática de coerência usando o modelo grade de entidades para o português. Dissertação de Mestrado, Universidade Estadual de Maringá - Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação.

Friedrich, A., M. Valeeva, & A. Palmer (2014). Lqvsumm: A corpus of linguistic quality violations in multi-document summarization. Em N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Givon, T. (1987). Beyond foreground and background. Em R. S. Tomlin (Ed.), *Coherence and Grounding in Discourse*. Benjamins, Amsterdam/Philadelphia.

Golub, G. & C. E. Reinsch (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik 14*(5), 403–420.

Gonçalves, P. N. (2008). Correfsum: Revisão de coesão referencial em sumários extrativos. Dissertação de Mestrado, Mestrado Computação Aplicada - Universidade do Vale do Rio dos Sinos.

Grosz, B. J., S. Weinstein, & A. K. Joshi (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics 21*, 203–225.

Guinaudeau, C. & M. Strube (2013). Graph-based local coherence modeling. Em *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Volume 1, Sofia, Bulgaria, pp. 93–103.

Halliday, M. & R. Hasan (1976). *Cohesion in English*. Longman.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist. 23*(1), 33–64.

Heike Telljohann, E. H. & S. Kübler (2003). Stylebook for the tübingen treebank of written german (tüba-d/z). Technical report, Universität Tübingen.

Hovy, E. H. & J. Lavid (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*.

Iida, R. & M. Poesio (2011). A cross-lingual ilp solution to zero anaphora resolution. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Stroudsburg, PA, USA, pp. 804–813. Association for Computational Linguistics.

Iida, R. & T. Tokunaga (2012, December). A metric for evaluating discourse coherence based on coreference resolution. Em *Proceedings of COLING 2012: Posters*, Mumbai, India, pp. 483–494. The COLING 2012 Organizing Committee.

Iruskieta, M., I. Cunha, & M. Taboada (2014). A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation 49*(2), 263–309.

Joachims, T. (1999). Making large-scale support vector machine learning practical. Em B. Schlkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods*, Capítulo: Making large-scale support vector machine learning practical, pp. 169–184. MIT Press. Cambridge, MA, USA.

Joachims, T. (2002). Optimizing search engines using clickthrough data. Em *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, New York, NY, USA, pp. 133–142. ACM.

Jones, K. S. (1993). Discourse modelling for automatic summarisation. Technical report, University of Cambridge.

Kaspersson, T., C. Smith, H. Danielsson, & A. Jönsson (2012, may). This also affects the context - errors in extraction based summaries. Em N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Koch, I. & L. C. Travaglia (1989). *Texto e coerência*. Editora Cortez.

Koch, I. G. V. (1998). *A coesão textual – Mecanismos de Constituição Textual, A organização do Texto, Fenômenos de Linguagem* (10ª Edição). Linguística Contexto – Repensando a Língua Portuguesa.

Koch, I. G. V. & L. C. Travaglia (2002). *A coerência textual*. Editora Contexto.

Konstas, I. & M. Lapata (2012). Concept-to-text generation via discriminative reranking. Em *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, Stroudsburg, PA, USA, pp. 369–378. Association for Computational Linguistics.

Landauer, T., P. Foltz, & D. Laham (1998). An introduction to latent semantic analysis. *Discourse processes 25*, 259–284.

Landauer, T. K., D. Laham, B. Rehder, & M. E. Schreiner (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans.

Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's tau. *Comput. Linguist. 32*(4), 471–484.

Li, J. & E. H. Hovy (2014). A model of coherence based on distributed sentence representation. Em A. Moschitti, B. Pang, & W. Daelemans (Eds.), *EMNLP*, pp. 2039–2048. ACL.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Em S. S. Marie-Francine Moens (Ed.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics.

Lin, C.-Y. & E. Hovy (2003). Automatic evaluation of summaries using n-gram cooccurrence statistics. Em *Language Technology Conference*.

Lin, R., M. Yang, S. Liu, S. Li, & T. Zhao (2015). A maximum entropy approach to discourse coherence modeling. Em J. Li, H. Ji, D. Zhao, & Y. Feng (Eds.), *NLPCC*, Volume 9362 of *Lecture Notes in Computer Science*, pp. 3–11. Springer.

Lin, Z., H. T. Ng, & M.-Y. Kan (2011). Automatically evaluating text coherence using discourse relations. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Stroudsburg, PA, USA, pp. 997–1006. Association for Computational Linguistics.

Louis, A. & A. Nenkova (2012a). A coherence model based on syntactic patterns. Em *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Stroudsburg, PA, USA, pp. 1157–1168. Association for Computational Linguistics.

Louis, A. & A. Nenkova (2012b). A coherence model based on syntactic patterns. Em *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Stroudsburg, PA, USA, pp. 1157–1168. Association for Computational Linguistics.

