Sumarização automática multidocumento: seleção de conteúdo com base no Modelo CST (*Cross-document Structure Theory*)

Maria Lucía del Rosario Castro Jorge

Orientador: ***Prof. Dr. Thiago Alexandre Salgueiro Pardo***

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

USP – São Carlos

Fevereiro/2010

Agradecimentos

Agradeço primeiramente a Deus e a minha família pelo apoio durante os anos de mestrado.

Também gostaria de agradecer ao meu orientador Prof. Thiago Alexandre Salgueiro Pardo pela paciência e pelos ensinamentos ao longo destes dois anos de trabalho.

Agradeço a CNPq pelo apoio financeiro durante o mestrado.

Agradeço o apoio dos meus colegas do NILC pelo aprendizado de todos os dias.

Agradeço também aos funcionários da USP pela disposição de ajuda oferecida ao longo destes dois anos.

Resumo

A sumarização automática multidocumento consiste em produzir um sumário ou resumo (como mais comumente é conhecido) a partir de um grupo de textos que versam sobre um mesmo assunto, contendo as informações mais relevantes de acordo com o interesse do usuário. No cenário atual, com a quantidade imensa de informação em constante crescimento e atualização, e o tempo cada vez mais reduzido disponível para apreender o conteúdo de interesse, sumários multidocumento têm se tornado um recurso importante. Nesta dissertação, foram explorados métodos de seleção de conteúdo para sumarização multidocumento com base no modelo de relacionamento multidocumento CST (*Cross-document Structure Theory*), proposto recentemente e já difundido na área de Processamento de Línguas Naturais. Em particular, neste trabalho, foram definidos e formalizados operadores de seleção de conteúdo para sumarização multidocumento com base no modelo CST. Estes operadores representam possíveis preferências de sumarização e focam-se no tratamento dos principais desafios presentes no processamento de múltiplos documentos: redundância, complementaridade e informações contraditórias. Estes operadores são especificados em *templates* contendo regras e funções que relacionam essas preferências às relações CST. Especificamente, foram definidos operadores para extrair a informação principal, apresentar informação de contexto, identificar autoria, tratar redundâncias e identificar informação contraditória. Também foi avaliado o impacto do uso do modelo CST em métodos de sumarização superficiais. Experimentos foram realizados com textos jornalísticos escritos em português brasileiro. Os resultados das avaliações mostram que o uso da teoria CST melhora a informatividade e a qualidade dos sumários gerados.

Abstract

Multidocument summarization consists in producing a summary from a group of texts on a same topic, containing the most relevant information according to the user's interest. Recently, with the huge amount of growing information over the internet and the short time available to learn and process the information of interest, automatic summaries have become a very important resource. In this work, we explored content selection methods for multidocument summarization based on CST (Cross-document Structure Theory) a recently proposed model and already investigated in the Computational Linguistics area. Particularly, in this work we defined and formalized content selection operators based on CST model. These operators represent possible summarization preferences and they focus on the treatment of the main challenges of multidocument summarization: redundancy, complementarity and contradiction among information. These operators are specified in templates containing rules and functions that relate the preferences to CST relations. Specifically, we define operators for extracting main information, context information, identifying authorship, treating redundancy and showing contradicted information. We also explored the impact of CST model over superficial summarization methods. Experiments were done using journalistic texts written in Brazilian Portuguese. Results show that the use of CST model helps to improve informativeness and quality in automatic summaries.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE QUADROS

Capítulo

1

1. Introdução

1.1. Contexto e motivação

Com o advento da internet e a enorme quantidade de informação disponível, principalmente *on-line*, e o tempo cada vez mais escasso que as pessoas têm para absorver a informação que lhes interessa, a sumarização mostra-se como uma atividade muito importante. Para se ter uma idéia, um estudo recente realizado pela equipe de pesquisa do *International Data Corporation* (IDC) estimou que no ano de 2009 o mundo gerou 800 exabytes[1] de informação digital, e, para 2012, estima-se uma quantidade 3 vezes maior do que em 2009. Nesse cenário, a sumarização automática (SA) mostra-se como uma tarefa que pode auxiliar significativamente as pessoas a apreenderem as informações que lhes interessem.

A tarefa de SA consiste na produção automática de uma versão mais curta de um texto-fonte, seu sumário, ou, como é mais conhecido, seu resumo (Mani, 2001), sendo que este deve conter as informações mais relevantes de acordo com a preferência do usuário. Chama-se de sumário monodocumento o sumário produzido a partir de um único texto/documento. Com a web e a grande quantidade de informação, surgiu a área de sumarização multidocumento (McKeown e Radev, 1995; Radev e McKeown, 1998), na qual se procura fazer um sumário de um conjunto de textos que abordam um mesmo tópico. Um sumário multidocumento deve conter as principais informações sobre um determinado evento no decorrer do tempo e/ou sob diferentes perspectivas. Por exemplo, poder-se-ia construir um único sumário sobre tudo que foi noticiado sobre o aquecimento global ou o terremoto do Haiti. Como ilustração, a Figura 1 mostra um sumário multidocumento retirado do trabalho de Radev e McKeown (1998), produzido

[1] 1 Exabyte corresponde a 1 bilhão de Gigabytes

a partir de 4 textos sobre atentados terroristas em Israel em Março de 1996. O sumário foi mantido em inglês, como na obra original.

> *Reuters reported that 18 people were killed in a Jerusalem bombing Sunday. The next day, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel radio. Reuters reported that at least 12 people were killed and 105 wounded. Later the same day, Reuters reported that the radical Muslim group Hamas had claimed responsibility for the act.*

Figura 1: Exemplo de sumário multidocumento (Radev e McKeown, 1998, p. 478)

Na atualidade, existem várias aplicações da sumarização automática na Web. Por exemplo, se uma pessoa quiser as informações das principais notícias em relação a um tópico em particular, o Google Notícias fornece um pequeno resumo da informação mais relevante ao tópico, além dos links correspondentes a cada uma das páginas das notícias originais. Em sites de venda de livros na internet, por exemplo, o resumo pode auxiliar o usuário na escolha do produto a comprar. Os sumários automáticos também são úteis em aplicações do tipo bibliotecas digitais, onde se fornecem os resumos referentes aos artigos de preferência do usuário.

Na sumarização multidocumento, existem outros desafios além da identificação de informação relevante no conjunto de textos. Entre os maiores desafios estão a manutenção da coerência e coesão dos sumários. A coerência é influenciada por fatores como informação redundante no sumário, ordenação (temporal ou não) dos segmentos textuais que compõem os sumários, fusão de segmentos textuais com informações complementares e tratamento de informações contraditórias. De outro lado, a coesão observa fatores de continuidade na superfície textual, como boa pontuação e uso de itens lexicais, uso apropriado de anáforas, dentre outros fenômenos. Estes últimos fatores também interferem na coerência do texto.

Além desses desafios, na sumarização multidocumento também se deve levar em conta que os textos podem se originar de fontes diferentes e, em geral, são escritos por pessoas diferentes e, portanto, têm estilos diversos. Finalmente, considera-se também a taxa de compressão aplicada ao sumário. A taxa de compressão indica o tamanho do sumário em relação aos textos de entrada (em geral, em relação ao número de palavras), o que pode influenciar na informatividade do sumário. Em particular, quanto mais alta é taxa de compressão, menor é o sumário e, portanto, incluem-se menos informações. Para ilustrar alguns desafios que se apresentam na sumarização multidocumento, as Figuras 2 e 3 mostram dois resumos multidocumento produzidos automaticamente com

taxas de compressão de 70% sobre os textos de entrada. Nas figuras, cada número destaca um dos desafios da sumarização multidocumento. As sentenças indicadas com o número 1 mostram informações redundantes; sentenças indicadas com o número 2 não estão em ordem; sentenças indicadas com o número 3 não têm coesão nem coerência; sentenças indicadas com o número 4 mostram informações contraditórias; por último, as informações indicadas com o número 5 são complementares.

BRASÍLIA – [**A Receita Federal intensificou a fiscalização e o resultado foi um aumento do número de contribuintes que caíram na malha fina.**]**[1]** **[A Receita Federal intensificou a fiscalização sobre as declarações das pessoas físicas neste ano.]**[1] [**A expectativa da Receita é que até o final do ano mais de 300 mil contribuintes sejam autuados pela malha fina.]**[2]
[A partir deste ano, o foco passou a ser o contribuinte e todos os parâmetros disponíveis sobre ele são cruzados.][2] **"**A produtividade da malha ficou maior", disse o coordenador.
Segundo o secretário-adjunto da Receita Federal, Paulo Ricardo, dois terços das autuações de contribuintes pela malha fina são por omissão de renda. Também é freqüente sonegação de despesas médicas e de omissão de renda de aluguel.
Entre janeiro e julho, os auditores fiscais identificaram 208.471 declarações que precisaram ser revisadas, um crescimento de 104,5% na comparação com o mesmo período do ano passado.
[Para saber se caiu na malha fina, o contribuinte deve acessar o site www.receita.fazenda.gov.br.][3] **[Elas podem recorrer administrativamente na própria administração tributária ou na Justiça.]**[3]

Figura 2: **Exemplo de sumário multidocumento**

RIO – [**Um dos destaques desta temporada do esporte brasileiro, a ginasta Jade Barbosa foi escolhida, na noite desta terça-feira, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim.]**[1]
[A ginasta Jade Barbosa, que obteve duas medalhas nos Jogos Pan-Americanos do Rio][4] **[em julho, venceu votação na internet]**[5] e [**será a representante brasileira no revezamento da tocha olímpica para Pequim-2008.**][1]
A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico. Por isso, Jade participará do evento em Buenos Aires, na Argentina, única cidade da América do Sul a receber o símbolo dos Jogos.
[Em votação pela internet, a ginasta recebeu mais de 100 mil votos e superou o nadador Thiago Pereira, que ganhou seis ouros nos Jogos Pan-Americanos.][5]
Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril. [**Aos 16 anos, Jade conquistou três medalhas no Pan]**[4]: ouro na disputa dos saltos, prata na apresentação por equipes e bronze no

Figura 3: Exemplo de sumário multidocumento

Nestes exemplos os desafios não foram tratados adequadamente o que faz que os textos sejam pouco coerentes. Na tarefa de SA estes desafios devem ser tratados de forma que, as redundâncias sejam eliminadas, os segmentos informativos complementares sejam ordenados de forma coerente e coesa e as informações contraditórias sejam apresentadas destacando a contradição de forma coerente (ou as informações sejam eliminadas).

Mani (2001) afirmou que a SA multidocumento não é intuitiva para humanos, mas McKeown et al. (2005) demonstraram que, apesar das dificuldades, sumários multidocumento produzidos tanto automaticamente quanto por humanos se mostraram muito úteis em experimentos que simulavam a apreensão de informação por humanos. Por exemplo, um sumário multidocumento referente ao conjunto de notícias coletadas pelo Google Notícias em relação à Bolsa de Valores ajuda a se ter uma idéia global sobre os eventos que se referem a tal assunto e, portanto, ajuda na compreensão e assimilação da informação.

Existem duas abordagens tradicionais para a SA em geral: a superficial (ou empírica/estatística) e a profunda (ou fundamental). A primeira abordagem é mais simples e barata, pois leva em consideração pouco ou nenhum conhecimento lingüístico para o processamento. A segunda abordagem faz uso de mais conhecimento lingüístico, levando em consideração regras gramaticais, semântica, conhecimento discursivo e de mundo.

Na abordagem profunda monodocumento, um dos principais modelos que guiam o processo de sumarização é a RST (Rhetorical Structure Theory) (Mann e Thompson, 1987). Por essa teoria, todo texto tem suas partes relacionadas por relações discursivas, ou seja, relações que se estabelecem entre o conteúdo de segmentos do texto, por exemplo, causa-efeito, contraste e elaboração. Tal teoria foi intensamente investigada para a SA monodocumento (por exemplo, O'Donnel, 1997; Marcu, 2000; Pardo e Rino, 2002; Uzêda et al., 2009) principalmente por sua característica de distinguir junto a cada relação estabelecida quais são os segmentos mais importantes.

Na sumarização multidocumento, no contexto da abordagem profunda, também se faz uso de conhecimento lingüístico mais profundo. Em geral, utilizam-se analisadores sintáticos, semânticos e/ou discursivos para estruturar múltiplos textos de entrada. Uma das formas mais usadas na abordagem profunda para estruturar um conjunto de textos que versam sobre um mesmo assunto é o estabelecimento de relações entre eles. Estas relações podem representar vínculos semânticos ou discursivos (se for o caso) entre

diferentes partes dos textos. Neste contexto, com o crescimento das pesquisas em SA multidocumento, surgiu o modelo CST (Cross-document Structure Theory) (Radev, 2000). Este modelo se inspira na RST e é o único dessa natureza para textos de qualquer domínio que se conhece. Neste modelo, propõe-se um conjunto de relações que permitem identificar similaridades, diferenças, contradições e informações complementares entre partes de textos sobre um mesmo tópico. Como exemplo, considere os trechos de dois textos sobre um mesmo tópico ilustrados na Figura 4 a seguir.

Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 13 pessoas na quinta-feira à tarde, informou nesta sexta feira um porta-voz das Nações Unidas.
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.

Figura 4: Trechos de textos sobre um mesmo tópico

Pode-se observar neste exemplo como duas sentenças que contêm informações sobre um mesmo assunto manifestam várias relações entre si, neste caso, de contradição (no número de pessoas mortas no acidente) e de complementaridade (informação a mais sobre o avião e o vôo, antes do acidente). Estas relações podem ser modeladas pela CST.

Apesar de sua criação recente, o modelo CST tem sido aplicado com sucesso na SA multidocumento (Radev et al., 2000, 2001a, 2001b; Zhang et al., 2002; Afantenos et al., 2004, 2007). Tais trabalhos utilizam-se do fato de que é possível tratar os desafios da sumarização multidocumento conhecendo-se as relações existentes entre os segmentos dos textos que estão sendo processados, como relações de equivalência semântica, elaboração e contextualização. Os resultados de alguns destes trabalhos mostram que a CST é útil para identificar similaridades, diferenças, informações complementares e contraditórias entre os textos, o que permite melhorar a informatividade e qualidade dos sumários.

Independente da abordagem que se segue, Mani e Maybury (1999) sugerem uma arquitetura genérica para a SA, a qual é apresentada na Figura 5. Na etapa de análise, os textos de origem são processados e seu conteúdo é representado em um ou mais níveis de análise lingüística: morfológico, sintático, semântico e/ou discursivo. O relacionamento entre os elementos pode ser visualizado como um espaço

multidimensional. Na etapa de transformação, o conteúdo representado é refinado, sendo que isto é feito por meio de operações de seleção de conteúdo relevante, eliminação de elementos irrelevantes e combinação de informações. Finalmente, na etapa de síntese, o conteúdo selecionado é organizado e expresso em língua natural, podendo-se utilizar métodos de geração de texto.

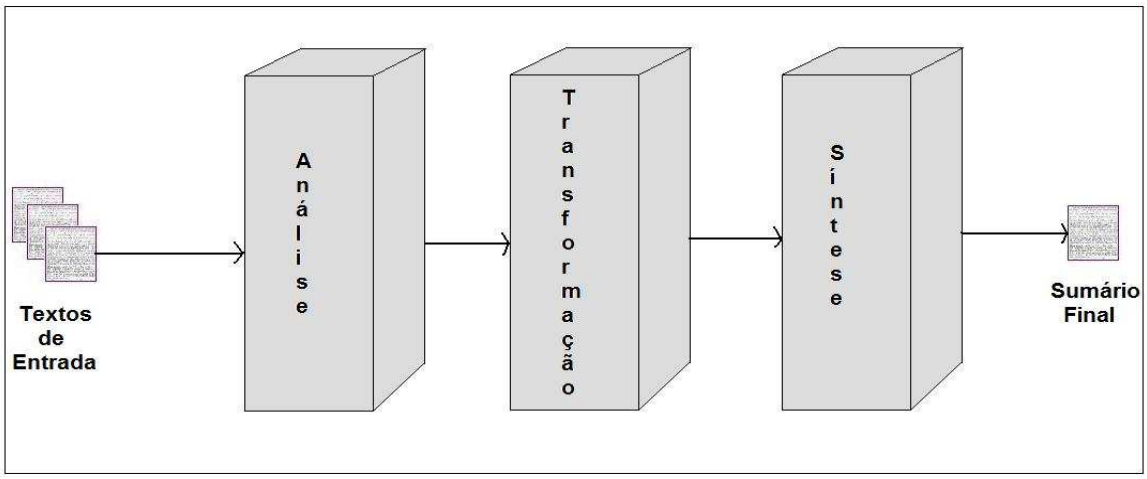

Figura 5: Arquitetura genérica de sistemas de SA

A seleção de conteúdo é a tarefa mais importante da etapa de transformação. Segundo Reiter e Dale (2000) a tarefa de seleção de conteúdo consiste em decidir quais informações devem ser incluídas no texto final. Neste trabalho, em particular, o foco é a etapa de transformação. A partir de textos já analisados, exploram-se métodos de seleção de conteúdo, conteúdo este que será expresso posteriormente na etapa de síntese. Todo o processo é modelado e dirigido pelo modelo CST. A seguir, são descritos lacunas, hipóteses e objetivos deste trabalho.

1.2. Lacunas, hipóteses e objetivos

A tarefa de seleção de conteúdo para sumarização multidocumento, no contexto da abordagem profunda, tem sido muito pouco investigada para o português do Brasil e para as outras línguas em geral. Para se ter uma idéia, para o português, em particular, só se conhece um sistema de sumarização multidocumento (Pardo, 2005), o qual é da abordagem superficial. Além disso, acredita-se que, para se lidar adequadamente com a riqueza dos fenômenos multidocumento (já mencionados anteriormente), abordagens que fazem pouco uso de conhecimento lingüístico são limitadas e, portanto, não são capazes de produzir sumários tão bons quanto os produzidos por humanos. Nessa

perspectiva, a abordagem profunda pode fornecer resultados melhores. Há, portanto, lacunas importantes a serem exploradas na área.

O objetivo principal deste trabalho é a investigação de métodos de seleção de conteúdo para sumarização multidocumento com base no modelo CST, avaliando o impacto do seu uso na sumarização de textos em português. Em particular, são definidos e formalizados operadores de seleção de conteúdo que representam possíveis preferências de sumarização especificadas pelo usuário. Estes operadores exploram as relações CST entre os textos e utilizam esse conhecimento para atender a demanda do usuário e para lidar com os fenômenos multidocumento.

As seguintes hipóteses permearam este trabalho:

- a CST permite modelar adequadamente textos em português brasileiro, estabelecendo relações semântico-discursivas entre os textos;
- as informações fornecidas pela CST permitem explorar o conteúdo dos textos e assim identificar as informações mais relevantes;
- a CST permite identificar redundâncias, contradições e informações complementares, ajudando a melhorar a qualidade dos sumários produzidos;
- o uso da CST ajuda a melhorar a informatividade dos sumários automáticos.

A CST é o modelo multidocumento adotado neste trabalho devido a sua generalidade e possibilidade de ampla aplicação. Além disso, é um dos poucos modelos multidocumento existentes, como será discutido nos capítulos seguintes desta dissertação.

Na etapa de análise foi feita a anotação dos textos segundo a CST, considerando um conjunto refinado das relações do modelo. Assume-se esta anotação como entrada na etapa de transformação onde são aplicados diversos métodos de seleção de conteúdo. O conteúdo selecionado pelos métodos explorados neste trabalho é expresso na etapa de síntese, que neste trabalho consiste basicamente em justapor e ordenar de forma simples as informações disponíveis. Também se faz uso, quando apropriado, de um sistema de fusão de sentenças, introduzido posteriormente neste trabalho.

Para a realização do trabalho proposto, foram definidos e formalizados operadores de seleção de conteúdo para sumarização multidocumento com base no modelo CST. Estes operadores representam possíveis preferências de sumarização e focam-se no tratamento dos principais desafios presentes no processamento de múltiplos documentos: redundância, complementaridade e informações contraditórias. Estes

operadores são especificados em *templates* contendo regras e funções que relacionam essas preferências às relações CST. Especificamente, são definidos operadores para extrair a informação principal, apresentar informação de contexto, identificar autoria, tratar redundâncias e identificar informação contraditória. A decisão sobre as preferências a seguir na elaboração do sumário e, portanto, que operadores de seleção de conteúdo aplicar, provêm do usuário/leitor.

Neste trabalho, avalia-se também o impacto do uso do modelo CST na seleção de conteúdo de dois sumarizadores superficiais: o MEAD (Radev et al., 2001a), que é um dos sumarizadores multilíngües mais amplamente conhecidos na área, e o Gistsumm (Pardo, 2005), que atualmente é o único sumarizador multidocumento dedicado ao português do Brasil. O propósito desta experiência foi mensurar o impacto do uso deste tipo de conhecimento na informatividade dos sumários produzidos por esses sistemas.

Todas as abordagens investigadas são avaliadas segundo os principais pontos tratados na sumarização multidocumento: informatividade, redundância, complementaridade e a contradição de informações.

Para esta dissertação foi usado o córpus CSTNews (Aleixo e Pardo, 2008a), que é composto por um conjunto de 50 coleções de textos jornalísticos em português do Brasil, sendo que cada coleção contém textos que versam sobre um mesmo assunto. Este córpus foi anotado com base na teoria CST e constitui o primeiro córpus anotado com a CST para o português do Brasil.

Esta investigação constitui a primeira pesquisa de abordagem profunda para sumarização multidocumento de textos em português, por isto as contribuições são várias. Primeiramente, em termos práticos, tem-se o primeiro protótipo de sumarização com métodos de seleção de conteúdo de abordagem profunda e o primeiro córpus anotado segundo a CST. Em termos teóricos, a CST é validada para a língua portuguesa e é estabelecida uma primeira formalização de métodos de seleção de conteúdo com base no modelo CST; também são apresentadas as vantagens e desvantagens do uso do modelo na estruturação multidocumento e na produção de sumários em português. Os fenômenos da redundância, complementaridade e contradição de informações são analisados sob a ótica do modelo.

1.3. Organização do texto

O trabalho está organizado da seguinte forma: no Capítulo 2, apresenta-se uma revisão literária sobre o modelo CST. No Capítulo 3, estudam-se a sumarização

multidocumento e as diversas metodologias relacionadas aplicadas na área. No Capítulo 4, relata-se a exploração da seleção de conteúdo com base no modelo CST e expõem-se os resultados obtidos para cada uma das abordagens desenvolvidas neste trabalho. Finalmente, no Capitulo 5, são feitas as considerações finais deste trabalho, destacando-se os principais trabalhos futuros.

Capítulo

2

2. CST*: Cross- document Structure Theory*

2.1. Trabalhos prévios

Como foi visto no capítulo anterior, a teoria discursiva CST (Radev, 2000) surge como um modelo de estruturação de textos, que estabelece relações entre unidades informativas que apresentam similaridades, diferenças, contradições e informações complementares, sendo que estas unidades informativas pertencem a textos que versam sobre um mesmo assunto.

A CST inspira-se em vários trabalhos anteriores. Uns dos primeiros e principais trabalhos que guiaram o estabelecimento de relações entre vários textos são os trabalhos de Trigg (1983) e Trigg e Weiser (1986). Trigg (1983) apresentou uma tipologia fixa de *links* entre documentos científicos. A estrutura primitiva fixa que Trigg propôs tem a intenção de cobrir todas as necessidades de relacionamentos entre os textos. Para isto, Trigg dividiu os *links* em dois tipos: normais e de comentário.

Os *links* de tipo normal são aqueles que trabalham diretamente com o conteúdo: a apresentação do problema, argumentação e especificações teóricas. Alguns exemplos desses *links* são: *support* (que dá apoio e argumentação a uma afirmação), *refutation* (que argumenta contra a idéia), *background* (fornece um cenário da idéia principal), etc.

Os *links* de tipo comentário, tal como o nome especifica, fazem comentários referentes ao texto ou ao autor do texto. Trigg divide os *links* de tipo comentário em três categorias de propósito geral: c*omment*, c*ritical* e *supportive*. Note que o *supportive* do tipo comentário é diferente ao *support* do tipo normal, pois o que caracteriza um *link* de tipo normal é que ele tem argumentos e informações novas em relação ao texto. Os *links* de tipo comentário só se limitam a oferecer uma opinião de apoio ou discordância por parte do leitor.

Além de classificar a tipologia de *links*, Trigg também sugere dois tipos de direcionamento dos *links*: o direcionamento físico e o direcionamento semântico. O primeiro tipo considera o direcionamento do link de acordo com a ordem em que se faz a leitura do link. Por exemplo, se há um *link* que vai de A a B, a leitura de A antecede à leitura de B. O segundo tipo de direcionamento depende do significado semântico do *link*, que não necessariamente concorda com o direcionamento físico Veja-se, por exemplo, o *link citation* (Trigg, 1983). Se A cita B, o direcionamento semântico é de A a B, mais o *link* físico não é o mesmo, pois B tem que ser lido antes de A. Segundo Trigg, os *links* de tipo comentário sempre apresentam os dois direcionamentos de forma oposta, o que não é padrão nos *links* de tipo normal.

Esta taxonomia é a base do TEXTNET (Trigg e Weisser, 1986). Esta é uma ferramenta que propõe uma nova forma de estruturar textos científicos relacionando trechos de documentos, e facilitando a exploração destes por parte do usuário. Radev (2000) afirma que a deficiência principal no trabalho do Trigg (1983) é que a taxonomia é limitada para um domínio específico e não para textos de domínio geral.

Outra pesquisa importante em relacionamento entre documentos surgiu na década de 90 com o trabalho de Allan (1996). Essa pesquisa mostra como identificar *hyperlinks* automaticamente. Para isso, os tipos de *links* são divididos em três categorias: mapeamento de padrões, *links* automáticos e *links* manuais. A categoria de mapeamento de padrões considera basicamente a similaridade lexical das palavras ou sentenças entre documentos para estabelecer os *links*. Os *links* manuais são aqueles que geralmente são feitos por humanos, pois requerem um processamento profundo de língua natural. Estes tipos de *links* conectam documentos que contêm componentes de debate, argumentação, implicações lógicas, circunstâncias dos acontecimentos, etc. Alguns exemplos de *links* manuais são: *caused_by*, *purpose*, *warning,* etc. Por ultimo, os *links* automáticos indicam similaridades semânticas ou elementos discursivos entre os textos, como relações conhecidas de equivalência, contraste, resumo, etc. Diferentemente dos outros dois tipos de *links*, os *links* automáticos não possuem a dificuldade de identificação dos *links* manuais, mas também não são tão triviais como os *links* de mapeamento de padrões.

A partir desses conceitos, a primeira tarefa no processo de estabelecimento de *links* proposto por Allan consiste em estabelecer a similaridade dos documentos e de suas partes em relação a uma consulta feita pelo usuário. Desta forma, só se relacionam documentos que tem maior relevância de acordo com um tópico definido pelo usuário.

Depois, cada parte de um documento é comparada com cada unidade dos outros documentos para calcular a similaridade entre eles. Os *links* de segmentos que têm similaridade fraca são descartados.

Mckeown e Radev (1995) exploram também as relações entre informações de vários textos com a ferramenta SUMMONS. Nesta pesquisa, os autores trabalham sobre uma base de dados de notícias de terrorismo. O sistema, que é focado no tópico, considera *templates* que contêm as principais informações em relação ao tema do terrorismo. Após os *templates* serem preenchidos com as informações da base de dados, estes são relacionados por meio de operadores. Estes operadores representam algumas das relações já revisadas em trabalhos prévios, tais como: *contradiction, equivalence e background.* Muitas destas relações foram a base das relações usadas na CST.

Mani e Bloedorn (1997) também exploram a análise relacional multidocumento, identificando similaridades e diferenças, tais como equivalências e contradições entre vários documentos, relações que posteriormente são utilizadas pela CST. Os autores propõem um método de sumarização usando uma representação de grafo para modelar os textos e as relações de similaridades e diferenças entre eles. Logo, mediante uma função, percorre-se cada documento para encontrar os nós relacionados semanticamente ao tópico. Após ter identificado esses nós, gera-se um novo grafo, mapeando similaridades e diferenças dos nós identificados de acordo com a similaridade semântica. Finalmente se gera o resumo multidocumento em língua natural.

A partir destes estudos, a CST pretende estender varias das tipologias apresentadas e generalizá-las para qualquer domínio de textos que se conheça. A CST é introduzida a seguir.

2.2. A teoria CST

Radev (2000) propôs o modelo multidocumento CST com um conjunto inicial de 24 relações para o tratamento multidocumento. No Quadro 1 a seguir, listam-se as 24 relações originais e suas respectivas descrições. Os nomes das relações são mantidos em inglês, como na obra original.

Quadro 1:Conjunto original de relações CST

Identity	O mesmo segmento aparece em mais de um lugar
Equivalence (paraphrasing)	Dois segmentos contêm a mesma informação
Translation	Dois segmentos com a mesma informação, mas em diferente língua.
Subsumption	Um segmento contém mais informação do que o outro
Contradiction	Informação contraditória
Historical background	Informação extra que coloca a informação recente num contexto dado
Cross-Reference	A mesma entidade é mencionada
Citation	Um segmento cita informação de outro documento
Modality	Versão qualificada de um segmento
Attribution	Um segmento repete a mesma informação que outro, mas adiciona algum atributo
Summary	Um segmento é um resumo de outro
Follow-up	Informação adicional que reflete que algum evento relacionado aconteceu
Elaboration	Informação adicional que não estava incluída no evento anterior
Indirect-Speech	Refere, de forma indireta, a outra informação
Refinement	Informação adicional mais específica
Agreement	Um segmento concorda com a informação de outro segmento
Judgment	Uma versão qualificada sobre um evento
Fulfilment	Uma previsão torna-se verdade
Description	Descreve-se um evento
Reader profile	Estilo e contexto referente ao público ao qual é dirigido o texto
Contrast	Faz um contraste entre dois eventos
Parallel	Compara dois eventos
Generalization	Generaliza alguma informação
Change of perspective	A mesma fonte apresenta a informação, mas com uma perspectiva diferente

As relações estabelecidas podem ser simétricas (sem direcionalidade), ou assimétricas (com direcionalidade). A relação *Equivalence* é um exemplo de relação simétrica, enquanto a relação *Histórical background* é assimétrica, pois um segmento fornece o cenário histórico para outro e, portanto, pode ser menos importante. Como parte da teoria, Radev (2000) propõe uma taxonomia em que qualquer unidade de informação pode ser considerada na análise e no estabelecimento de relações CST.

Podem-se relacionar proposições expressas por palavras, expressões multipalavra, sintagmas, orações, sentenças, parágrafos ou blocos de textos maiores. Esse esquema geral de relacionamento é ilustrado na Figura 6. Pode-se perceber que o resultado da análise CST é um grafo, ou seja, um conjunto de elementos/nós relacionados, sem restrição quanto à forma de relacionamento.

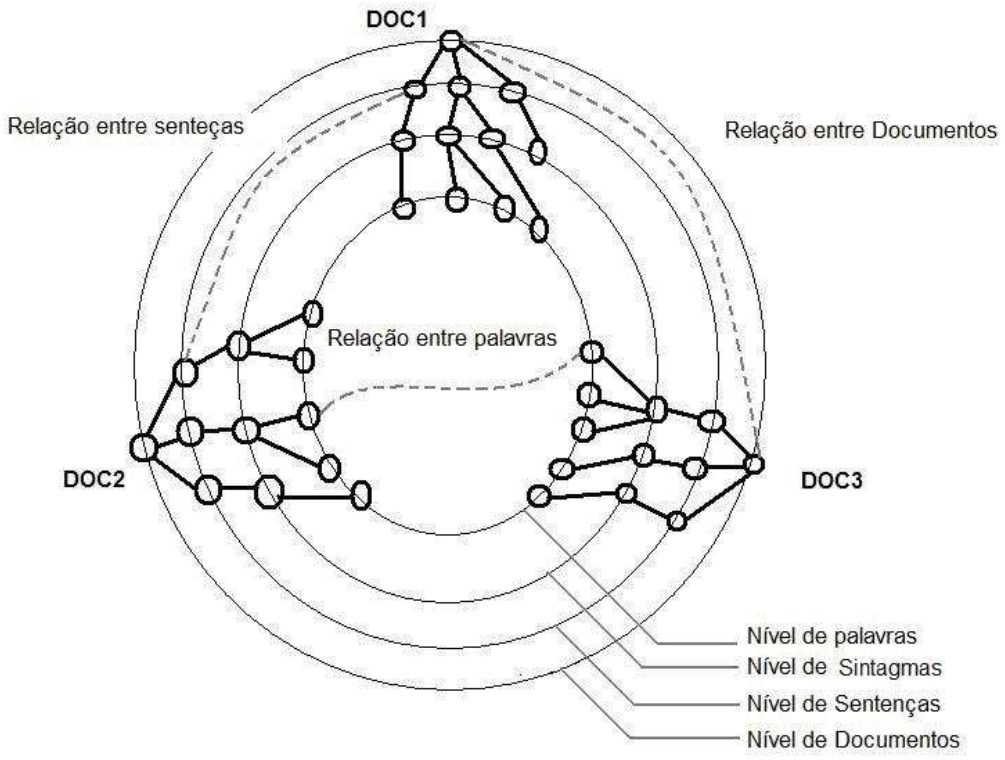

Figura 6: Ilustração dos relacionamentos CST (Radev , 2000, p. 5)

De acordo com a figura, os documentos similares são representados numa hierarquia de palavras, sintagmas, sentenças e os próprios documentos. Em cada nível desta hierarquia, podem se estabelecer relações CST. São considerados todos esses níveis, pois, apesar de orações e sentenças serem tradicionalmente os segmentos mais utilizados, tarefas particulares podem exigir um relacionamento entre unidades menores. Por exemplo, para a fusão de informações, o relacionamento de sintagmas pode ser mais adequado do que orações ou sentenças. Também se deve considerar que nem todas as unidades informativas contidas nos textos têm relações CST com outras unidades, portanto é possível que o grafo seja desconexo.

É importante dizer que existe ambigüidade na análise CST, assim como em qualquer análise mais subjetiva. Analisadores humanos diferentes podem identificar diferentes relações entre os mesmos segmentos ou, ainda, podem selecionar segmentos diferentes

para relacionar. Veja, por exemplo, os dois trechos de texto extraídos do córpus CSTBank (Radev et al., 2003) e apresentados na Figura 7 onde, ao relacionar duas sentenças, os anotadores encontraram três relações diferentes, devido à ambigüidade das relações propostas pela teoria.

Trecho de texto 1: *"The crash put a hole in the 25th floor of the Pirelli building, and smoke was seen pouring from the opening."*
Trecho de texto 2: *"(ABCNEWS.com) 8212 A small plane crashed into the 25th floor of a skyscraper in downtown Milan today."*
Relações encontradas: **- Anotador 1:** *Follow-up* **- Anotador 2:** *Elaboration* **- Anotador 3:** *Subsumption*

Figura 7: Exemplo de discordância na anotação de relações CST

Neste sentido, algumas pesquisas têm encontrado algumas deficiências na teoria CST. Por exemplo, Afantenos et al. (2004) e Afantenos (2007) consideram que há uma baixa concordância na anotação das relações dada a subjetividade das relações. Afantenos et al. também consideram que nem todas as relações são utilizadas e que somente segmentos do mesmo nível na taxonomia podem ser conectados por meio das relações (sentenças com sentenças, palavras com palavras, etc.). Segundo o critério dos autores, isso se deve ao fato de que a CST se inspira na RST (citada no Capítulo 1), sem levar em conta que no caso monodocumento se assume a coerência no texto. No caso multidocumento, têm-se diversos estilos de redação e não faz sentido se falar em coerência. Uma outra crítica de Afantenos et al. é que a CST deveria se focar num tópico em particular, para evitar a generalização das relações.

No trabalho de Zhang et al. (2002), os pesquisadores também concluíram que houve um baixo nível de concordância na anotação de relações. Eles argumentam que isso se deve às ambigüidades de algumas relações e à subjetividade própria da língua. Por exemplo, as relações *elaboration* e *refinement* têm definições similares, o que geraria ambigüidade e reduziria a concordância entre os anotadores. Para reduzir a ambiguidade, Zhang e colaboradores propuseram um refinamento do conjunto de relações do modelo original, considerando somente 18 relações. As relações consideradas similares na sua definição foram reduzidas a uma relação só, por exemplo, as relações *equivalence* e *paraphrase* têm definições similares, portanto *paraphrase* foi eliminada e *equivalence* permaneceu.

Existem vários recursos e ferramentas para o modelo CST, sendo que na seção seguinte são explicados alguns destes.