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

Mani, I. & M. Maybury (1999). *Advances in automatic text summarization*. The MIT Press.

Mann, W. C. & S. A. Thompson (1987). Rhetorical structure theory: A theory of text organization. Technical report, ISI/RS-87-190.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Tese de Doutorado, Department of Computer Science, University of Toronto.

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

Marcus, M. P., M. A. Marcinkiewicz, & B. Santorini (1993). Building a large annotated corpus of english: The penn treebank. *Comput. Linguist. 19*(2), 313–330.

Marcuschi, L. A. (1983). *Lingüística de texto: que é e como se faz?* Editora Universitária da UFPE.

Martins, C., T. Pardo, A. Espina, & L. Rino (2001). Introdução à sumarização automática. Technical report, Departamento de Computação, Universidade Federal de São Carlos. 38 p.

Maziero, E., M. Jorge, & T. A. S. Pardo (2010). Identifying multidocument relations. Em *Proceedings of the 7th International Workshop on Natural Language Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science - NLPCS,*, pp. 60–90.

Maziero, E. & T. Pardo (2008). Aprimoramento e avaliação do analisador discursivo automático dizer para o português do brasil. Em *16o Simpósio Internacional de Iniciação Científica da Universidade de São Paulo - SIICUSP*, pp. 1.

Maziero, E. & T. Pardo (2012). Cstparser - a multi-document discourse parser. Em *Proceedings of the PROPOR 2012 Demonstrations*, Coimbra, Portugal, pp. 17–20.

Maziero, E. & T. A. S. Pardo (2009). Automatização de um método de avaliação de estruturas retóricas. Em *Proceedings of the RST Brazilian Meeting. 9 p.*

Maziero, E. G., G. Hirst, & T. A. S. Pardo (2015). Semi-supervised never-ending learning in rhetorical relation identification. Em *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pp. 436–442.

Mckeown, K., R. J. Passonneau, D. K. Elson, A. Nenkova, & J. Hirschberg (2005). Do summaries help? a task-based evaluation of multi-document summarization. Em *28th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM 38*(11), 39–41.

Nenkova, A., S. Maskey, & Y. Liu (2011). Automatic summarization. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*, HLT '11, Stroudsburg, PA, USA, pp. 3:1–3:86. Association for Computational Linguistics.

Ng, V. & C. Cardie (2002). Improving machine learning approaches to coreference resolution. Em *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Stroudsburg, PA, USA, pp. 104–111. Association for Computational Linguistics.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature 506*, 150 – 152.

O'Donnell, M. (2000). Rsttool 2.4 - a markup tool for rhetorical structure theory. Em *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pp. 253–256.

Otterbacher, J. C., D. R. Radev, & A. Luo (2002). Revisions that improve cohesion in multi-document summaries: A preliminary study. Em *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. & T. H. Dang (2010). Overview of the tac 2010 summarization track. Em *Proceedings of the Text Analysis Conference*, pp. 1.

Papineni, K., S. Roukos, T. Ward, & W.-J. Zhu (2002). Bleu: a method for automatic evaluation of machine translation. Em *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Stroudsburg, PA, USA, pp. 311–318. Association for Computational Linguistics.

Pardo, T. (2008). Sumarização automática: Principais conceitos e sistemas para o português brasileiro. Technical report, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 13 p.

Pardo, T. & M. Nunes (2004). Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em português do brasil. Technical report, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Pardo, T. & M. Nunes (2006). Dizer - an automatic discourse analyzer for brazilian portuguese. Em *V Best MSc Dissertation/PhD Thesis Contest - CTDIA*, Ribeirão Preto-SP, Brazil.

Pardo, T., M. Nunes, & L. Rino (2004). Dizer - an automatic discourse analyzer for brazilian portuguese. Em *17th Brazilian Symposium on Artificial Intelligence - SBIA*, São Luis-MA, Brazil, pp. 224–234. Lecture Notes in Artificial Intelligence.

Pardo, T., L. Rino, & M. Nunes (2003). Gistsumm: A summarization tool based on a new extractive method. Em *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken - PROPOR*, Faro, Portugal, pp. 210–218. Lecture Notes in Artificial Intelligence 2721.

Pardo, T. & E. Seno (2005). Rhetalho: um corpus de referência anotado retoricamente. Em *V Encontro de Corpora*, pp. 1.

Pardo, T. A. S. (2002). Gistsumm: Um sumarizador automático baseado na idéia principal de textos. Technical report, NILC-TR-02-13. 22 p.

Pardo, T. A. S. & M. d. G. V. Nunes (2008). On the development and evaluation of a brazilian portuguese discourse parser. *Journal of Theoretical and Applied Computing 15*(2), 43–64.

Pitler, E., A. Louis, & A. Nenkova (2010). Automatic evaluation of linguistic quality in multi-document summarization. Em *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 544–554. Association for Computational Linguistics.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, & B. Webber (2008a). The penn discourse treebank 2.0. Em *In Proceedings of LREC*.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, & B. Webber (2008b). The penn discourse treebank 2.0. Em *Proceedings of the 6th Internacional Conference on Language Resources an Evaluation (LREC 2008)*.

Prince, E. F. (1981). Toward a taxonomy of given-new information. Em P. Cole (Ed.), *Syntax and semantics: Vol. 14. Radical Pragmatics*, pp. 223–255. New York: Academic Press.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Radev, D. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. Em *1st ACL SIGDIAL Workshop on Discourse and Dialogue, Hong Kong*.

Radev, D., S. Teufel, H. Saggion, W. Lam, J. Blitzer, A. Celebi, H. Qi, D. Liu, & E. Drabek (2002). Evaluation challenges in large-scale multi-document summarization: the mead project. Em *Proceedings of the SIGIR*, pp. 1.

Radev, D. R., J. Otterbacher, & Z. Zhang (2004). Cst bank: A corpus for the study of cross-document structural relationships. Em *LREC*. European Language Resources Association.

Rassi, A., A. Zacarias, J. Maziero, E.G.; Souza, M. Dias, M. Castro Jorge, P. Cardoso, P. Balage Filho, R. Camargo, V. Agostini, A. Di Felippo, E. Seno, L. Rino, & T. Pardo (2013). Anotação de aspectos textuais em sumários do córpus cstnews. Technical report, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Ribaldo, R. (2013). Investigação de mapas de relacionamento para sumarização multidocumento. Monografia de Conclusão de Curso, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Novembro, 61p.

Ribeiro, G. F. & L. H. M. Rino (2005). A sumarização automática com base em a sumarização automática com base em estruturas rst. Technical report, NILC - ICMC-USP.

Rino, L. H. M. & T. A. S. Pardo (2006). A coleção temário e a a coleção temário e a avaliação de sumarização automática. Technical report, ICMC-USP. 15 p.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language 10*, 187–228.

Rossi, D., C. Pinheiro, N. Feier, & R. Vieira (2001). Resolução de correferência em textos da língua portuguesa. *Revista Eletrônica de Iniciação Científica, 1*(2), 1.

Salton, G., A. Singhal, M. Mitra, & C. Buckley (1997). Automatic text structuring and summarization. *Inf. Process. Manage. 33*(2), 193–207.

Schwarm, S. E. & M. Ostendorf (2005). Reading level assessment using support vector machines and statistical language models. Em *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Stroudsburg, PA, USA, pp. 523–530. Association for Computational Linguistics.

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev. 5*(1), 3–55.

Silva, L. L. & V. D. Feltrim (2015). Análise automática de coerência textual em resumos científicos: Avaliando quebras de linearidade. Em *Symposium in Information and Human Language Technology, Natal, Brazil*, pp. 45–49.

Souza, V. M. A. & V. D. Feltrim (2012). A coherence analysis module for scipo: providing suggestions for scientific abstracts written in portuguese. *Journal of the Brazilian Computer Society 19*(1), 59–73.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *JOURNAL OF DOCUMENTATION 28*, 11–21.

Stede, M. (2004). The potsdam commentary corpus. Em *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, DiscAnnotation '04, Stroudsburg, PA, USA, pp. 96–102. Association for Computational Linguistics.

Strube, M. & S. P. Ponzetto (2006). Wikirelate! computing semantic relatedness using wikipedia. Em *Proceedings of the 21st National Conference on Artificial Intelligence*.

Taboada, M. & J. Renkema (2008). Discourse relations reference corpus. Em *Conference Name*, Conference Location. Simon Fraser University and Tilburg University, http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

Tan, P.-N., M. Steinbach, & V. Kumar (2005). *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Triantafillou, E., A. Pomportsis, & E. Georgiadou (2002). Aes-cs: Adaptive educational system based on cognitive styles.

Vieira, R. & V. Lima (2001). Linguística computacional: princípios e aplicações. Em *X Escola Regional de Informática - 2001 (ERI2001)*, Porto Alegre, Brazil, pp. 27–58.

Vliet, N. V. D., I. Berzlánovich, G. Bouma, M. Egg, & G. Redeker (2011). Building a discourse-annotated dutch text corpus. Em *In S. Dipper & H. Zinsmeister (Eds.), Beyond Semantics, Bochumer Linguistische Arbeitsberichte 3*, pp. 157–171.

Webber, B. (2004). D-ltag: extending lexicalized tag to discourse. *Cognitive Science 28*(5), 751 – 779. 2003 Rumelhart Prize Special Issue Honoring Aravind K. Joshi.

Witten, H. I. & E. Frank (2005). *Data mining - practical machine learning tools and techniques*. Morgan Kaufmann - Elsevier.