2.3. Recursos e ferramentas para CST

Com o surgimento da CST, algumas ferramentas e recursos têm sido desenvolvidos, e alguns outros ainda estão em desenvolvimento. Um dos primeiros recursos desenvolvidos para pesquisas com CST foi o CSTBank (Radev et al., 2003). Este é um córpus de documentos na língua inglesa composto por 6 coleções de textos jornalísticos, onde cada coleção tem em media 8 textos sobre um mesmo tópico. O córpus está anotado manualmente com relações CST, sendo que esta anotação foi feita por oito juízes. O nível de concordância na anotação foi baixo. Esta concordância foi calculada usando a medida kappa (Siegel e Castellan, 1988). No caso da anotação do CSTBank, o nível de concordância kappa foi de 0,4, sendo que boas concordâncias são observadas a partir de 0,6. Os juízes argumentaram que, devido à existência de múltiplas relações que podem conectar um mesmo par de sentenças, é difícil que todos os anotadores cheguem a um mesmo critério.

Para o português do Brasil, foi criado o córpus CSTNews (Aleixo e Pardo, 2008a), o primeiro córpus anotado com base no modelo CST para esta língua. Os textos que compõem o córpus foram extraídos manualmente dos jornais online: Folha de São Paulo, O Estado de São Paulo, O Globo, Jornal do Brasil e Gazeta do Povo. O córpus é formado por 50 coleções de textos e cada uma tem em média 3 documentos que versam sobre um mesmo assunto. A seguir, no Quadro 2, mostram-se as estatísticas do córpus, extraídas na íntegra do trabalho de Aleixo e Pardo (2008a).

Quadro 2: Estatísticas gerais do córpus CSTNews.

Coleção	Categoria	Número de documentos	Número de sentenças	Número de palavras
1	Mundo	3	24	432
2	Política	4	78	1.405
3	Cotidiano	4	143	2.864
4	Cotidiano	3	40	846
5	Cotidiano	2	24	574
6	Cotidiano	4	50	1.253
7	Ciência	2	23	587
8	Esportes	3	24	600
9	Política	4	64	1.543
10	Mundo	5	79	1.987

11	Cotidiano	5	128	2.320
12	Mundo	3	34	974
13	Mundo	3	37	962
14	Mundo	4	54	1.402
15	Mundo	4	43	986
16	Política	3	43	1.033
17	Política	2	49	965
18	Mundo	3	74	1.301
19	Esportes	2	13	299
20	Política	5	120	2.516
21	Política	3	41	870
22	Cotidiano	5	127	2.300
23	Mundo	2	25	572
24	Esportes	4	52	1.091
25	Esportes	5	159	2.788
26	Mundo	5	116	2.621
27	Esportes	5	181	2.985
28	Esportes	5	70	1.336
29	Mundo	3	48	1.167
30	Dinheiro	3	46	1.136
31	Esportes	2	10	217
32	Mundo	4	112	2.354
33	Cotidiano	5	131	2.803
34	Cotidiano	3	60	1.139
35	Mundo	5	90	1.976
36	Cotidiano	4	124	2.134
37	Cotidiano	2	27	475
38	Esportes	5	79	1.470
39	Cotidiano	4	54	1.324
40	Política	5	73	1.881
41	Esportes	5	109	1.945
42	Política	2	40	1.075
43	Política	5	141	1.643
44	Política	2	28	737
45	Cotidiano	3	50	1.226
46	Mundo	5	78	1.519
47	Mundo	5	99	2.753
48	Esportes	2	43	800
49	Cotidiano	5	69	575
50	Política	4	108	2.388
Total		185	3.534	72.149
Média		3,7	70,68	1.442,98

Para a anotação deste córpus, Aleixo e Pardo (2008a) consideram 14 relações das 24 propostas pela CST. De fato, para a anotação, as relações foram refinadas a partir do conjunto já refinado proposto por Zhang (2002), onde só foram consideradas 18 das

relações propostas originalmente por Radev (2000). As relações que foram mantidas são: *elaboration*, *overlap*, *subsumption*, *historical background*, *attribution*, *equivalence, follow up, contradiction*, *summary*, *identity*, *modality*, *indirect speech*, *citation* e *translation.*

O refinamento das relações foi feito porque algumas delas foram consideradas similares na sua definição, o que causa mais ambigüidade ainda, e outras nunca foram observadas. Particularmente, foram 4 as relações eliminadas: *description*, *fulfillment*, *reader profile* e *change of perspective*. *Description* foi considerada como uma forma de *elaboration*, pois, ao igual que *elaboration*, ela considera que, dadas duas sentenças, uma delas traz mais detalhes e informação de alguma entidade da outra sentença. Com o mesmo critério, a relação *fulfillment* foi considerada uma forma de *follow up*. *Reader profile* e *change of perspective* foram consideradas desnecessárias.

A seguir, no Quadro 3, mostram-se as estatísticas da freqüência das relações no córpus, extraídas na íntegra do relatório do córpus (Aleixo e Pardo, 2008a).

Quadro 3:Freqüência das relações no córpus

Relações	**Freqüência no Córpus**
Elaboration	23,98%
Overlap	19,85%
Subsumption	15,24%
Background	6,49%
Atributtion	5,68%
Equivalence	5,09%
Follow up	4,72%
Contradiction	4,35%
Summary	4,35%
Identity	3,69%
Modality	3,54%
Indirect Speech	2,73%
Citation	0,29%

A concordância das anotações para o córpus foi calculada com a medida Kappa, obtendo, em geral, uma baixa concordância por parte dos anotadores, como aconteceu para a língua inglesa. A medida foi calculada para algumas das relações encontradas no córpus. A seguir, no Quadro 4, são apresentadas as medidas de concordância para algumas relações, segundo o próprio relatório do córpus (Aleixo e Pardo, 2008a).

Quadro 4:Concordância entre os anotadores e as relações

Relação	Kappa
Elaboration	0,321
Overlap	0,562
Subsumption	0,006
Follow-up	0,009
Summary	0,003
Indirect Speech	0,013
Não há relação	0,279
Média	0,258

Junto com o CSTNews, foi desenvolvida a CSTTool (Aleixo e Pardo, 2008b). Esta é uma ferramenta semi-automática para a anotação de textos com relações CST. Esta ferramenta, que ainda está em desenvolvimento (no NILC – Núcleo Interinstitucional de Lingüística Computacional), segmenta os textos extraindo sentenças para serem analisadas. Em seguida, a similaridade lexical entre as sentenças é calculada para determinar quais sentenças são candidatas a ter relações CST. Por ultimo, as relações CST são atribuídas pelo usuário às sentenças lexicalmente similares (anotação humana). A arquitetura da CSTTool é proposta no relatório da ferramenta (Aleixo e Pardo 2008b) e mostrada na Figura 8, a seguir.

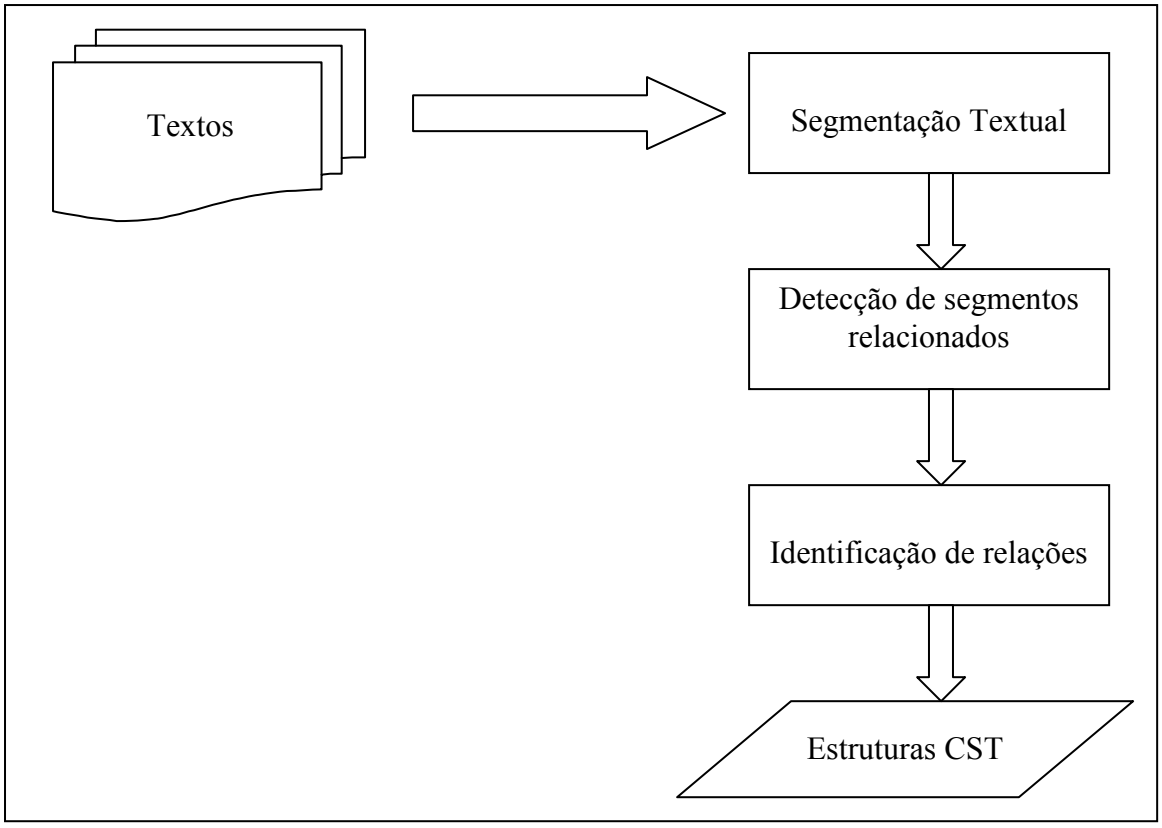

Figura 8: Arquitetura da *CSTTool*

Na primeira etapa, a segmentação automática é feita pelo SENTER (Pardo, 2006), que é um segmentador automático desenvolvido no NILC. Na Figura 9, pode-se observar um exemplo de segmentação de sentenças com a ferramenta *CSTTool.*

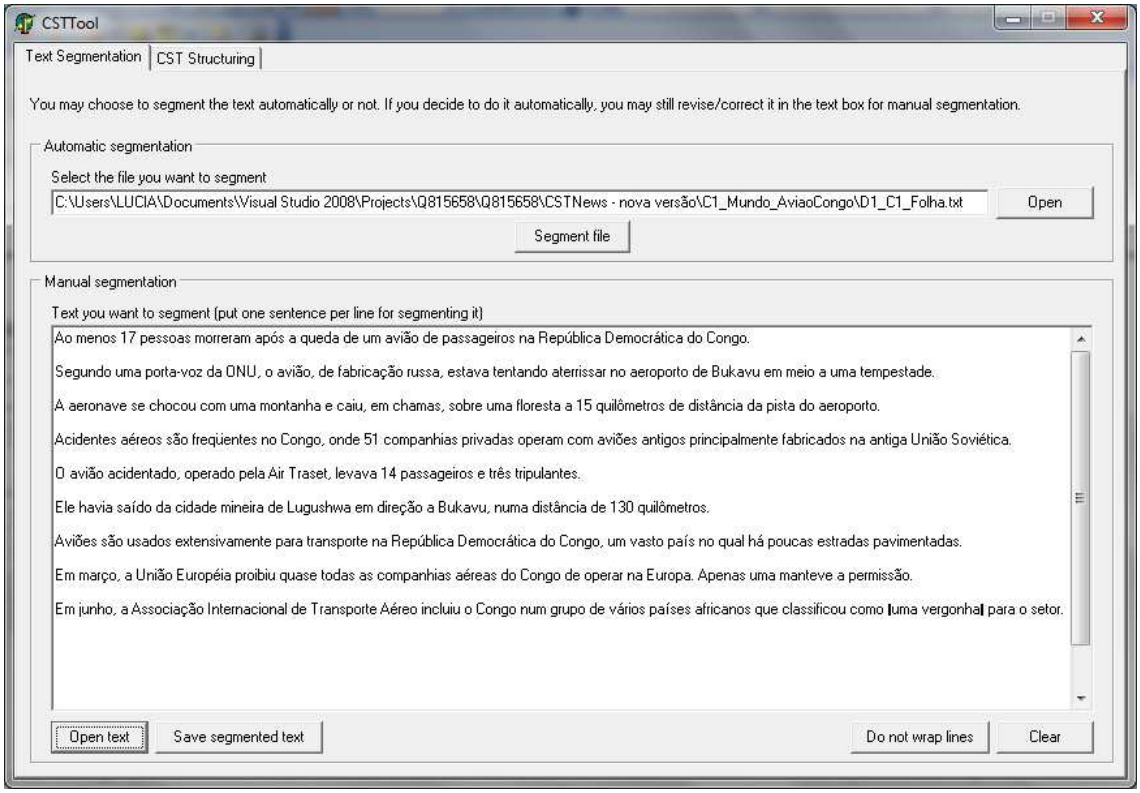

Figura 9: Segmentação de sentenças com *CSTTool*

A ferramenta permite abrir qualquer arquivo de texto e depois segmentá-lo com o botão Segment file. Também permite guardar o texto segmentado. Na segunda etapa da arquitetura, a ferramenta CSTTool identifica automaticamente pares de sentenças que são candidatas a serem relacionadas por meio de uma relação CST. Isso é necessário porque nem todas as sentenças dos textos abordam tópicos relacionados. Além disso, tentar estabelecer relações em todo o conjunto de possíveis pares de sentenças torna a anotação humana um processo muito trabalhoso (Radev et al., 2008).

Aleixo e Pardo (2008c) investigam algumas das medidas de similaridade lexical para encontrar sentenças relacionadas num conjunto de textos sobre um mesmo tópico. Algumas das medidas exploradas são *WordOverlap* e a medida do Coseno.

A medida de similaridade lexical usada na CSTTool é *WordOverlap*, que é a mesma utilizada no CSTBank. Radev et al. (2008) afirmam que essa medida é a mais

eficiente dentre muitas outras. Esta medida é expressa pela seguinte fórmula, como descrita no trabalho de Aleixo e Pardo (2008b):

Wol (S1, S2) = #Palavras em comum / (#Palavras (S1) + #Palavras (S2))

O resultado da fórmula figura são valores entre 0 e 1. Quanto mais próximo de 1, mais relacionado é o par de segmentos. Na Figura 10 mostra-se o modulo para anotação de relações na CSTTool.

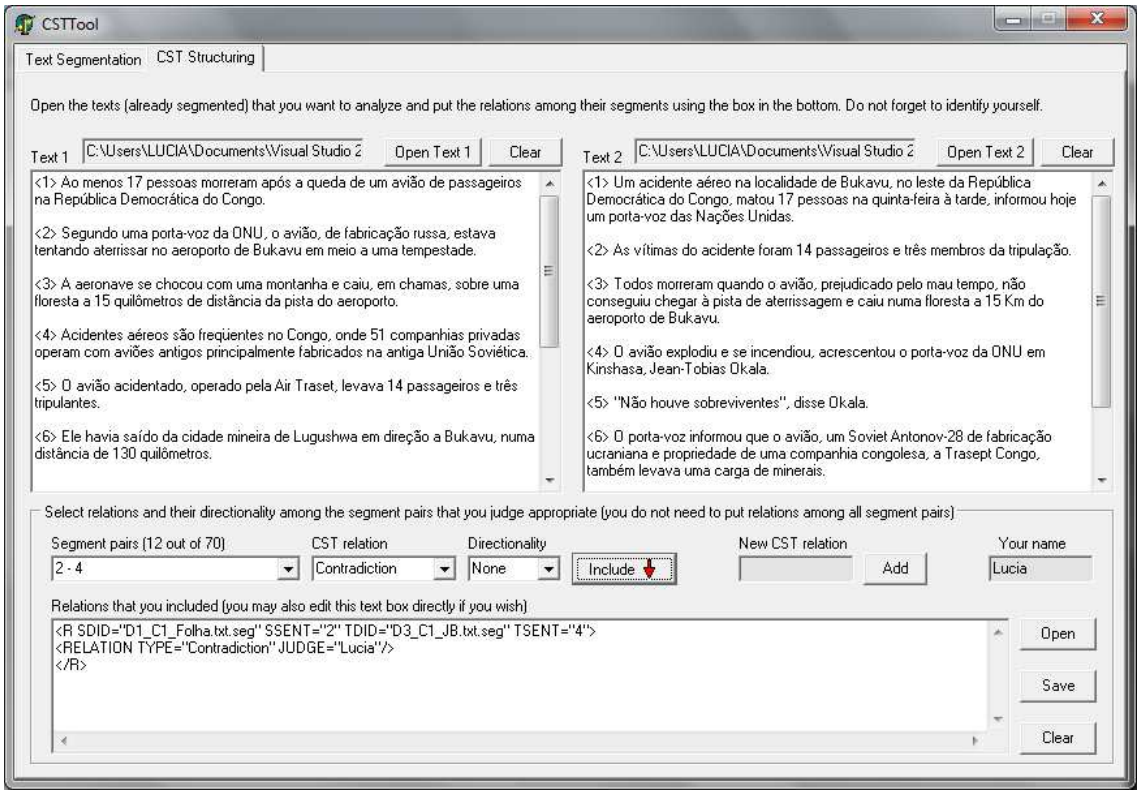

Figura 10: Seleção de sentenças candidatas usando *CSTTool*

A ferramenta permite que, a partir dos pares de sentenças candidatas, o usuário determine as relações CST (se houver) para cada par de sentenças candidatas, as quais estão enumeradas na caixa de texto principal, do módulo de anotação da ferramenta. O objeto gerado depois da anotação é um arquivo XML que contém todas as informações da anotação feita. Este arquivo segue o formato XML utilizado no CSTBank (Radev et al., 2003). Este formato pode ser observado na caixa inferior da tela mostrada na Figura 10. Os códigos “SDID” e “TDID” indicam dados como o número de documento, número de coleção e a fonte daquele documento. A diferença entre ambos é que

“SDID” refere-se ao documento de onde parte a relação CST e “TDID” refere-se ao documento em que incide a relação CST. Por exemplo, na parte inferior da Figura 10, mostra-se um trecho de texto em código XML indicando a anotação correspondente a um par de sentenças do texto. O código contém a seguinte informação:

<R SDID-“ D1_C1_Folha.txt” SSENT-“2” TDID-“D3_C1_JB.txt” TSENT-4>
<RELATION TYPE= “Contradiction” JUDGE= “Lucia”/>
</R>

Neste exemplo, o código SDID-“D1_C1_Folha.txt”, indica o documento 1, da coleção 1, extraído da Folha de São Paulo. Do mesmo modo o código correspondente a TDID indica o documento 3, da coleção 1, extraído do Jornal do Brasil. O código SDID indica o documento origem da relação e TDID o documento de destino da relação. Quando a relação não tem direcionalidade, a definição de SDID e TDID é arbitrária. Os códigos “SSENT” e “TSENT” indicam as sentenças relacionadas aos documentos referidos por “SDID” e “TDID” respectivamente, o seja às sentenças origem e destino da relação. “RELATION TYPE” indica a relação CST que vincula as duas sentenças e “JUDGE” indica ao nome do juiz que anotou a relação CST correspondente. O exemplo completo da figura pode ser lido da seguinte forma: a sentença 2 do documento 1 da coleção 1 tem uma relação de Contradiction com a sentença 4 do documento 3 da coleção 1.