Witten, I. H., E. Frank, & M. A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd^{a} Edição). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yokono, H. & M. Okumura (2010). Incorporating cohesive devices into entity grid model in evaluating local coherence of japanese text. Em *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing*, CICLing'10, Berlin, Heidelberg, pp. 303–314. Springer-Verlag.

Zaccara, R. C. C. (2012). Anotação e classificação automática de entidades nomeadas em notícias esportivas em português brasileiro. Dissertação de Mestrado, Instituto de Matemática e Estatística da Universidade de São Paulo.

Zhang, R. (2011). Sentence ordering driven by local and global coherence for summary generation. Em *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, Stroudsburg, PA, USA, pp. 6–11. Association for Computational Linguistics.

Zhang, Z., S. Blair-Goldensohn, & D. Radev (2002). Towards cst-enhanced summarization. Em *AAAI 2002 Conference*.

Zhang, Z., J. Otterbacher, & D. R. Radev (2003). Learning cross-document structural relationships using boosting. Em *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8*, pp. 124–130.

Appendices

Apêndice

A

APÊNDICE A - Definições das Relações RST

Nome da Relação:	ANTITHESIS
Restrições sobre o N:	o escritor julga N válido
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	N e S estão em contraste; por causa da aparente incompatibilidade, não se pode julgar N e S válidos ao mesmo tempo; a compreensão de S e da incompatibilidade entre N e S faz o leitor aceitar melhor N
Efeito:	o leitor aceita melhor N

Nome da Relação:	ATTRIBUTION
Restrições sobre o N:	N apresenta uma expressão, fala ou pensamento de alguém ou algo
Restrições sobre o S:	S apresenta alguém ou algo que produz N
Restrições sobre o N + S:	S e N indicam, respectivamente, a fonte de uma mensagem e a mensagem
Efeito:	o leitor é informado sobre a mensagem e sobre quem ou o que a produziu

Nome da Relação:	BACKGROUND
Restrições sobre o N:	o leitor não compreenderá suficientemente N antes de ler S
Restrições sobre o S:	S Nenhuma
Restrições sobre o N + S:	S aumenta a habilidade do leitor em compreender algum elemento em N
Efeito:	a habilidade do leitor para compreender N aumenta

Nome da Relação:	CIRCUMSTANCE
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	S apresenta uma situação (realizável)
Restrições sobre o N + S:	S provê uma situação na qual o leitor pode interpretar N
Efeito:	o leitor reconhece que S provê uma situação na qual N deve ser interpretado

Nome da Relação:	COMPARISON
Restrições sobre o N:	apresenta uma característica de algo ou alguém
Restrições sobre o S:	apresenta uma característica de algo ou alguém comparável com o que é apresentado em N
Restrições sobre o N + S:	as características de S e N estão em comparação
Efeito:	o leitor reconhece que S é comparado a N em relação a certas características

Nome da Relação:	CONCESSION
Restrições sobre o N:	o escritor julga N válido
Restrições sobre o S:	o escritor não afirma que S pode não ser válido
Restrições sobre o N + S:	o escritor mostra uma incompatibilidade aparente ou em potencial entre N e S; o reconhecimento da compatibilidade entre N e S melhora a aceitação de N pelo leitor
Efeito:	o leitor aceita melhor N

Nome da Relação:	CONCLUSION
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	S baseia-se no que é apresentado em N
Restrições sobre o N + S:	S apresenta um fato concluído a partir da interpretação de N
Efeito:	o leitor reconhece que S é uma conclusão produzida devido à interpretação de N

Nome da Relação:	CONDITION
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	S apresenta uma situação hipotética, futura ou não realizada
Restrições sobre o N + S:	a realização de N depende da realização de S
Efeito:	o leitor reconhece como a realização de N depende da realização de S

Nome da Relação:	ELABORATION
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S apresenta detalhes adicionais sobre a situação ou algum elemento apresentado em N
Efeito:	L reconhece que S fornece detalhes adicionais sobre N

Nome da Relação:	ENABLEMENT
Restrições sobre o N:	apresenta uma ação do leitor não realizada
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	a compreensão de S pelo leitor aumenta sua habilidade para realizar a ação em N
Efeito:	a habilidade do leitor para realizar a ação em N aumenta

Nome da Relação:	EVALUATION
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S se relaciona a N pelo grau de avaliação positiva do escritor por N
Efeito:	o leitor reconhece que S avalia N e reconhece o valor que ele atribui

Nome da Relação:	EVIDENCE
Restrições sobre o N:	o leitor poderia não acreditar em N de forma satisfatória para o escritor
Restrições sobre o S:	o leitor acredita em S ou o achará válido
Restrições sobre o N + S:	a compreensão de S pelo leitor aumenta sua convicção em N
Efeito:	a convicção do leitor em N aumenta

Nome da Relação:	EXPLANATION
Restrições sobre o N:	apresenta um evento ou situação
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S explica como e/ou porque o evento ou situação apresentado em N ocorre ou veio a ocorrer
Efeito:	o leitor reconhece que S é a razão para N ou que S explica como N ocorre

Nome da Relação:	INTERPRETATION
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S apresenta um conjunto de ideias que não é expresso em N propriamente, mas derivado deste
Efeito:	o leitor reconhece que S apresenta um conjunto de ideias que não é propriamente expresso no conhecimento fornecido por N

Nome da Relação:	JUSTIFY
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	a compreensão de S pelo leitor aumenta sua prontidão para aceitar o direito do escritor de apresentar N
Efeito:	a prontidão do leitor para aceitar o direito do escritor de apresentar N aumenta

Nome da Relação:	MEANS
Restrições sobre o N:	uma atividade
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S apresenta um método ou instrumento que faz com que a realização de N seja mais provável
Efeito:	o leitor reconhece que o método ou instrumento em S faz com que a realização de N seja mais provável

Nome da Relação:	MOTIVATION
Restrições sobre o N:	uma ação volitiva não realizada
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	a compreensão de S motiva a realização de N
Efeito:	o leitor reconhece que S motiva a realização de N

Nome da Relação:	NON-VOLITIONAL CAUSE
Restrições sobre o N:	apresenta uma ação não volitiva
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S apresenta uma situação que pode ter causado N; sem S, o leitor poderia não reconhecer o que causou a ação em N; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em S como a causa da ação apresentada em N

Nome da Relação:	NON-VOLITIONAL RESULT
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	apresenta uma ação não volitiva
Restrições sobre o N + S:	N apresenta uma situação que pode ter causado S; sem N, o leitor poderia não reconhecer o que causou a ação em S; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em N como a causa da ação

Nome da Relação:	OTHERWISE
Restrições sobre o N:	apresenta uma situação não realizada
Restrições sobre o S:	apresenta uma situação não realizada
Restrições sobre o N + S:	a realização de N impede a realização de S
Efeito:	o leitor reconhece que a realização de N impede a realização de S

Nome da Relação:	PARENTHETICAL
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	apresenta informação extra relacionada a N que não está expressa no fluxo principal do texto
Restrições sobre o N + S:	S apresenta informação extra relacionada a N, complementado N; S não pertence ao fluxo principal do texto
Efeito:	o leitor reconhece que S apresenta informação extra relacionada a N, complementando N

Nome da Relação:	PURPOSE
Restrições sobre o N:	apresenta uma ação
Restrições sobre o S:	apresenta uma situação não realizada
Restrições sobre o N + S:	S apresenta uma situação que pode realizar N
Efeito:	o leitor reconhece que a atividade em N pode ser iniciada por meio de S

Nome da Relação:	RESTATEMENT
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S se relaciona a N; ambos apresentam conteúdo comparável; N é mais importante para a satisfação do objetivo do escritor
Efeito:	o leitor reconhece que S expressa o mesmo conteúdo de N, mas de forma diferente

Nome da Relação:	SOLUTIONHOOD
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	apresenta um problema
Restrições sobre o N + S:	N é uma solução para o problema em S
Efeito:	o leitor reconhece N como uma solução para o problema em

Nome da Relação:	SUMMARY
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S apresenta o conteúdo de N resumido
Efeito:	o leitor reconhece S como um resumo do conteúdo de N

Nome da Relação:	VOLITIONAL CAUSE
Restrições sobre o N:	apresenta uma ação volitiva ou uma situação que poderia surgir de uma ação volitiva
Restrições sobre o S:	Nenhuma
Restrições sobre o N + S:	S apresenta uma situação que pode ter acarretado o fato do agente da ação volitiva em N ter realizado a ação; sem S, o leitor poderia não reconhecer a motivação da ação; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em S como a causa da ação apresentada em

Nome da Relação:	VOLITIONAL RESULT
Restrições sobre o N:	Nenhuma
Restrições sobre o S:	apresenta uma ação volitiva ou uma situação que poderia surgir de uma ação volitiva
Restrições sobre o N + S:	N apresenta uma situação que pode ter acarretado o fato do agente da ação volitiva em S ter realizado a ação; sem N, o leitor poderia não reconhecer a motivação da ação; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em N como a causa da ação apresentada em

Nome da Relação:	CONTRAST
Restrições sobre os Ns:	não mais do que dois Ns; as situações nos Ns são (a) compreendidas como similares em vários aspectos, (b) compreendidas como diferentes em vários aspectos e (c) comparadas em relação a uma ou mais dessas diferenças
Efeito:	o leitor reconhece as similaridades e diferenças resultantes da comparação sendo feita