Para se ter uma idéia mais clara do relacionamento entre textos por meio de relações CST, no Apêndice A é mostrado um exemplo completo de relacionamento entre dois textos pequenos do córpus CSTNews.

A seguir, no Capítulo 3, são apresentados os conceitos básicos sobre sumarização multidocumento e as metodologias relacionadas empregadas nesta tarefa.

Capítulo

3

3. Sumarização automática multidocumento

3.1. Conceitos básicos

O objetivo da sumarização automática multidocumento é condensar a informação mais importante de um conjunto de textos e apresentar essa informação de forma coerente e coesa ao usuário (Mani, 2001). Para isto, os principais desafios da sumarização multidocumento descritos no Capítulo 1 devem ser tratados: eliminação da redundância, tratamento de informações complementares e contraditórias e ordenação das sentenças, dentre outros.

Neste contexto, existem vários fatores que influenciam o processo de sumarização. Um dos fatores mais importantes a considerar é a taxa de compressão do sumário final. A taxa de compressão indica o tamanho do sumário final em relação ao tamanho dos textos de entrada, em termos de número de palavras. Este tamanho pode abranger entre 0% e 100%, o que significa que um sumário pode ser tão grande quanto à própria entrada ou tão pequeno que não inclua informação nenhuma. Isto representa um limitante importante ao processo de sumarização já que pode influenciar negativamente na qualidade do sumário, pois, se o sumário ficar muito pequeno, pode não ser tão informativo, ou, se ficar muito grande, perde sua utilidade. Em ambos os casos, não atinge o objetivo da sumarização automática (Nenkova et al., 2008).

Outros fatores importantes do processo de sumarização multidocumento são:

- A língua, sendo que o sumarizador pode ser monolíngüe, se processa documentos em uma só língua, ou multilíngüe, se processa documentos em várias línguas;

- A função do sumário final, podendo ser indicativo (referencia partes dos textos de entrada que possam ser de interesse do usuário) ou informativo (elabora um sumário que contém o conteúdo principal, dispensando a leitura do(s) texto(s)-fonte);
- O tipo de sumário de saída, sendo que pode ser um extrato (as sentenças que comporão o sumário final são as mesmas dos textos de entrada) ou um abstrato (reescreve-se a informação para apresentá-la no sumário final).

Além dos fatores mencionados, a sumarização multidocumento também é influenciada pelas abordagens da sumarização automática mencionadas no Capítulo 1. Numa visão profunda do tratamento multidocumento, procura-se identificar similaridades e diferenças semânticas entre os textos, usando conhecimento lingüístico. Numa visão superficial, faz-se uso de métodos estatísticos.

O modelo CST, revisado no capítulo anterior, é uma das técnicas utilizadas na abordagem profunda. Com a proposta deste modelo pretende-se guiar o tratamento de múltiplos documentos. A seguir, apresentam-se os principais trabalhos relacionados de sumarização multidocumento. Os que aplicam CST são de especial interesse para este trabalho.

3.2. Métodos de sumarização multidocumento

Várias pesquisas têm utilizado a teoria CST para fins de sumarização multidocumento. Por exemplo, Radev (2000), além de propor a CST, também propôs uma metodologia de sumarização a partir de um conjunto de documentos anotados de acordo com a teoria.

O processo de sumarização proposto por Radev (2000) consiste em quatro etapas importantes. A primeira delas é a etapa de agrupamento, onde os textos similares são agrupados por conteúdo. Na segunda etapa, realiza-se uma análise interna dos textos, levando em conta a estrutura das sentenças, sintagmas ou palavras contidas no texto. Na terceira etapa procede-se ao estabelecimento das relações de CST. Por último, na quarta etapa, extrai-se o sumário final. A Figura 11 a seguir ilustra estas quatro etapas.

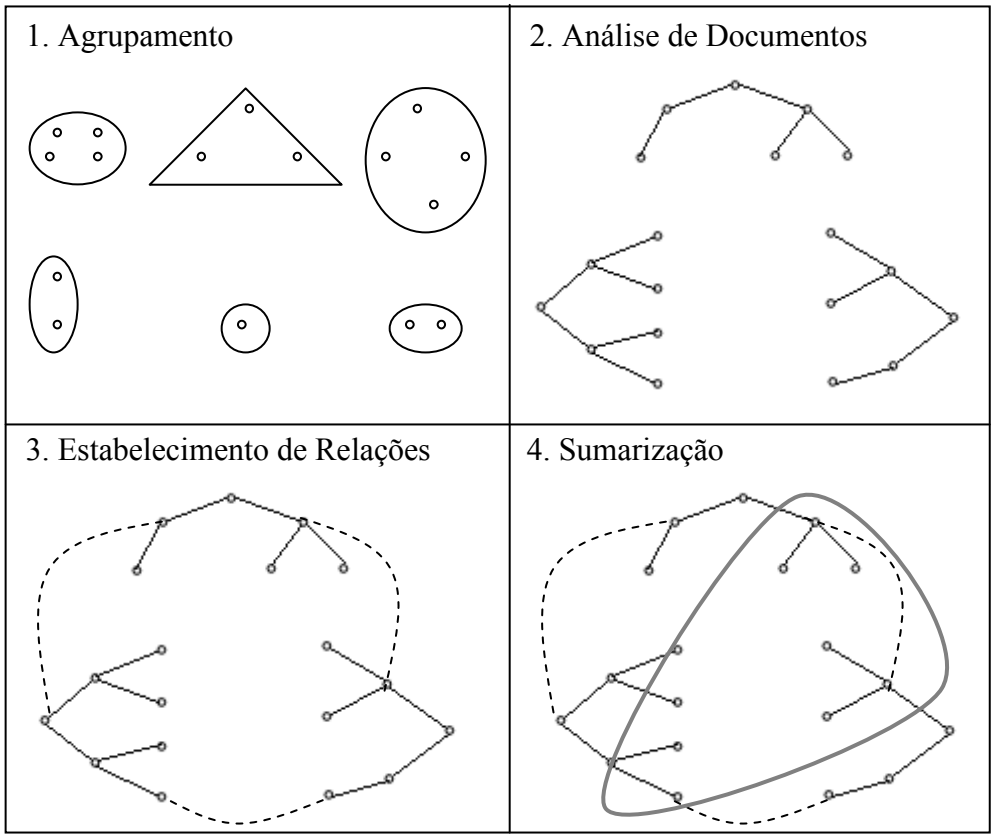

Figura 11: Etapas do processo de sumarização CST proposto por Radev (2000)

O autor sugere a criação de operadores ativados por preferências de sumarização do usuário, que percorram o grafo formado pelos documentos e suas relações e selecionem os segmentos textuais que comporão o sumário final. Exemplos de preferências são: preferência sobre "contradição", que reporta as contradições encontradas nos documentos e suas respectivas fontes, ou preferência sobre "autoria", que reporta informações atribuídas a uma fonte determinada pelo usuário.

Como exemplo, a Figura 12 mostra um sumário em português brasileiro, produzido num dos experimentos realizados neste trabalho de Mestrado, com base na metodologia descrita acima. O sumário foi gerado com uma taxa de compressão de 70% sobre o texto fonte de maior tamanho do grupo.

Um incêndio afetou a central nuclear de Kashiwazaki-Kariwa, situada próxima do epicentro do tremor, causando o vazamento de água com restos de material radioativo, segundo a companhia elétrica.
[Além disso, várias regiões afetadas continuam sem água corrente e sem eletricidade.] Ainda não há outros detalhes sobre danos a pessoas ou ao meio ambiente.
Nesta terça, o ministro da Defesa enviou 450 soldados para as áreas devastadas. O hipocentro do terremoto se localizou 17 km abaixo do nível do mar.
[Milhares de residências ficaram sem energia elétrica e sem água durante todo o dia.] Aquele foi o terremoto mais devastador no Japão, um dos países mais propensos a terremotos no mundo, desde o de Kobe, com magnitude 7,3, que matou mais de 6.400 pessoas em 1995.
O violento fenômeno foi sentido em quase toda a ilha principal de Honshu e sacudiu os arranha-céus de Tóquio, ao mesmo tempo em que o serviço de trens de alta velocidade Shinkansen foi interrompido.

Figura 12: Sumário multidocumento com representação de grafos proposta por Radev (2000)

Para a criação deste sumário não foi considerada ainda nenhuma preferência do usuário. O critério usado para selecionar a informação foi a quantidade de arestas (maior quantidade de relações CST) dos nós. Segundo (Mani, 2001) nós com maior quantidade de arestas contêm informações mais relevantes já que são mais elaboradas pois têm mais relações com outras partes dos textos. Os segmentos textuais, representados pelos nós são ranqueados de acordo com este critério e, os segmentos melhores pontuados são incluídos no sumário final.

Radev et al. (2000, 2001a, 2001b) apresentaram o sumarizador automático multidocumento MEAD. O sistema produz sumários a partir de um conjunto de textos sobre um mesmo tópico extraindo as sentenças mais importantes deles. O cálculo da importância (pontuação) de uma sentença baseia-se na sua distância em relação ao centróide do conjunto de textos (ou seja, o conjunto de palavras mais representativas do assunto tratado no conjunto de textos), na distância em relação ao início do documento (pode ser o título) e sua distância lexical em relação à primeira sentença do documento. Diversos parâmetros são levados em consideração no processo. Dentre eles, destaca-se o parâmetro que indica o máximo de informação em comum que as sentenças que comporão o sumário podem ter. Tal parâmetro visa identificar e tratar adequadamente redundância, simulando a relação *Subsumption* da CST. A idéia é que, se uma sentença X subsume uma sentença Y, então Y pode ser excluída do sumário, pois X já contém sua informação. Os autores ainda apresentam duas variações do MEAD: os sistemas NewsInEssence (Radev et al., 2001b) e WebInEssence (Radev et al., 2001c). Ambos os sistemas recuperam informação da Web de acordo com a consulta realizada pelo

usuário. Os textos recuperados formam um grupo e, a partir dele, é elaborado o sumário multidocumento, extraindo-se as sentenças mais relevantes de acordo com os critérios do MEAD. A diferença entre os processos é que o primeiro sistema, NewsInEssence, seleciona os centróides e elabora o resumo a partir de todo o grupo gerado pela consulta. No entanto, o WebInEssence seleciona os centróides e gera o resumo somente a partir dos documentos selecionados pelo usuário de acordo com o critério do próprio usuário.

Zhang et al. (2002) propõem o enriquecimento de sumários multidocumento produzidos automaticamente pelo uso de relações CST. Basicamente, a metodologia apresentada pelos autores consiste primeiramente em identificar sentenças do sumário que tenham uma baixa pontuação (pontuação esta determinada por qualquer método de sumarização que tenha sido utilizado, por exemplo, o método do sistema MEAD) e não seja relacionada a nenhuma outra sentença do sumário via alguma relação CST. Depois se substitui tais sentenças identificadas por sentenças que apresentem alguma relação CST em particular com alguma sentença do sumário, ou sentenças que têm alto número de relações CST. Os autores executaram alguns experimentos e demonstraram que: os sumários cujas sentenças são mais relacionadas (pelas relações CST) são melhores em relação à informatividade do sumário final; as inclusões de sentenças relacionadas por diferentes relações CST afetam de formas diversas a qualidade do sumário, sendo que algumas relações melhoram (por exemplo, as relações *Equivalence* e *Subsumption*) e outras pioram a qualidade (por exemplo, as relações *Historical background* e *Description*) do sumário; não faz diferença se a relação CST considerada é simétrica ou não.

Otterbacher et al. (2002) conduzem um estudo sobre os problemas de coesão em sumários multidocumento e sobre como sobrepujá-los utilizando-se teorias como a CST. A entrada do sistema é uma coleção de documentos sobre um mesmo assunto. A partir disso, as sentenças mais relevantes dos textos são selecionadas. O passo seguinte é identificar relações CST entre essas sentenças. As sentenças com relações entre elas aparecem juntas no sumário final e podem ser reordenadas de acordo com a ordem temporal que indicam as relações.

Com base na CST, Afantenos et al. (2004) discutem como determinar de forma mais adequada as relações multidocumento e como usá-las para a sumarização focada em interesses dos usuários. Como foi visto no capítulo anterior, diferentemente da CST, os autores dividem suas relações em dois tipos: sincrônicas e diacrônicas. O sistema de sumarização que propõem consiste de um componente de extração de informação (a

partir de um conjunto de textos, extrai os fatos/eventos que serão sumarizados), um componente para detectar as relações entre os fatos/eventos e um componente de geração textual (que seleciona a informação relevante e produz o sumário final). No componente de extração de informação, o autor faz uso de ontologias relacionadas ao tópico para extrair informações dos textos em formato de mensagens. Cada mensagem tem parâmetros de entrada que representam elementos da ontologia. Uma vez que as mensagens são extraídas, os tipos de relações entre as mensagens são identificados e estabelecidos. Finalmente, esta informação é utilizada pelo componente de geração textual. A Figura 13 ilustra um exemplo extraído do estudo de caso de Afantenos et al. (2004). O exemplo mostra duas mensagens baseadas numa ontologia para o domínio de futebol e extraídas de duas fontes de informação distintas. Os parâmetros das mensagens contêm informações do tipo: nome da equipe (ou jogador), posição, o tempo de jogo e o desempenho. Em seguida as mensagens são vinculadas por meio da relação sincrônica *Contradiction*, pois têm informação contraditória referente ao desempenho. Os exemplos foram traduzidos do inglês.

Fonte de Informação 1: F1.1: *performance* (Georgeas, Geral, Tempo_1, **Excelente**). **Fonte de Informação 2:** F2.1: *performance* (Georgeas, Geral),Tempo_1, **Ruim**).
Relação Sincrônica: *Contradiction* (F1.1, F2.1)

Figura 13: Exemplo de construção de relações a partir de mensagens

Há vários outros trabalhos importantes em sumarização multidocumento que não lidam diretamente com a CST. Mani e Bloedorn (1997) constroem grafos do conjunto de documentos a serem sumarizados e identificam relações entre grupos de palavras apenas, indicando similaridades e diferenças para determinar os grupos mais salientes para a produção de sumários. A metodologia do sistema consiste em três etapas: análise, refinamento e síntese. Na etapa de análise, uma representação com base em grafos é feita. Basicamente, cada nó do grafo representa uma palavra em diferentes posições. Logo, a informação extraída é comparada com a consulta feita pelo usuário, para assim determinar quais são as informações (nós) mais similares à consulta, o que indicaria a sua importância. Na etapa de refinamento, exploram-se as similaridades e diferenças

entre as porções informativas mais importantes (de acordo com a etapa anterior) de cada texto. Para calcular as similaridades e diferenças, utilizam-se medidas de similaridade que incluem relações de coesão. As relações de coesão são aquelas que relacionam sinônimos, antônimos, repetições, adjacências e co-referências. Finalmente as similaridades e diferenças encontradas são exploradas para a construção do sumário final.

Um trabalho prévio a CST que deu as bases para o desenvolvimento dessa teoria é o proposto por Radev e McKeown (1998). Eles propuseram o sumarizador multidocumento chamado SUMMONS (*Summarizing Online News Articles*) para sumarização de artigos jornalísticos que versam sobre um mesmo evento. Os dados de entrada para o sistema consistem em um conjunto de informações salientes em cada artigo organizadas em *templates*, ou seja, informações na forma atributo-valor. O sistema opera sobre o domínio de terrorismo. Os *templates* são preenchidos com atributos do tipo: vítimas, lugar do crime, etc. No total, cada *template* tem 25 campos. O sumarizador aplica uma série de operadores sobre os *templates*, atribuindo-lhes pontos de acordo com sua importância (em função da repetição de informação em diferentes *templates* e do tipo de informação que contêm), combinando-os em novos *templates* mais genéricos ou específicos (dependendo do operador aplicado), excluindo *templates*, etc. Ao final, os *templates* mais bem pontuados são selecionados e, com base nos dados contidos neles, o sistema produz o sumário final utilizando técnicas de geração de língua natural.

Os operadores aplicados sobre os *templates* são de diversos tipos e são disparados por heurísticas que identificam as relações entre os *templates*. O operador de "contradição", por exemplo, é aplicado se a informação factual sobre o número de mortos em um atentado é diferente em 2 *templates* sendo processados; o resultado desse operador pode ser a criação de um novo *template* contendo a contradição com suas respectivas fontes jornalísticas. Outros operadores detectam relações de *refinement*, *elaboration* e *agreemen*t, dentre várias outras. Os autores indicam como uma limitação de seu trabalho a ausência de informação sobre conexões diretas entre artigos, limitação esta que pode ser suprida pela CST.

Carbonell e Goldstein (1998) definem a medida *Maximal Marginal Relevance* para detectar e tratar redundância na sumarização multidocumento. Esta técnica mede o grau de similaridade entre os diferentes elementos extraídos do conjunto de documentos. Assim, quando dois elementos são muito similares, de acordo com uma medida de

similaridade dada, a informação é redundante e pode ser removida. Posteriormente surgiram algumas melhorias para essa técnica, como a proposta por Goldstein et al. (2000), que, além de medir o grau de similaridade, adiciona um fator que determina quão boa é a cobertura de um elemento extraído em relação aos tópicos dos textos.

Barzilay et al. (1999) produzem sumários multidocumento utilizando técnicas de detecção de intersecção temática entre sentenças. Numa primeira fase, a similaridade entre sentenças é calculada. Depois, as sentenças que foram julgadas similares são contrastadas para encontrar a intercessão delas, ou seja, encontrar os trechos que contêm informação comum ao conjunto dessas sentenças, descartando as diferenças. A informação comum é utilizada para gerar uma nova sentença.

Barzilay et al. (2001) atacam o problema da ordenação de sentenças na produção de sumários multidocumento. Basicamente, o algoritmo que eles propõem consiste, numa primeira etapa, em identificar as unidades que têm a mesma informação ou informação relacionada. Estas unidades se juntam para formar tópicos ou temas, considerando que os temas podem conter informação repetida. Na segunda etapa, dado um par de tópicos, se duas sentenças de cada tópico aparecerem num mesmo texto e num mesmo segmento do texto, então se considera que há um alto grau de relacionamento temático. A partir desta informação pode-se medir o grau de similaridade entre dois tópicos de acordo com o número de sentenças que têm as propriedades mencionadas entre os tópicos. A terceira etapa agrupa em blocos os tópicos que são similares. O objetivo final do algoritmo é reordenar primeiro os blocos e, em seguida, os tópicos dentro de cada bloco e, por último, as sentenças dentro de cada tópico.