Nome da Relação:	JOINT
Restrições sobre os Ns:	Nenhuma
Efeito:	Nenhuma

Nome da Relação:	LIST
Restrições sobre os Ns:	itens comparáveis apresentados nos Ns
Efeito:	o leitor reconhece como comparáveis os itens apresentados

Nome da Relação:	SAME-UNIT
Restrições sobre os Ns:	os Ns apresentam informações que, juntas, constituem uma única proposição
Efeito:	o leitor reconhece que as informações apresentadas constituem uma única proposição separadas, não fazem sentido

Nome da Relação:	SEQUENCE
Restrições sobre os Ns:	as situações apresentadas nos Ns são realizadas em sequência
Efeito:	o leitor reconhece a sucessão temporal dos eventos apresentados

Apêndice

B

APÊNDICE B - Definições das Relações CST

Nome da Relação: Identity
Direcionalidade: Nula
Restrições: As sentenças devem ser idênticas
Comentários:

Nome da Relação: Equivalence
Direcionalidade: Nula
Restrições: As sentenças apresentam o mesmo conteúdo, mas expresso de forma diferente
Comentários:

Nome da Relação: Summary
Direcionalidade: S1 $\leftarrow$ S2
Restrições: Summary é um tipo de equivalence, mas summary deve haver diferença significativa de tamanho entre as sentenças.
Comentários: S1 contém X e Y, S2 contém X

Nome da Relação: Subsumption
Direcionalidade: S1 $\rightarrow$ S2
Restrições: S1 apresenta as informações contidas em S2 e informações adicionais.
Comentários: S1 contém X e Y, S2 contém X

Nome da Relação: Overlap
Direcionalidade: Nula
Restrições: S1 e S2 apresentam informações em comum e ambas apresentam informações adicionais distintas entre si.
Comentários: S1 contém X e Y, S2 contém X e Z.

Nome da Relação: Historical background
Direcionalidade: S1 $\leftarrow$ S2
Restrições: S2 apresenta informações históricas sobre algum elemento presente em S1.
Comentários: O elemento elaborado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (por exemplo, overlap).

Nome da Relação: Follow-up
Direcionalidade: S1 $\leftarrow$ S2
Restrições: S2 apresenta acontecimentos que acontecem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si.
Comentários:

Nome da Relação: Elaboration
Direcionalidade: S1 $\leftarrow$ S2
Restrições: S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1.
Comentários: O elemento elaborado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (por exemplo, overlap).

Nome da Relação: Contradiction
Direcionalidade: Nula
Restrições: S1 e S2 divergem sobre algum elemento das sentenças.
Comentários:

Nome da Relação: Citation
Direcionalidade: S1 $\leftarrow$ S2
Restrições: S2 cita explicitamente informação proveniente de S1 em S1.
Comentários: Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.

Nome da Relação: Attribution
Direcionalidade: S1 $\leftarrow$ S2
Restrições: SS1 e S2 apresentam informação em comum e S2 atribui essa informação a uma fonte/autoria presente em S1.
Comentários: Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.

Nome da Relação: Modality
Direcionalidade: S1 ← S2
Restrições: S1 e S2 apresentam informação em comum e em S2 a fonte/autoria da informação é indeterminada/relativizada/amenizada
Comentários: Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total

Nome da Relação: Indirect speech
Direcionalidade: S1 ← S2
Restrições: S1 e S2 apresentam informação em comum; S1 apresenta essa informação em discurso direto e S2 em discurso indireto.
Comentários:

Nome da Relação: Translation
Direcionalidade: Nula
Restrições: S1 e S2 apresentam informação em comum em línguas diferentes.
Comentários:

Apêndice

C

APÊNDICE C - Exemplos de Sumários Anotados com Erros da QL

(S1) O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac).
(S2) <e TYPE=MD CONEC="Mas">Mas, diante da dificuldade para encontrar pessoas que aceitassem assumir uma das diretorias da agência reguladora, após a renúncia de três diretores, Jobim decidiu indicar a economista para o cargo.**</e>**
(S3) <e TYPE=SEM_REL>Como os diretores de agências têm mandato de cinco anos, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa.**</e>**
(S4) <e TYPE=SEM_REL>Inicialmente, Solange Vieira, que é assessora especial de Jobim havia sido escolhida para comandar a Secretaria Nacional de Aviação Civil, a ser criada na estrutura do ministério, segundo a assessoria de imprensa do ministério.**</e>**

Figura C.1: Sumário Anotado da coleção 5 do córpus CSTNews

(S1) <e TYPE=1M-EXP>Lula**</e>** disse que além de melhorar a qualidade de vida dos brasileiros, **<e TYPE=SNdef-REF>**as obras**</e>** vão gerar empregos.
(S2) "Algumas (obras) já estão em andamento, outras vão começar a andar agora, outras ainda precisam de licenciamento".
(S3) <e TYPE=SNdef-REF>O presidente**</e>** afirmou ainda que o critério para os municípios e Estados contemplados com obras é técnico.
(S4) <e TYPE=SEM_REL><e TYPE=SEN_INC>E a novidade é o compromisso dos governadores.**</e></e>**
(S5) Lula disse que a prioridade é a realização de obras nas regiões metropolitanas de grandes centros urbanos.
(S6) <e TYPE=SEN_INC>O dado concreto é que nós vamos fazer deste país um verdadeiro canteiro de obras em se tratando de infra-estrutura, disse.**</e>**
(S7) <e TYPE=RED SENT=S2>Algumas já estão em andamento, outras vão começar a andar agora, outras ainda precisam de licenciamento.**</e>**

Figura C.2: Sumário Anotado da coleção 6 do córpus CSTNews

(S1) Quinze voluntários da **<e TYPE=ACR-EXP>**ONG**</e>** francesa Ação Contra a Fome (ACF) foram assassinados no nordeste do Sri Lanka, informou hoje um porta-voz da organização.
(S2) Segundo um representante do grupo **<e TYPE=Outros EXPLANATION="Referência em Francês para termo introduzido em Português">**Action Contre la Faim**</e>**, os corpos foram encontrados no escritório da organização.
(S3) O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".
(S4) Até o momento, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.
(S5) <e TYPE=SEM_REL><e TYPE=SNdef-REF>Os rebeldes**</e>** afirmaram que consideram **<e TYPE=SNdef-REF>**o novo bombardeio do Exército**</e>** equivalente "uma declaração de guerra".**</e>**

Figura C.3: Sumário Anotado da coleção 13 do córpus CSTNews

(S1) O hipocentro **<e TYPE=SNdef-REF>**do terremoto**</e>** se localizou 17 km abaixo do nível do mar.
(S2) Foi o pior do país desde 1995, quando um tremor de magnitude 7,3 matou mais de 6.400 pessoas na cidade de Kobe.
(S3) O Japão é um dos países mais atingidos por terremotos no mundo, com um tremor a cada pelo menos cinco minutos.
(S4) <e TYPE=SEM_REL><e TYPE=Outros EXPLANATION="Declaração sem fonte">- Se pudermos restaurar os serviços de água, mais gente poderá ir para casa, então é isso que queremos fazer primeiro.**</e></e>**
(S5) - Os danos são piores que o esperado - afirmou a repórteres Hiroshi Kaeda, prefeito de Kashiwazaki.
(S6) Há preocupação com a saúde dos desabrigados, muitos dos quais idosos.
(S7) Nove idosos morreram e uma pessoa está desaparecida.
(S8) <e TYPE=Outros EXPLANATION="Declaração sem fonte">- Não sei quando poderei voltar para casa.**</e>**
(S9) Cerca de 9.000 pessoas passariam uma segunda noite em escolas e outros centros de resgate improvisados.
(S10) <e TYPE=SEM_REL>Nesta terça-feira, **<e TYPE=SNdef-REF>**a empresa**</e>** também admitiu que uma pequena quantidade de materiais radioativos escapou para a atmosfera.**</e>**

Figura C.4: Sumário Anotado da coleção 32 do córpus CSTNews

(S1) <e TYPE=SNdef-REF>Os dois**</e>** estiveram na casa **<e TYPE=SNdef-REF>**do ministro**</e>** na manhã da última terça-feira.
(S2) 'O Globo' confirmou que semana passada houve um encontro secreto de Mantega, com Tasso Jereissati e Sergio Guerra.
(S3) <e TYPE=SEM_REL>Segundo o líder do governo no Senado, Romero Jucá (**<e TYPE=ACR-EXP>**PMDB**</e>**-**<e TYPE=ACR-EXP>**RR**</e>**), **<e TYPE=SNdef-REF>**esses abatimentos**</e>** resultariam em uma desoneração de cerca de 2 bilhões de reais por ano.**</e>**
(S4) Na reunião entre Mantega e tucanos, o governo também reafirmou a disposição de conceder outras desonerações para empresas, que juntas somariam outros 2 bilhões de reais.
(S5) A proposta de reduzir a alíquota da contribuição paga pelas pessoas jurídicas **<e TYPE=SNdef-REF>**ao Sistema S**</e>**, no entanto, não entrará no acordo com o **<e TYPE=ACR-EXP>**PSDB**</e>**, disse Jucá.
(S6) - Continuamos estudando o assunto, mas será algo mais para frente - afirmou o líder a jornalistas.
(S7) - Achamos que a questão da (desoneração da) **<e TYPE=ACR-EXP>**CPMF**</e>** deve ser mais abrangente.
(S8) - Consideramos a proposta um avanço, mas ainda insuficiente - afirmou **<e TYPE=nM+EXP SENT=S2 TEXT="Tasso Jereissati">**o presidente do PSDB, Tasso Jereissati (**<e TYPE=ACR-EXP>**CE**</e>**)**</e>**, a jornalistas após reunião com **<e TYPE=nM+EXP SENT=S1 TEXT="do ministro">**o ministro da Fazenda, Guido Mantega**</e>**.
(S9) <e TYPE=SEM_REL>Para salários acima **<e TYPE=SNdef-REF>**desse limite**</e>**, haverá um desconto fixo no valor de 214 reais, também na declaração do imposto de renda.**</e>**