McKeown et al. (2002) apresentam o sistema NewsBlaster de sumarização multidocumento, o qual se caracteriza por unificar diversas ferramentas de PLN para o tratamento do problema. Neste trabalho, o objetivo é fazer um sistema multilíngüe para sumarização multidocumento. Primeiramente, a ferramenta extrai textos em várias línguas, especificamente: Inglês, Russo e Japonês. Para esta primeira etapa, eles utilizam um extrator de textos multilíngüe, desenvolvido como parte desse trabalho. Numa segunda fase, um tradutor automático é utilizado para traduzir todos os textos a uma mesma língua: o Inglês. Numa terceira fase, é feito o processo de agrupamento dos textos. Por último, as sentenças dos textos são extraídas, e a similaridade semântica é calculada. As sentenças mais similares à consulta principal são apresentadas no sumário final.

Lin e Hovy (2002) e Leuski et al. (2003) apresentam os sistemas de sumarização multidocumento NeATS e iNeATS, respectivamente, que abordam o relacionamento multidocumento apenas indiretamente. NeATS é um sistema de extração que tem três componentes básicos. O primeiro componente identifica os conceitos principais da coleção de documentos. Estes conceitos são identificados em unigramas, bigramas e trigramas e agrupados para identificar tópicos e sub-tópicos. O segundo componente filtra a informação da coleção com base em três filtros: posição das sentenças, palavras inúteis e parâmetros de redundância. As sentenças mais relevantes em relação aos conceitos principais são preservadas. O terceiro e último componente justapõe as sentenças selecionadas seguindo a ordem cronológica dos eventos. O iNeATS se baseia no NeATS, mas permite de uma forma mais interativa maior controle do usuário sobre o processo de sumarização. O usuário tem a possibilidade de controlar parâmetros como o tamanho do sumário e os elementos textuais que devem ser considerados na elaboração do sumário. Também se pode ter um acesso visual aos documentos, para saber a quais documentos pertencem as sentenças do sumário. Pode-se clicar numa sentença do sumário e o link nela leva o usuário até o documento de onde foi extraída.

Mihalcea e Tarau (2005) apresentam uma técnica de sumarização multidocumento baseada em percurso em grafos, independente de língua e com resultados bons. Os autores estudam dois algoritmos baseados em grafos: o algoritmo de Kleinberg e o algoritmo *PageRank* de Google. O primeiro utiliza só as arestas de entrada num nó. A importância reside no número de arestas que apontam pra ele. O algoritmo *PageRank* considera arestas de entrada e saída, e faz um ponderado. Para o processo de sumarização, as sentenças dos textos representam os nós, e os links entre os textos são as arestas. Os autores validam sua técnica para a língua portuguesa, inclusive.

Farzindar et al.(2005) apresentam o sistema CATS (*Cats is an Answering Text Summarizer*), que é um sistema de sumarização multidocumento que gera sumários automáticos a partir de requerimentos feitos pelo usuário, incluindo o grau de detalhe que ele deseja no sumário. O sistema se divide em cinco etapas de processamento. A primeira etapa consiste numa análise da consulta feita pelo usuário e das sentenças dos textos de entrada. Nesta etapa são identificadas as entidades nomeadas, expressões temporais e tópicos relevantes na consulta do usuário. Também nesta etapa, as sentenças da consulta e do córpus são segmentadas em elementos básicos para que logo possam ser comparadas com facilidade. Na segunda etapa, os elementos dos textos e da consulta são comparados e é dada uma pontuação para cada sentença, de acordo com o grau de

similaridade com os elementos da consulta. A similaridade é dada por cálculos simples de freqüência de palavras e pela medida do coseno (Salton, 1989). A terceira etapa corresponde ao pós-processamento, onde elementos irrelevantes são eliminados. Na quarta etapa, as sentenças mais importantes são selecionadas de acordo com a pontuação obtida por elas na segunda etapa. Finalmente, na quinta etapa do processo, o sumário é elaborado de acordo com as informações selecionadas. O sistema foi avaliado na data da DUC 2005 e os resultados mostram que em termos de qualidade e informatividade o sistema está entre os 5 melhores dos 32 participantes da DUC.

Para o Português do Brasil, foi desenvolvido o sumarizador GistSumm (Pardo 2005). O sistema é baseado numa abordagem superficial e simples para sumarização mono e multidocumento. O GistSumm consiste de três tarefas principais: identificação de sentenças, ranqueamento de sentenças e produção do extrato. Primeiramente, as sentenças são identificadas por meio de sinais de pontuação, como os pontos finais. Depois disso, as sentenças são pré-processadas para remover *stopwords* e fazer o processo de *stemming*. Posteriormente, a importância das palavras é calculada, processo que pode ser feito por meio dos métodos *Keywords* (Luhn, 1958) ou TF-ISF *(Term frequency - Inverse Sentence frequency*) (Larroca et al., 2000). De acordo com essas medidas, as sentenças são ranqueadas. A primeira sentença é considerada a mais relevante em relação ao tópico. Finalmente, o resto das sentenças que comporão o sumário final são selecionadas com base em dois critérios: a sentença deve conter ao menos uma palavra ou raiz lexical que corresponda a alguma palavra contida na sentença principal; a sentença deve ter uma pontuação maior do que a média das pontuações das outras sentenças. A quantidade de sentenças que serão incluídas no sumário final depende da taxa de compressão. Para fazer a sumarização multidocumento, o GistSumm une todos os textos em um só arquivo e aplica a metodologia descrita acima.

Wan e Yang (2006) propõem uma metodologia de sumarização multidocumento com base em grafos de afinidade, que são grafos que representam os textos e as relações entre as unidades informativas (sentenças) dos textos. No grafo os nós representam as unidades informativas e as arestas, não direcionadas, representam links semânticos entre as sentenças, sendo que os links podem ser inter-sentenciais ou intra-sentenciais. Os autores fazem esta diferenciação para dar maior ênfase aos links inter-sentenciais ao se calcular a informatividade de uma sentença. A metodologia é composta de três etapas principais: na primeira etapa se constrói o grafo identificando as relações entre as

sentenças, o que é feito por meio da medida de similaridade coseno (Salton, 1989). Além de usar esta medida, também é aplicado um processo de difusão que consiste em calcular a similaridade semântica entre duas sentenças que não estão diretamente relacionadas, mas estão relacionadas a sentenças intermediárias em comum. Este processo de difusão é feito por meio do cálculo da somatória dos valores de similaridade semântica de todos os nós que, no grafo, estão entre duas sentenças i e j. Após se ter montado o grafo, a informatividade de cada sentença é calculada com base em três critérios: 1) quanto mais vizinhos uma sentença tem no grafo, mais informativa ela é; 2) quanto mais informativos são os vizinhos de uma sentença, mais informativa ela é; 3) quanto mais alto é o valor da conexão semântica de uma sentença com outras sentenças informativas do grafo, mais informativa ela é. A partir destes critérios, a informatividade para cada sentença é calculada, sendo que as sentenças que têm links inter-sentenciais terão maior importância ao se calcular a informatividade. Finalmente, as sentenças com maiores pontuações são as selecionadas para compor o sumário final. O sistema foi avaliado com os dados da DUC 2002 e da DUC 2004, obtendo melhores resultados na informatividade em comparação com outros sistemas que também foram avaliados com as mesmas bases de dados.

No capítulo seguinte detalha-se o trabalho realizado neste Mestrado.

Capítulo 4

4. Seleção de conteúdo com base no modelo CST

Seguindo a linha de abordagem profunda para a produção de sumários multidocumento, nesta dissertação foram investigados e desenvolvidos métodos de seleção de conteúdo relevante para a tarefa de sumarização automática com base no modelo CST. A investigação contempla a análise do efeito das relações do modelo CST na informatividade e coerência do conteúdo extraído dos textos. Em particular, foram explorados os seguintes fatores de coerência da sumarização multidocumento: redundância, tratamento de informação complementar e tratamento de informação contraditória.

4.1. Objetivos, hipóteses e metodologia de desenvolvimento

O objetivo principal desta investigação foi explorar o uso das relações CST em textos escritos em português do Brasil e o seu impacto na informatividade e qualidade de sumários multidocumento. Para atingir este objetivo, alguns métodos propostos na literatura foram formalizados e outros apenas explorados. Em particular, foram formalizados seis operadores de seleção de conteúdo com base no trabalho de Radev (2000).

Além de desenvolver operadores de seleção de conteúdo, também foi explorada a metodologia proposta por Zhang et al. (2002), usando como base dois sumarizadores superficiais já estudados na literatura: o MEAD (Radev, 2001) e o GistSumm (Pardo et al., 2005). O MEAD é um dos sumarizadores multilíngüe mais conhecidos e utilizados, e o GistSumm é o único sumarizador multidocumento para o português do Brasil que se conhece.

Os resultados dos métodos explorados confirmam as hipóteses apresentadas neste trabalho. Em particular, foram confirmadas as seguintes hipóteses:

- a CST permite modelar textos em português brasileiro, estabelecendo relações semânticas e discursivas entre os textos;
- as informações fornecidas pelo modelo CST permitem explorar o conteúdo dos textos e assim identificar as informações mais relevantes ao tópico que está sendo tratado;
- o modelo CST permite identificar redundâncias, contradições e informações complementares, ajudando assim a melhorar a qualidade dos sumários produzidos.
- o uso do modelo CST ajuda a melhorar a informatividade dos sumários produzidos

Neste trabalho, o processo de sumarização é dividido em três tarefas principais: 1) análise dos textos de entrada de acordo com o modelo CST; 2) aplicação de métodos de seleção de conteúdo sobre os textos analisados; 3) organização do conteúdo selecionado para formar o sumário final. Na Figura 14 mostram-se estas três tarefas no contexto geral do processo de sumarização descrito no Capítulo 1.

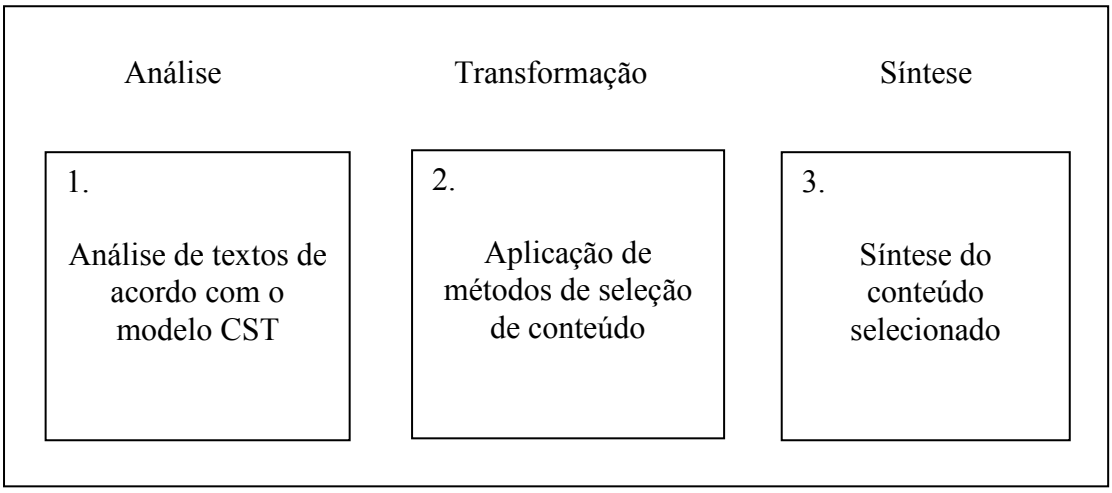

Figura 14: Arquitetura do processo de sumarização com base em CST

Na etapa de análise, as relações do modelo CST foram refinadas a partir das relações consideradas pelo trabalho de Aleixo e Pardo (2008a). De acordo com este novo refinamento, foi realizada uma nova anotação do córpus CSTNews. Na etapa de transformação, foram formalizados e desenvolvidos operadores de seleção de conteúdo, além de serem explorados dois sumarizadores de abordagem superficial (MEAD e

GistSumm). Finalmente, na etapa de síntese, o conteúdo selecionado foi organizado para ser apresentado de forma coerente e coesa.

Na Figura 15 a seguir é ilustrada a arquitetura da metodologia de sumarização desenvolvida no presente projeto de pesquisa.

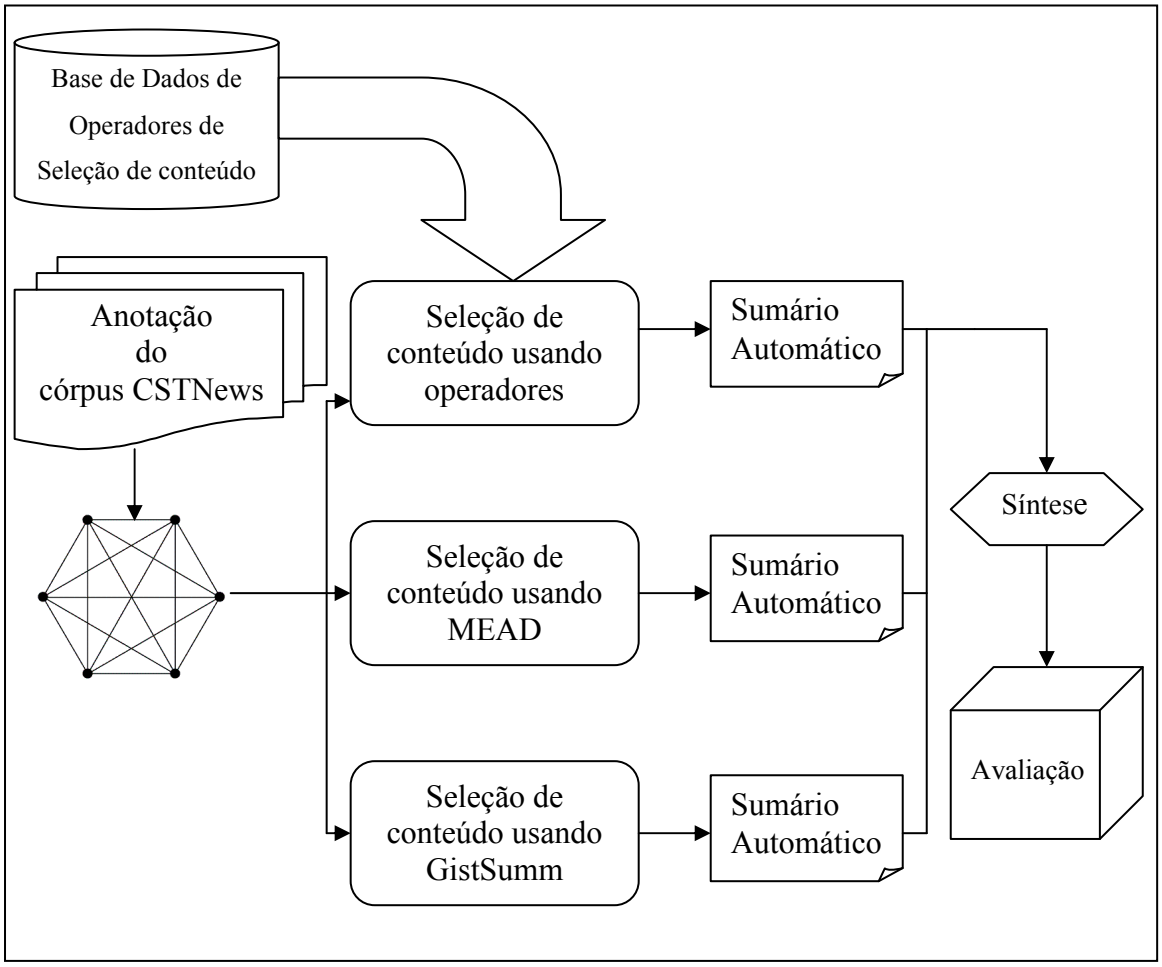

Figura 15: Metodologia de sumarização com base em CST

De acordo com a figura, são considerados três módulos de seleção de conteúdo. O primeiro módulo corresponde aos métodos com operadores de seleção de conteúdo, os quais utilizam as informações contidas na base de dados onde estão especificados e formalizados os operadores definidos neste trabalho. O segundo módulo corresponde ao sumarizador MEAD e à metodologia que incorpora o modelo CST no processo de sumarização do MEAD. O terceiro módulo corresponde ao sumarizador GistSumm e à metodologia que incorpora o modelo CST no processo de sumarização do GistSumm. Cada um destes módulos recebe como entrada o grafo CST dos textos que serão processados, o qual é construído a partir do córpus anotado. Neste grafo, os nós

representam as sentenças do córpus e as arestas representam as relações estabelecidas de acordo com a anotação.

Cada método utilizado gera o respectivo sumário automático, o qual é levado à etapa de síntese. Neste trabalho, a etapa de síntese envolve a tarefa de fusão de sentenças usando o sistema de fusão para o português de Seno e Nunes (2009) e a tarefa de ordenação de sentenças selecionadas. Finalmente, na etapa de avaliação, o sumário pós-processado é avaliado automaticamente pela ferramenta ROUGE (Lin e Hovy, 2003) e também por meio de uma avaliação humana, a qual será descrita na seção 4.6.

A seguir são relatadas a definição e formalização dos operadores de seleção de conteúdo desenvolvidos neste trabalho.

4.2. Anotação do córpus CSTNews

Como foi visto no Capitulo 2, o CSTNews (Aleixo e Pardo, 2008a) é um córpus anotado em português brasileiro composto de 50 coleções de textos jornalísticos, sendo que cada coleção contém em média 3 documentos que versam sobre um mesmo assunto.

Para o presente trabalho se fez uma nova anotação do córpus com base em um novo refinamento das relações do modelo CST, com vistas a se atingir uma maior concordância e melhor formalização e definição das relações desse modelo. Acredita-se que, em qualquer trabalho de PLN, são necessários dados representativos e confiáveis, provenientes de tarefas bem definidas.

Nesta nova anotação colaboraram 4 anotadores durante 3 meses. Os anotadores, com formação em lingüística computacional, foram treinados por um período de 1 mês antes de iniciarem a anotação.

O novo refinamento proposto para este trabalho foi feito considerando como base as 14 relações utilizadas no trabalho de Aleixo e Pardo (2008a), a partir das quais foi proposta uma classificação/tipologia das relações. A Figura 16 a seguir mostra esta tipologia.

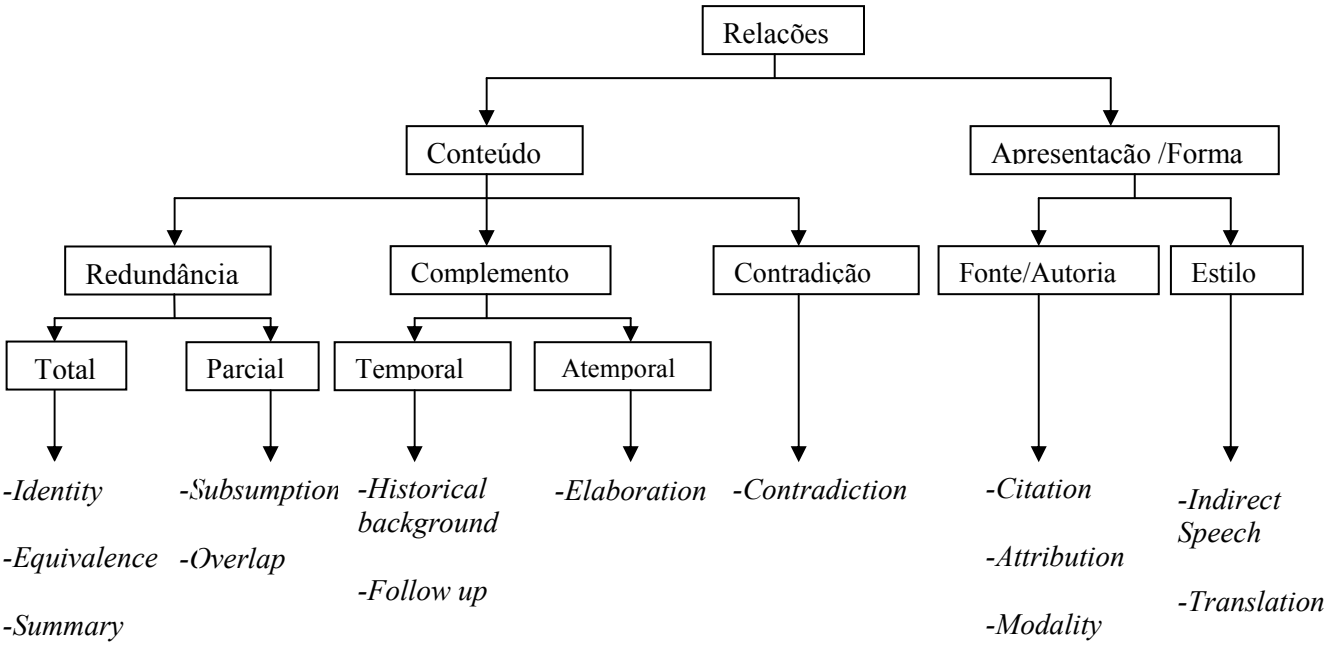

Figura 16: Classificação de relações CST

Pode-se ver que, em seu nível mais alto, a tipologia tem duas subdivisões principais: relações de conteúdo e relações de apresentação/forma. Na categoria de conteúdo, estão as relações que expressam similaridade, contradição ou complementaridade entre as informações textuais, ou seja, relações que de fato se referem ao conteúdo dos segmentos relacionados. Esta categoria se divide, por sua vez, em três subcategorias: redundância, complemento e contradição. Na subcategoria redundância, consideram-se as relações que expressam similaridades parciais ou totais das informações. Por exemplo, as relações *Identity*, *Equivalence* e *Summary* expressam uma similaridade total entre segmentos, já que as informações podem ser idênticas ou equivalentes em conteúdo, enquanto as relações *Overlap* e *Subsumption* indicam somente algumas similaridades entre as informações, já que podem haver informações diferenciadas nos segmentos que estão sendo relacionados. Na subcategoria complemento, estão todas as relações que elaboram as informações principais, quer seja com informações históricas, fatos que dão continuidade a um evento, ou com informação contextual. As relações que indicam informações históricas e de seguimento de um evento (*Historical background* e *Follow up*) são consideradas temporais, enquanto as relações que indicam um contexto de um fato (atual, em geral) são consideradas não temporais. A última subcategoria dentro da categoria conteúdo é Contradição, onde está a relação *Contradiction*, que indica informações contraditórias entre dois segmentos.

A segunda grande categoria na tipologia é de Apresentação/Forma, em que são incluídas todas as relações que lidam com aspectos secundários da informação, como a atribuição de uma informação a um determinado autor ou fonte (*Attribution*, *Citation*), o estilo de escrita e o posicionamento do autor do texto (*Indirect Speech*, *Modality*) e a língua utilizada *(Translation*).

É importante dizer que, de acordo com esta tipologia, mais de uma relação pode acontecer para um mesmo par de unidades informativas se e somente se as relações pertencerem a diferentes categorias. Por exemplo, a relação *Attribution* pode acontecer com qualquer relação que pertence à categoria conteúdo: *Subsumption*, *Overlap*, etc. O que não pode acontecer é que mais de uma relação de uma mesma categoria seja estabelecida para um mesmo par de unidades informativas. Tal decisão ajuda a evitar a ambigüidade e a controlar, na medida do possível, a subjetividade envolvida na anotação textual.

As relações utilizadas nesta anotação são definidas com base em dois atributos principais: direcionalidade e restrições. Dadas duas sentenças S1 e S2, a direcionalidade pode ser nula (S1−S2), à esquerda (S1←S2) ou à direita (S1→S2). As restrições especificam as situações em que deve acontecer a relação indicada. As Figuras 17 e 18 mostram dois exemplos de definições de duas relações CST. Além da direcionalidade e das restrições, também é incluído um exemplo que ajuda a entender melhor a definição em cada caso. A definição de todas as relações consideradas neste trabalho é mostrada no Apêndice A.

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 as informações X e Y,e, S2 contém as informações X e Z.

Figura 17:Definição da relação *overlap* de acordo com o novo refinamento

Exemplo de relação *Overlap* entre duas sentenças:

> **S1.** TOQUIO- Um terremoto de 6.8 graus na escala Richter, com epicentro a 17 quilômetros de profundidade, atingiu a costa noroeste do Japão às 10h13m desta segunda-feira (22h13m de domingo em Brasília).
>
> **S2**. Um forte terremoto matou ao menos cinco pessoas no noroeste do Japão nesta segunda-feira.

No exemplo, as sentenças S1 e S2 apresentam informação em comum sobre o terremoto em Japão acontecido na segunda-feira, mas a sentença S1 apresenta informações particulares sobre a magnitude, lugar de origem e horário de acontecimento do terremoto; a sentença S2 apresenta informações particulares do numero de mortos.

Nome da Relação: *Historical background*
Direcionalidade: S1←S2
Restrições: S2 apresenta informações históricas sobre algum elemento presente em S1.
Comtentários: O elemento elaborado em S1 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (por exemplo, *overlap*).

Figura 18:Definição da relação *historical background* de acordo com o novo refinamento

Exemplo de relação *historical background* entre duas sentenças:

> **S1.** TOQUIO- Um terremoto de 6.8 graus na escala Richter, com epicentro a 17 quilômetros de profundidade, atingiu a costa noroeste do Japão às 10h13m desta segunda-feira (22h13m de domingo em Brasília).
> **S2.** O Japão é um dos países do mundo mais suscetíveis a terremotos, com um tremor ocorrendo ao menos cada cinco minutos.

No exemplo, a sentença S2 apresenta informações históricas sobre terremotos no Japão, o que complementa o foco da sentença S1 que é o terremoto acontecido na segunda-feira.

A partir desta nova anotação, as freqüências observadas das relações no córpus utilizado neste trabalho são mostradas no Quadro 5 a seguir.

Quadro 5: Freqüência das relações no córpus

Relações	Freqüência relações (%)
Overlap	28
Elaboration	21
Follow up	17
Subsumption	13
Historical background	5
Identity	5
Attribution	4
Contradiction	3
Equivalence	2
Indirect speech	1
Summary	0.5
Modality	0.4
Translation	0.1
Citation	0

Para a nova anotação também foi calculada a medida kappa (Fleiss, 1971) das relações, direcionalidade e relações agrupadas (relações que pertencem a uma mesma categoria dentro da nova classificação proposta neste trabalho). A medida kappa, diferente da concordância simples, mede o nível de concordância excluindo a concordância por azar. No Quadro 6 são mostradas as médias dos valores kappa, sendo possível notar que os valores para as relações são melhores que os valores observados em Aleixo e Pardo (2008a).

Quadro 6: Valor kappa das relações, direcionalidade e relações agrupadas

	Média do valor Kappa
Kappa de relações	0.5094
Kappa direcionalidade	0.4459
Kappa de relações agrupadas	0.6141

Além da medida kappa, também foram avaliadas as concordâncias totais, parciais e nulas das relações, direcionalidade e relações agrupadas. A diferencia entre a concordância medida pela kappa e a concordância simples é que a medida kappa exclui a concordância acontecida pelo azar. Os Quadros 7, 8 e 9 mostram os resultados médios da concordância simples, neles estão codificadas as porcentagens de vezes que os anotadores concordaram totalmente (todos indicaram as mesmas relações), parcialmente

(a maioria dos anotadores indicaram as mesmas relações) ou não concordaram (cada anotador indicou relações diferentes).

Quadro 7: Concordância das relações

	Média da concordância (%)
Concordância total relações	54
Concordância parcial relações	28
Concordância nula relações	18

Quadro 8: Concordância da direcionalidade

	Média da concordância (%)
Concordância total direcionalidade	59
Concordância parcial direcionalidade	27
Concordância nula direcionalidade	14

Quadro 9: Concordância de relações agrupadas

	Média da concordância (%)
Concordância total relações agrupadas	70
Concordância parcial direcionalidade	21
Concordância nula direcionalidade	9

Pode-se notar que, para as relações, houve concordância total ou parcial de 82%, o que é muito bom. Para a língua inglesa, Zhang et al. (2002) reportaram que, dadas 88 sentenças onde pelo menos foi identificada uma relação CST, somente houve 57% de concordância parcial ou total das relações. Neste trabalho, para a análise da concordância total se considera que os anotadores devem concordar em todas as relações observadas, enquanto na análise da concordância parcial se considera que os anotadores devem concordar na maioria das relações observadas. Para a direcionalidade houve uma concordância total ou parcial de 86% o que também é muito bom.

Apesar de ter melhorado na concordância, ainda se observaram valores consideráveis para as concordâncias nulas, isto é devido a subjetividade da análise. No caso das relações agrupadas foi observado um 91% de concordância total ou parcial o que mostra que a nova classificação feita ajuda a ter uma idéia mais clara e uniforme das relações.

Com base nos resultados obtidos a partir desta anotação, é confirmada a hipótese de que os textos em português brasileiro podem ser modelados pela CST, ajudando assim a identificar similaridades, diferenças e informações complementares entre os textos.

Na seção seguinte é descrita a formalização e definição dos operadores de seleção de conteúdo desenvolvidos neste trabalho.

4.3. Definição e formalização de operadores de seleção de conteúdo

Formalmente, definimos um operador de seleção de conteúdo como um artefato computacional que processa uma representação de conteúdo previamente fornecida e produz uma versão mais curta contendo as informações mais relevantes segundo os critérios especificados.

Os operadores são aplicados após os textos-fonte terem sido analisados segundo a CST (na etapa de análise). Até o momento, a análise foi realizada manualmente, visto que o analisador discursivo CST para a língua portuguesa ainda está em desenvolvimento.

A partir da análise CST dos textos, é construído automaticamente um grafo com base na anotação em XML dos arquivos correspondentes. A partir do grafo, é, então, produzido um ranque inicial das unidades informativas. Esse ranque é um dos dados de entrada que os operadores de seleção de conteúdo recebem.

O ranque inicial deve conter as unidades informativas do texto na ordem de preferência em que devem ser inseridas no sumário final. Quanto mais relevante for a unidade informativa, mais acima no ranque ela deve estar. A função dos operadores é, a partir do ranque inicial e da preferência do usuário, produzir um ranque refinado, de tal forma que as unidades informativas mais relevantes segundo o critério especificado pelo usuário melhorem no ranque e, portanto, ganhem preferência para estar no sumário. Por fim, dada uma taxa de compressão (ou seja, o tamanho do sumário desejado em relação ao tamanho dos textos-fonte), são selecionadas tantas sentenças do ranque quanto possível para que a taxa seja respeitada.

O ranque inicial é construído considerando todas as unidades informativas contidas no grafo CST. A relevância das unidades informativas depende do número de relações CST que elas apresentam, pois se assume que as informações mais importantes são aquelas que se repetem e são elaboradas ao longo dos textos, apresentando, portanto, mais relações. Tal suposição é padrão na área de SA (Mani, 2001) e, de fato, pode ser facilmente verificada.

Na Figura 19, mostra-se um exemplo hipotético de um grafo CST e o ranque inicial formado a partir deste. As relações CST extraídas do grafo também são incluídas no ranque, não sendo necessário que se consulte o grafo constantemente, portanto. Como

se pode notar, a unidade informativa mais importante é a 4, pois apresenta 3 relações CST, seguida pelas unidades 2 e 1 (que apresentam a mesma quantidade de relações), que, por sua vez, são seguidas pela unidade 5 (com apenas 1 relação), terminando-se na unidade 3 (sem relação alguma). Note que a direcionalidade das relações (indicada pela direção das setas) não tem influência alguma no processo de construção do ranque inicial.

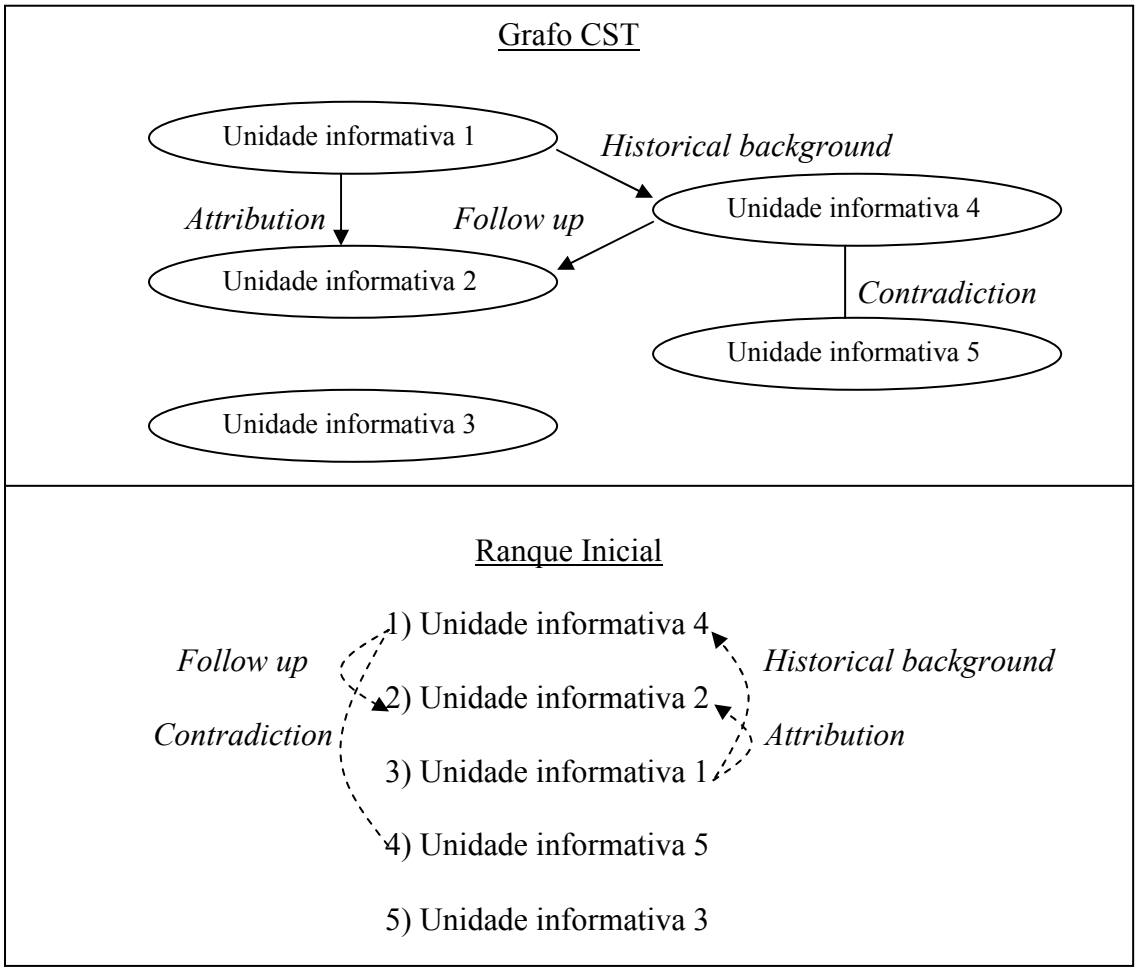

Figura 19: Grafo CST e ranque inicial

Quando algumas unidades apresentam o mesmo número de relações, elas são ranqueadas na ordem em que são lidas do grafo. Neste trabalho, as sentenças são consideradas como unidades informativas.

Os operadores de seleção de conteúdo com base na CST estão definidos em formato de *templates*, contendo um conjunto de regras. As regras são especificadas por meio de condições e restrições, as quais, caso sejam satisfeitas, dispararão funções primitivas de manipulação da informação no ranque. Cada regra é definida da seguinte forma:

CONDIÇÕES, RESTRIÇÕES ⇒ AÇÕES

Cada condição tem o formato CONDIÇÃO(Si, Sj, Direcionalidade, Relação) e uma dada condição é satisfeita se existem a relação e a direcionalidade (de Si até Sj: →; o caso oposto: ←; ou nenhuma direcionalidade: —) especificadas entre duas sentenças Si e Sj. As restrições são opcionais, pois representam possíveis requisitos extras para que o operador seja aplicado. Atualmente, só são consideradas restrições sobre o tamanho das sentenças, como será mostrado mais adiante.

Se todas as condições e restrições forem satisfeitas, então as ações serão aplicadas ao ranque inicial, produzindo assim uma versão refinada do ranque. As ações são definidas em termos de pelo menos uma das três funções primitivas definidas a seguir:

- SOBE(Si,Sj): a sentença j é colocada em uma posição imediatamente após a sentença i no ranque; é importante notar que a sentença i sempre estará em uma posição superior a sentença j no ranque;
- TROCA(Si,Sj): trocam-se as posições das sentenças i e j no ranque;
- ELIMINA(Sj): elimina-se a sentença j do ranque.