Figura C.5: Sumário Anotado da coleção 50 do córpus CSTNews

(S1) Foi **<e TYPE=SNdef-REF>**o maior aumento de autuações**</e>** entre os setores econômicos fiscalizados em 2007.
(S2) <e TYPE=SEM_REL>Para fazer **<e TYPE=SNdef-REF>**a consulta**</e>**, basta entrar **<e TYPE=SNdef-REF>**no site**</e>** e clicar na seção IRPF - Extrato Simplificado do Processamento.**</e>**
(S3) <e TYPE=SEM_REL>Os **<e TYPE=PRO-REF>**outros**</e>** motivos mais constantes são a omissão de rendimentos imobiliários e despesas médicas incompatíveis.**</e>**
(S4) Segundo **<e TYPE=1M-EXP>**Cardoso**</e>**, esta omissão representa cerca de dois terços das declarações que estão em revisão.
(S5) <e TYPE=RED SENT=S3,S4>O motivo que mais leva os contribuintes à malha fina é a omissão de renda própria ou de dependentes.**</e>**
(S6) <e TYPE=PRO-REF>Elas**</e>** podem recorrer administrativamente na própria administração tributária ou na Justiça.
(S7) <e TYPE=SNdef-REF>O valor**</e>** não significa que essas pessoas físicas precisam fazer o pagamento de forma imediata.
(S8) <e TYPE=Outros EXPLANATION="Falta informação na sentença">Nos sete primeiros meses do ano foi R$ 1,339 bilhão, um aumento de 316,5%.**</e>**
(S9) Por conta do maior número de dados disponíveis, também subiu o valor do quanto esses contribuintes podem ter deixado de recolher à Receita.
(S10) No ano passado, tiveram suas declarações revistas 209.364 contribuintes.
(S11) Um dos motivos é que cada parâmetro era analisado isoladamente, como por exemplo os dados referentes a rendimentos imobiliários.

Figura C.6: Sumário Anotado da coleção 34 do córpus CSTNews

(S1) <e TYPE=Outros EXPLANATION="Falta sujeito e objeto">Fez mais.**</e>**
(S2) Entrou pela direita e cruzou rasteiro.
(S3) <e TYPE=SEM_REL>Dessa vez, nem foi necessária, a presença dos astros de Barcelona e Milan.**</e>**
(S4) Impecável, a Seleção, comandada por Ronaldinho Gaúcho e Kaká aplicou uma goleada humilhante por 4 a 1 e trouxe a taça.
(S5) <e TYPE=SEM_REL>O **<e TYPE=PRO-REF>**outro**</e>** revés aconteceu em 2005, na Copa das Confederações, na Alemanha.**</e>**
(S6) <e TYPE=RED SENT=S4><e TYPE=SEM_REL>No final, vitória brasileira.**</e></e>**
(S7) <e TYPE=SEM_REL><e TYPE=SNdef-REF>No ataque seguinte**</e>**, **<e TYPE=1M-EXP>**Adriano**</e>**, que terminou **<e TYPE=SNdef-REF>**a competição**</e>** como artilheiro, com sete gols, empatou e levou a decisão para os pênaltis.**</e>**
(S8) Em 2004, também pela Copa América, os argentinos venciam por 2 a 1, quando Tevez resolveu "fazer graça" perto da bandeirinha de escanteio.
(S9) É a terceira final seguida que eles perdem para a Seleção.
(S10) <e TYPE=SEM_REL><e TYPE=1M-EXP>Vágner Love**</e>** puxou um contra-ataque rápido e tocou na medida, nas costas da zaga, para **<e TYPE=1M-EXP>**Daniel Alves**</e>**, que chutou forte, colocado, sem chances para **<e TYPE=1M-EXP>**Abbondanzieri**</e>**.**</e>**
(S11) Aos 23 da segunda etapa, o golpe fatal.
(S12) <e TYPE=Outros EXPLANATION="Inclusão de metadados">RIO - **</e>**Foi um domingo especial e inesquecível para o esporte brasileiro.
(S13) <e TYPE=SEM_REL><e TYPE=MD CONEC="Mas">Mas a partir **<e TYPE=SNdef-REF>**do segundo**</e>**, tudo voltou ao normal.**</e></e>**

Figura C.7: Sumário Anotado da coleção 25 do córpus CSTNews