Para o presente trabalho, são definidos e formalizados 5 operadores que representam possíveis estratégias de seleção de conteúdo. São eles: apresentação de informação de contexto, exibição de informação contraditória, identificação de autoria, tratamento de redundância, e apresentação de eventos que evoluem com o tempo. O processo de construir o ranque inicial também pode ser representado como um operador, no qual a preferência é pela informação principal. Chama-se este último operador de "operador genérico" ou "operador de informação principal".

Cada operador é definido por três campos: um nome de referência, uma breve descrição e um conjunto de regras. Na Figura 20, mostra-se o operador para apresentação de informação contextual. Nesse operador, procuram-se por pares de sentenças (ao longo do ranque) que apresentem relações CST do tipo *Historical background* e *Elaboration*, já que essas relações são as que fornecem informação contextual. Caso essas informações sejam encontradas, elas sobem no ranque, obtendo, assim, maior preferência para estarem no sumário.

Nome	Apresentação de informação contextual
Descrição	Preferência por informações históricas e complementares
Regras	CONDIÇÃO(S_i, S_j, ←, *Elaboration*) ⇒ SOBE(S_i, S_j) CONDIÇÃO(S_i, S_j, ←, *Historical background*) ⇒ SOBE(S_i, S_j)

Figura 20: Operador de apresentação de informação contextual

A aplicação deste operador ao ranque inicial da Figura 19 irá produzir o ranque refinado da Figura 21, na qual também se exibe o ranque inicial (para facilitar a comparação). É possível notar que a informação histórica da unidade informativa 1 sobe de posição no ranque, sendo posicionada imediatamente depois da sentença a qual se refere.

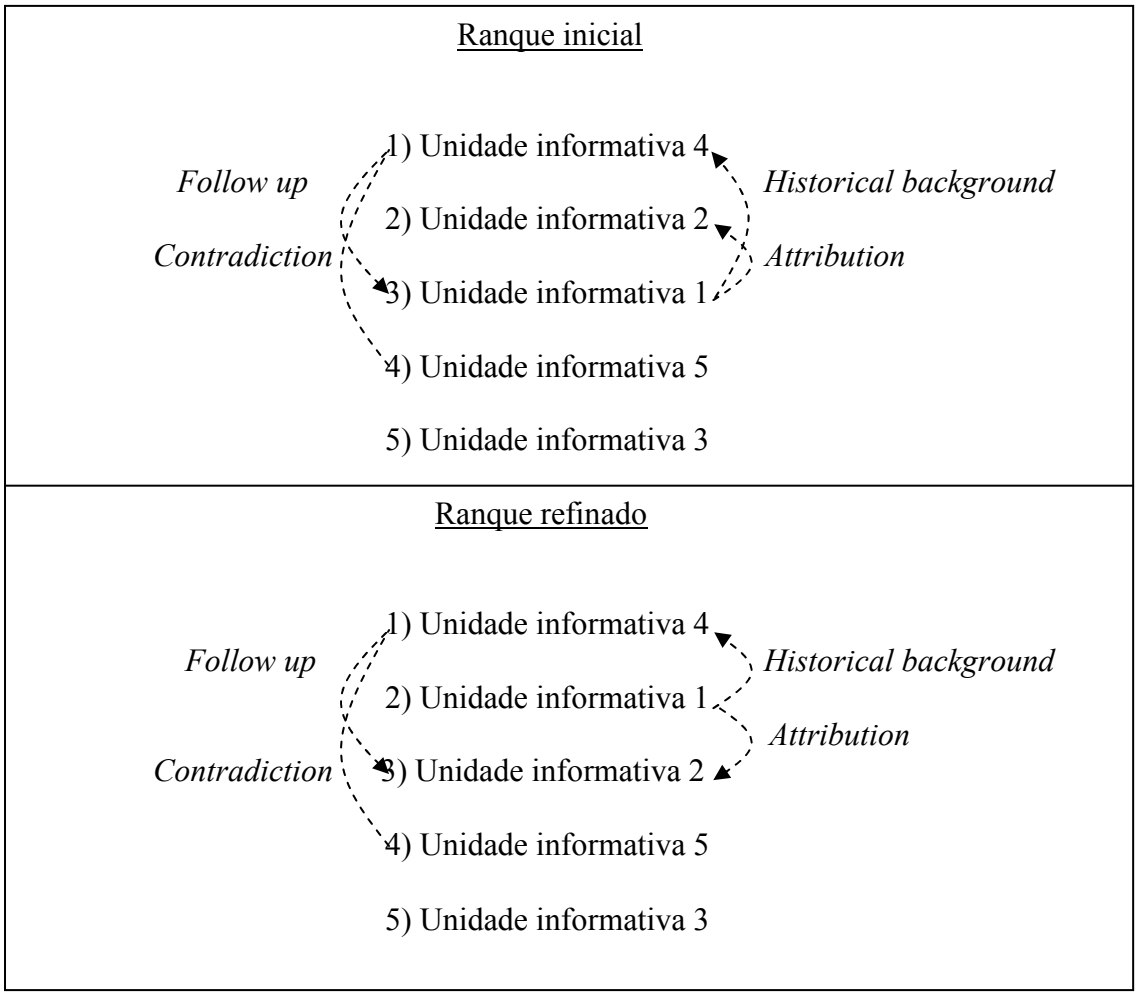

Figura 21: Ranque inicial e ranque refinado a partir do operador de apresentação de informação contextual

Na Figura 22 a seguir, é ilustrado um exemplo de sumário multidocumento gerado automaticamente usando o operador de apresentação de informação contextual. Como pode ser observado na figura, a segunda e a terceira sentença (grifadas) contêm informação contextual e histórica, respectivamente, em relação à primeira sentença. De fato, pode-se notar que a segunda sentença é redundante em relação à primeira, já que nenhum tratamento de redundância está sendo feito. Para resolver esse problema, faz-se necessário aplicar o operador de tratamento de redundância, detalhado posteriormente.

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. O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito. A maior tragédia ferroviária da história do Egito ocorreu em fevereiro de 2002, após o incêndio de um trem que cobria o trajeto entre Cairo e Luxor (sul), lotado de passageiros, e que deixou 376 mortos, segundo números oficiais.

Figura 22: Sumário gerado pelo operador de apresentação de informação contextual

O próximo operador prioriza a evolução de um evento no tempo. Esta evolução é modelada na CST por meio das relações *Historical background* e *Follow-up*. A Figura 23 mostra o operador correspondente. A forma de interpretação deste operador é a mesma do operador anterior. É interessante notar que a direcionalidade não importa neste caso, já que a preferência sobre eventos que evoluem no tempo pretende informar sobre todos os fatos sobre um evento dado, incluindo tanto os elementos elaborados quanto os que elaboram outros fatos. Contrariamente, no caso do operador de contexto, somente é preciso informar sobre os dados de contexto. No operador de apresentação de eventos que evoluem no tempo todas as regras devem ser respeitadas para todas as possíveis direcionalidades.

Nome	Apresentação de eventos que evoluem no tempo
Descrição	Preferência por informações sobre eventos que evoluem no tempo
Regras	CONDIÇÃO(S_i, S_j, $\leftarrow$, *Historical background*) $\Rightarrow$ SOBE(S_i, S_j) CONDIÇÃO(S_i, S_j, $\rightarrow$, *Historical background*) $\Rightarrow$ SOBE(S_i, S_j) CONDIÇÃO(S_i, S_j, $\leftarrow$, *Follow-up*) $\Rightarrow$ SOBE(S_i, S_j) CONDIÇÃO(S_i, S_j, $\rightarrow$, *Follow-up*) $\Rightarrow$ SOBE(S_i, S_j)

Figura 23: Operador de apresentação de eventos que evoluem no tempo

A Figura 24 mostra um sumário produzido pelo uso desse operador sobre um conjunto de três textos do córpus CSTNews. Pode-se notar que a segunda sentença

(grifada) contém informação sobre fato anterior ao fato narrado na primeira sentença, foco dos textos-fonte.

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**.** Pouco antes Thiago Pereira já havia conquistado a segunda medalha de ouro brasileira no dia na final dos 400m medley, superando o norte-americano Robert Margalis e o canadense Keith Beavers.

Figura 24: Sumário gerado pelo operador de apresentação de eventos que evoluem no tempo

A Figura 25 mostra o operador para exibir informações contraditórias, as quais são expressas por meio da relação *Contradiction*, enquanto a Figura 26 mostra o operador para identificação de autoria, expressadas pelas relações *Attribution* e *Citation*. Pode-se perceber que as regras deste último operador contêm mais de uma condição, sendo que todas elas devem ser satisfeitas para que o operador seja aplicado. Este caso em particular se deve ao fato de que as relações *Attribution* e *Citation* sempre envolvem a presença de alguma outra relação, neste caso, a relação de conteúdo *Subsumption*. Também neste operador deveria incluir-se a relação Overlap, pois esta relação também pode aparecer junto com as relações *Attribution* e *Citation*. Em particular, a relação *Overlap* precisa de um tratamento mais complexo para eliminar a redundância nas sentenças, pois é necessário juntar todas as informações sem repetição. Este processo requer métodos de processamento de língua que ainda não são parte da tarefa de seleção de conteúdo deste trabalho. Por isso, este processo é feito posteriormente à tarefa de seleção de conteúdo, particularmente, na etapa de síntese.

Nome	Exibição de informações contraditórias
Descrição	Preferência por informações contraditórias
Regras	CONDIÇÃO(S_i, S_j, —, *Contradiction*) ⇒ SOBE(S_i, S_j)

Figura 25: Operador de exibição de informações contraditórias

Nome	Identificação de autoria
Descrição	Preferência por informações atribuídas a uma fonte
Regras	CONDIÇÃO(S_i, S_j, ←, *Attribution*), CONDIÇÃO(S_i, S_j, ←, *Subsumption*) ⇒ TROCA(S_i,S_j), ELIMINA(S_i) CONDIÇÃO(S_i, S_j, ←, *Citation*), CONDIÇÃO(S_i, S_j, ←, *Subsumption*) ⇒ TROCA(S_i,S_j), ELIMINA(S_i)

Figura 26: Operador de identificação de autoria

As Figuras 27 e 28 mostram sumários produzidos pelos operadores de informações contraditórias e identificação de autoria respectivamente. A informação privilegiada está grifada.

Cairo - 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. No entanto, o ministro da Saúde, Hatem El-Gabaly, insistiu que até o momento foram recuperados apenas 36 cadáveres e que 133 feridos foram encaminhados a hospitais da região. 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.

Figura 27: Sumário gerado pelo operador de exibição de informações contraditórias

Quinze voluntários da ONG francesa Ação Contra a Fome (ACF) foram assassinados no nordeste do Sri Lanka, informou hoje um porta-voz da organização. 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".

Figura 28: Sumário gerado pelo operador de identificação de autoria

Pode-se notar no sumário da Figura 27 que as 2 últimas sentenças apresentam informações contraditórias entre si e também em relação a primeira sentença. A contradição, neste caso, tem origem da narração da notícia em momentos diferentes, quando números mais precisos vão surgindo conforme a passagem do tempo. Também é possível notar que o sumário produzido não é coerente nem coeso, pois as sentenças contraditórias são justapostas de forma simples. No sumário da Figura 28, a segunda sentença apresenta o nome do diretor de uma organização, atribuindo a ele algumas informações ditas.

Finalmente, o operador de tratamento de redundância é mostrado na Figura 29. Em particular, neste operador, também são definidas algumas restrições em relação ao cumprimento das unidades informativas (representado pelas barras verticais | |). Como a relação *Equivalence* indica que duas sentenças têm o mesmo conteúdo, elimina-se a sentença maior, mantendo-se a menor no sumário.

Nome	Tratamento de redundâncias
Descrição	Preferência por informações não redundantes
Regras	CONDIÇÃO(S_i, S_j, —, *Identity*) ⇒ ELIMINA(S_j) CONDIÇÃO(S_i, S_j, —, *Equivalence*), $\|S_i\| \leq \|S_j\|$ ⇒ ELIMINA(S_j) CONDIÇÃO(S_i, S_j, —, *Equivalence*), $\|S_i\| > \|S_j\|$ ⇒ TROCA(S_i,S_j), ELIMINA(S_i) CONDIÇÃO(S_i, S_j, ←, *Subsumption*) ⇒ TROCA(S_i,S_j), ELIMINA(S_i) CONDIÇÃO(S_i, S_j, →, *Subsumption*) ⇒ ELIMINA(S_j)

Figura 29: Operador de tratamento de redundâncias

É desejável que o operador de tratamento de redundâncias seja aplicado antes de qualquer outro operador, pois ele evita que conteúdo redundante seja incluído nos sumários.

No operador de redundância também deveria ser incluída uma regra para o tratamento de redundância quando há relação de *Overlap* entre duas sentenças. Este processo é feito na etapa de síntese, após a tarefa de seleção de conteúdo pois requer um tratamento especial. Nesta etapa, é aplicado o sistema de fusão de sentenças de Seno e Nunes (2009) o qual gera uma única sentença a partir de um conjunto de sentenças originais. A fusão é baseada em um processo de alinhamento de informações comuns (por exemplo, sinônimos e paráfrases) entre as sentenças de um conjunto. O alinhamento é responsável por guiar o processo de fusão na identificação das informações que devem ser preservadas na produção da nova sentença. Uma vez identificadas essas informações, o modelo faz uso de regras para combinar essas informações numa sentença de saída.

Este sistema possui duas técnicas de fusão: fusão por intersecção e fusão por união. A fusão por intersecção considera as informações comuns entre as sentenças de entrada e gera uma única sentença com essas informações, enquanto a fusão por união considera todas as informações (comuns e não comuns) das sentenças de entrada e gera uma única sentença contendo todas essas informações. Para este trabalho somente foi usada a técnica de fusão por união já que não pode-se perder nenhuma informação após a seleção de conteúdo.

Além da fusão de sentenças, na etapa de síntese também é feita a ordenação das sentenças que comporão o sumário final. Esta ordenação é feita com base num critério simples, considerando somente a posição das sentenças nos textos originais. Por exemplo, dadas as sentença 1 do documento 1 e a sentença 4 do documento 2 e, ambas

sentenças são selecionadas para compor o sumário final, então a sentença 1 será lida antes da sentença 4. Em caso duas sentenças tiverem a mesma posição, a ordem é dada pelo número do documento no córpus. Por exemplo, dadas a sentença 1 do documento 1 e a sentença 1 do documento 2, a sentença 1 do documento 1 será lida antes da sentença 1 do documento 2.

Os operadores estudados nesta seção são armazenados de forma simples em um arquivo XML que pode ser facilmente manipulado, podendo-se adicionar, remover ou alterar operadores de maneira trivial. O formato do arquivo XML usado é mostrado na Figura 30 a seguir:

```
<Operadores>
 <OperadorID OID="Contexto">
            <Descrição> Operador que seleciona informação complementar, histórica ou argumentativa
            sobre as informações  principais do sumário final. </Descrição>
          <Regra REID="1">
                <Condição CID="1">
                  <Direcionalidade>Esq</Direcionalidade>
                  <RelaçãoCST>Elaboration</RelaçãoCST>
                </Condição>
                <Restrição RID="1">Nenhuma</Restrição>
                 <Ação AID="1">SOBE<Parametro_Ação >2</ Parametro_Ação > </Ação>
          </Regra>
          <Regra REID="2">
                < Condição CID="1">
                  < Direcionalidade>Esq</ Direcionalidade>
                  < RelaçãoCST >Historical-background</ RelaçãoCST >
                </Condição>
                < Restrição RID="1">Nenhuma</ Restrição >
                <Ação AID="1">SOBE<Parametro_Ação>2</Parametro_Ação> </Ação>
          </Regra>
        </OperadorID>
</Operadores>
```

Figura 30: Exemplo de operador codificado em formato XML

Pode observar que todas as condições, restrições e ações das regras estão codificadas neste arquivo. O exemplo da Figura 30 mostra a codificação para o operador de apresentação de informações contextuais. As duas regras deste operador estão codificadas com suas respectivas condições, restrições e ações. Estas regras são identificadas pelos valores do parâmetro REID; as condições são identificadas pelo parâmetro CID; as restrições correspondem ao parâmetro RID e as ações ao parâmetro AID. A direcionalidade também é codificada e pode assumir três valores: **Esq, Dir e Nenhuma.** Estes valores podem ser aplicados da seguinte forma: S_i **Valor_Direcionalidade** S_j, assim o valor **Esq** é lido: $S_i \leftarrow S_j$, o valor **Dir** é lido: $S_i \rightarrow S_j$ e **Nenhuma** é lido: $S_i - S_j$. A primeira regra é ativada quando se satisfaz a condição

onde a relação é *Elaboration* e a direcionalidade é $S_i \leftarrow S_j$, isto significa que serão consideradas somente as sentenças que estejam vinculadas pela relação *Elaboration* e, onde S_j elabore a S_i ou seja, a aresta incide da direita para a esquerda.

Se as condições são satisfeitas, avaliam-se logo as restrições. As restrições podem assumir três valores como: **Nenhuma**, **Maior** e **MenorIgual**, estes dois últimos valores somente são usados para o operador de redundância de acordo com a definição dada. O valor **Maior** corresponde à restrição $|S_i|>|S_j|$ e o valor **MenorIgual** corresponde à restrição $|S_i|\leq|S_j|$. No exemplo da Figura 30 o valor da restrição é **Nenhuma** já que para o operador de informação contextual não foram definidas restrições.

Após avaliar condições e restrições, são aplicadas as ações. Os valores das ações no arquivo XML correspondem ao nome das funções definidas neste trabalho: **SOBE**, **ELIMINA** e **TROCA**. Elas podem receber parâmetros, os quais correspondem à sentença sobre a qual é aplicada a função. Neste caso, a ação da regra 1 do operador da Figura 30, é lida da seguinte forma: SOBE (S_j). Se o parâmetro fosse 1 a ação seria aplicada sobre S_i.

A seguir é mostrado o protótipo de sumarização implementado neste trabalho.

4.4. Protótipo

Utilizando os operadores descritos na seção anterior, foi implementado um protótipo do sumarizador. A arquitetura deste protótipo é mostrada na Figura 31.

Inicialmente, tem-se como entrada do protótipo o arquivo XML da anotação do córpus e, a partir deste arquivo, é construído o grafo CST. Logo de ter construído o grafo, o ranque inicial é construído também.

O módulo de seleção de conteúdo recebe como entrada a preferência de sumarização escolhida pelo usuário, o ranque inicial e o arquivo XML dos operadores. A partir disso a tarefa de seleção de conteúdo é realizada e produz como saída um ranque refinado de sentenças. Na Figura 32 é mostrado o algoritmo para seleção de conteúdo com base nos operadores descritos neste trabalho. Logo de ter produzido o ranque refinado, a taxa de compressão é aplicada e as sentenças melhor pontuadas são selecionadas, obtendo assim o conjunto de sentenças que serão incluídas no sumário final, sendo que estas sentenças são armazenadas num arquivo de texto. A taxa de compressão utilizada neste trabalho é de 70% sobre texto de maior tamanho do grupo. Isto foi feito porque os sumários de referencia também foram elaborados com a mesma consideração Este arquivo de sentenças selecionadas é a entrada do processo de fusão de Seno e Nunes (2009) que

produz um novo arquivo de sentenças fundidas. Finalmente, as sentenças produzidas pelo sistema de Seno e Nunes (2009) são ordenadas de acordo com o critério descrito na seção anterior produzindo assim, o arquivo do sumário final.

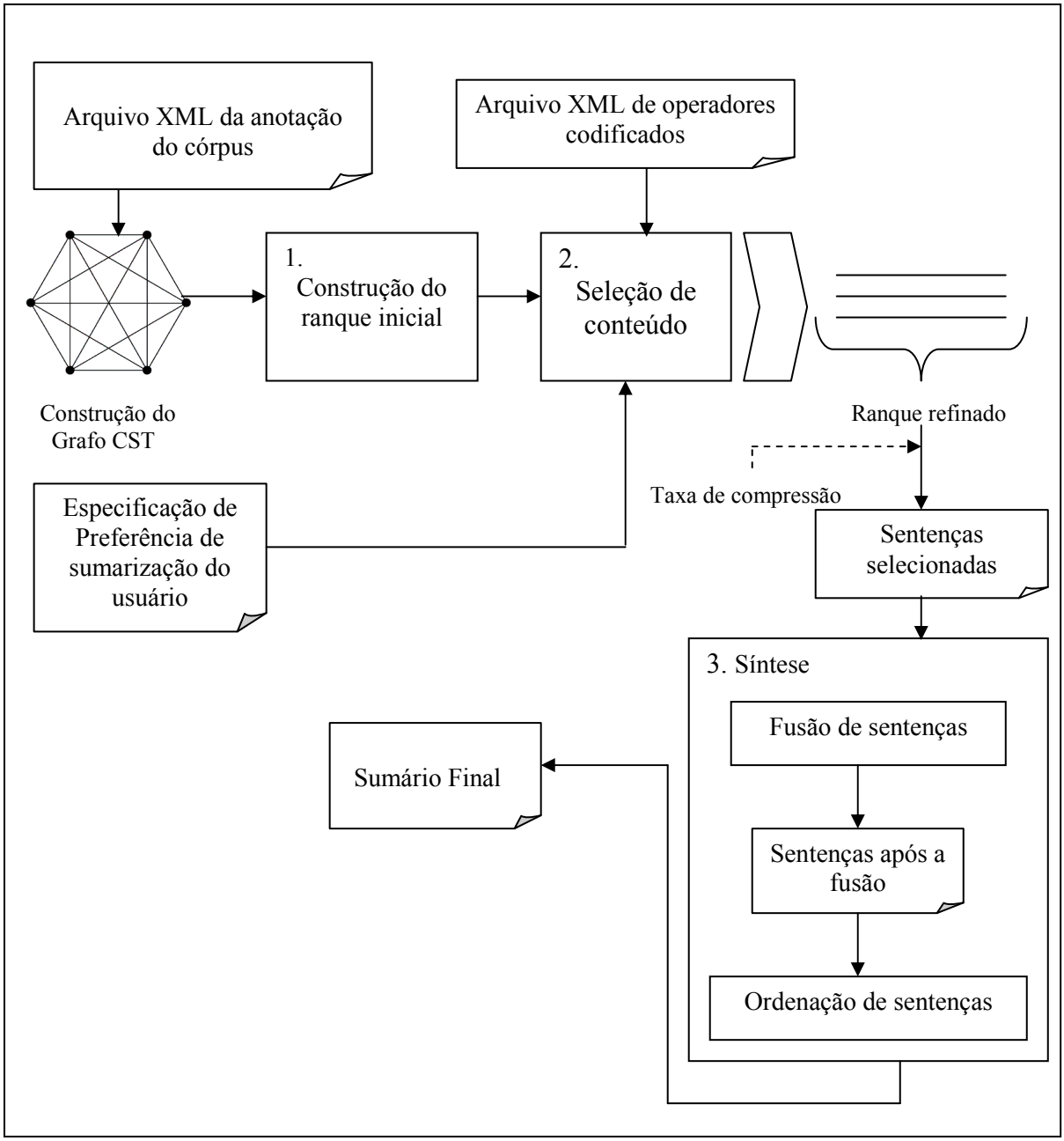

Figura 31: Protótipo de sumarização usando operadores de seleção de conteúdo

Procedimento para a aplicação de operadores de seleção de conteúdo
Entrada: Grafo CST
Saída: Ranque refinado
Construir o ranque inicial a partir do grafo CST (usando o operador geral/de informação principal)
Ler preferência de sumarização do usuário (se houver alguma)
Selecionar operador de seleção de conteúdo de acordo com a preferência de sumarização do usuário
Para cada regra do operador selecionado
    Para i=unidade informativa na primeira posição no ranque **até** a última posição do ranque
        Para j=unidade informativa na posição i+1 no ranque **até** a última posição no ranque
            Se as condições e restrições da regra são satisfeitas **então** aplicar as ações correspondentes nas sentenças i e j

Figura 32: Algoritmo geral para aplicação de operadores de seleção de conteúdo

Na seção seguinte é descrita a metodologia de avaliação usada neste trabalho e os resultados são discutidos.

4.5. Avaliação dos operadores de seleção de conteúdo

Existem duas formas de avaliar um sumário automático: de forma automática, onde um sistema determina a informatividade do sumário, e de forma humana, onde se avaliam aspectos da qualidade do sumário que não podem ser avaliados por uma ferramenta automática, por exemplo, gramaticalidade, coesão, coerência, etc. Para o presente trabalho, utilizaremos as duas formas de avaliação, pois assim podemos avaliar os aspectos da informatividade e os da qualidade.

Particularmente, neste trabalho, a informatividade dos sumários é avaliada em relação ao sumário de referência (elaborado por humanos) (Mani, 2001). Este sumário é composto pelas sentenças do texto fonte, que, de acordo com um critério humano, são consideradas essenciais para compor o sumário. Os sumários de referência utilizados na avaliação dos experimentos deste trabalho foram elaborados pelos próprios anotadores do córpus CSTNews, produzindo assim um conjunto de 50 sumários de referência, cada um deles correspondente a uma coleção de textos do córpus. É importante dizer que estes sumários foram elaborados de acordo com um critério geral de informatividade, ou

seja, eles simplesmente contêm as informações mais relevantes, segundo o anotador que elaborou o sumário, sem considerar nenhuma preferência de sumarização.

A informatividade dos sumários automáticos, gerados pelos diferentes métodos utilizados neste trabalho, é avaliada usando a medida ROUGE (Lin e Hovy, 2003), por ser a mais comumente utilizada para sumários automáticos. Esta é uma medida automática com julgamento próximo do humano. Ela computa a co-ocorrência de n-gramas entre o sumário automático e um ou mais sumários de referência feitos por humanos. Na medida ROUGE os n-gramas são considerados como seqüências de palavras que podem variar de 1 até 4 palavras. Neste trabalho, utilizamos somente unigramas, pois os autores da medida demonstraram que isso já é suficiente para discriminar bem a qualidade dos sumários.

Os resultados do ROUGE são dados em termos de precisão e cobertura em relação ao sumário de referência. A precisão e a cobertura definidas pelo ROUGE são:

$$P = \frac{\text{Número de n-gramas em comum com o sumário de referência}}{\text{Número de n-gramas do sumário automático}}$$

$$C = \frac{\text{Número de n-gramas em comum com o sumário de referência}}{\text{Número de n-gramas do sumário de referência}}$$

Existem outros fatores, além da informatividade, que não podem ser avaliados pela ROUGE, por exemplo, a coerência, a coesão e a redundância. Além disso, no caso dos sumários gerados automaticamente usando operadores de seleção de conteúdo, a avaliação da informatividade com os sumários de referência pode não ser o método mais adequado, já que os sumários de referência não consideram as mesmas preferências que foram consideradas nos sumários gerados a partir do uso de operadores de seleção de conteúdo. Para poder avaliar todos estes fatores, foi elaborado um cenário de avaliação humana onde os sumários são avaliados para cada preferência de sumarização utilizada. Esta avaliação foi feita por 6 juízes lingüistas computacionais. Cada avaliador recebeu um caderno contendo uma folha de instruções, uma tabela de avaliação, 8 sumários gerados automaticamente (em que 4 sumários correspondem a uma preferência de sumarização diferente) e os textos originais a partir dos quais foram

gerados estes sumários. Os juízes podiam atribuir uma nota de 0 e 4 para cada fator avaliado. Os valores são classificados em 5 categorias:

0. péssimo
1. ruim
2. regular
3. bom
4. excelente

No caso da informatividade, os avaliadores deviam considerar as preferências de sumarização do leitor. Por exemplo, se o leitor tem interesse em um sumário que aponte as contradições entre várias notícias sobre um mesmo assunto (por exemplo, relatos diferentes sobre número de mortos em alguma tragédia), quanto mais informações contraditórias o sumário tiver, mais informativo ele será; se o interesse for por informações sobre eventos referentes a um mesmo tópico que se desenvolvem em um período de tempo (por exemplo, os fatos que marcaram a crise econômica mundial), quanto mais dessas informações o sumário tiver, mais informativo ele será; se o interesse for por informações contextuais, quanto mais informações históricas, contextuais e detalhes o sumário tiver, mais informativo ele será;se o interesse for por informações de autoria e citações, quanto mais informações de fonte e autoria (ou seja, quem/que agência/que jornal/que organização noticiou ou afirmou alguma coisa) estiverem no sumário, mais informativo ele será.

Para a avaliação de cada sumário foram fornecidos os textos originais de onde foi elaborado o resumo. Os avaliadores deviam ler os textos originais para assim poder determinar se o resumo era realmente informativo ou não.

A seguir são mostrados os resultados das avaliações automáticas e humanas de acordo com a metodologia de avaliação descrita nesta seção.

4.5.1. Resultados da ROUGE

Os resultados obtidos a partir da avaliação com a ferramenta ROUGE são dados em termos de precisão, cobertura e medida-f, como é mostrado na Figura 33 a seguir:

	Precisão	Cobertura	Medida-F
Informação Geral	0.5564	0.5303	0.5356
Tratamento de Redundância	0.5761	0.5065	0.5297
Informação de Contexto	0.5196	0.4938	0.4994
Informação de autoria	0.5563	0.5224	0.5310
Informação contraditória	0.5503	0.5379	0.5365
Informações de Eventos Temporais	0.5159	0.5222	0.5140

Figura 33: Resultados da ROUGE para sumários gerados pelos operadores de seleção de conteúdo com base na CST

Pode-se observar que, em relação à avaliação da ROUGE, os melhores resultados correspondem aos operadores: geral, informação contraditória e informações de autoria. No caso do operador geral os resultados são bons já que os sumários de referência são genéricos também. No caso dos operadores de informação contraditória e de autoria os resultados são similares aos do operador geral, pois estes operadores tendem a gerar sumários mais genéricos, próximos ao sumário gerado pelo operador geral. Isto é porque as relações que neles se exploram não acontecem com muita freqüência no córpus, fazendo com que o ranque refinado fique muito próximo ao ranque inicial. Note, por exemplo, que as relações *Contradiction* e *Attribution* acontecem muito pouco no córpus (ver tabela de freqüência de relações na seção 4.2), portanto, a probabilidade de re-ranquear uma sentença com base nestas relações é baixa.

Contrariamente aos operadores de informação contraditória e de autoria, a medida-f é mais baixa para os operadores de redundância, contexto e eventos temporais, pois estes operadores têm mais probabilidade de mostrar informações que não acontecem em sumários genéricos. Note que as relações do tipo *Elaboration*, *Historical background* e *Subsumption* acontecem com muita freqüência no córpus, o que incrementa a probabilidade de re-ranquear as sentenças com base nestas relações. Isto resulta no fato do ranque inicial não ser genérico.

Para medir o grau de confiança destes resultados, foi aplicado o teste estatístico ANOVA (Zar,1999) de valor único para a precisão, a cobertura e a medida-f. Os resultados da ANOVA indicam que os resultados obtidos pela ROUGE são estatisticamente significativos para a precisão e a cobertura, com confiança de 95%. Para a medida-f, os resultados obtidos são estatisticamente significativos com 70% de confiança.

Os resultados obtidos confirmam a hipótese de que o uso do modelo CST na sumarização multidocumento ajuda a melhorar a informatividade dos sumários automáticos.

A seguir são mostrados os resultados da avaliação humana para os operadores de seleção de conteúdo.

4.5.2. Resultados da avaliação humana

Os resultados da avaliação humana para os operadores de seleção de conteúdo são mostrados em cinco tabelas, correspondentes aos cinco fatores avaliados: informatividade, coerência, coesão, redundância e gramaticalidade. Em cada tabela se mostra o desempenho dos operadores com base nas 5 possíveis notas consideradas pelos juízes na avaliação: (0) péssimo, (1) ruim, (2) regular, (3) bom e (4) excelente. O valor de cada célula da tabela indica o valor em porcentagem da atribuição daquela nota para cada operador de seleção de conteúdo. Nas Figuras 34-37 são mostrados os resultados

	Péssimo(%)	Ruim(%)	Regular(%)	Bom(%)	Excelente(%)	Nota média
Informações Genéricas	0	0	25	17	58	3.6 (Excelente)
Informações Contraditórias	0	0	16	16	68	3.7 (Excelente)
Informações Eventos Temporais	0	0	16	50	34	3.2 (Bom)
Informações Contextuais	0	16	0	50	34	2.2 (regular)
Informações de Autoria	0	0	16	50	34	3 (Bom)

Figura 34: Avaliação humana para o fator de informatividade

O fator de informatividade teve uma boa pontuação. Em mais do 50% dos casos avaliados para todos os operadores, os juízes acharam que a informatividade foi excelente. Na avaliação humana, diferentemente da avaliação automática, a preferência de sumarização pode ser avaliada. Isto quer dizer que, de acordo com o critério dos juízes, mais dos 50% dos sumários avaliados de cada operador foram considerados informativos em relação à preferência de sumarização escolhida. Este resultado confirma novamente a hipótese de que o modelo CST ajuda a melhorar a informatividade dos sumários automáticos por meio da exploração das informações entre os textos.

	Péssimo (%)	Ruim (%)	Regular (%)	Bom (%)	Excelente (%)	Nota média
Informações Genéricas	0	0	0	42	58	3.6 (Excelente)
Informações Contraditórias	0	16	33	33	16	2.4 (Regular)
Informações Eventos Temporais	0	16	16	34	34	2.1 (Regular)
Informações Contextuais	0	16	16	34	34	2.1 (Bom)
Informações de Autoria	0	0	0	66	34	3.3 (Bom)

Figura 35: Avaliação humana para o fator de coerência

	Péssimo (%)	Ruim (%)	Regular (%)	Bom (%)	Excelente (%)	Nota média
Informações Genéricas	0	0	25	17	58	3.2 (Bom)
Informações Contraditórias	0	0	17	50	33	2.7 (Bom)
Informações Eventos Temporais	0	0	50	50	0	2.5 (Regular)
Informações Contextuais	0	17	0	50	33	2.7 (Bom)
Informações de Autoria	0	0	50	17	33	(2.4) Regular

Figura 36: Avaliação humana para o fator de coesão

Nos casos de coerência e coesão, os resultados foram um pouco piores em relação à informatividade, mas nenhum deles foi considerado ruim. Entre os fatores que podem ter influenciado este resultado, podemos considerar as sentenças geradas pelo sistema de fusão, já que estas podem ser geradas com erros. Por outro lado, o critério de ordenação de sentenças é bastante básico e pode não mostrar sempre a ordenação mais coerente das sentenças. A pesar destes fatores, em termos gerais a coerência e a coesão observam um resultado bom em média. Neste trabalho assumimos que isto se deve a que o modelo CST ajuda a explorar semanticamente as informações o que permite melhorar a coerência e coesão dos sumários automáticos.

	Péssimo (%)	Ruim (%)	Regular (%)	Bom (%)	Excelente (%)	Nota média
Informações Genéricas	0	8	25	25	42	1.85 (Regular)
Informações Contraditórias	0	0	50	50	0	2.5 (Regular)
Informações Eventos Temporais	0	0	17	50	33	2.6 (Bom)
Informações Contextuais	0	0	17	30	53	3.6 (Bom)
Informações de Autoria	0	17	50	16	17	2.8 (Bom)

Figura 37: Avaliação humana para o fator de redundância

Finalmente os resultados para o fator de redundância foram um pouco piores em relação aos outros fatores avaliados. Igualmente ao caso da coerência e da coesão isto pode ser devido aos resultados do sistema de fusão de sentenças que pode não resolver todas as redundâncias presentes nas sentenças que são fundidas. De outro lado, existem outras relações CST além das tratadas no operador de redundância que podem trazer algumas informações repetidas. Em particular veja o exemplo a seguir de duas sentenças relacionadas pela relação *Contradiction:*

> **S1.** Cairo - 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.
>
> **S2.** CAIRO - 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.

Pode-se observar que nas duas sentenças há uma contradição no que se refere ao número de mortos no acidente, mas as duas sentenças tambem têm informações repetidas entre elas. Portanto, dependendo da natureza das informações algumas relações podem-se estabelecer entre sentenças de informação redundante. Isto faz com que a aplicação de alguns operadores possa trazer informações repetidas ao sumário final.

Os resultados obtidos confirmam as hipóteses nas quais se afirma que o modelo CST permite explorar as informações redundantes, contraditórias e complementares dos textos ajudando assim a melhorar a informatividade e qualidade dos sumários.

4.6. MEAD, GistSumm e CST

Como foi visto no Capitulo 3, os sumarizadores MEAD (Radev, 2001a) e GistSumm (Pardo et al., 2005) são sumarizadores superficiais, sendo que o MEAD é um sumarizador multilíngüe e o GistSumm é um sumarizador para inglês e português. Nos experimentos realizados, a metodologia proposta por Zhang et al. (2002), descrita no Capítulo 3, é aplicada para cada um destes sumarizadores.

Inicialmente o sumarizador MEAD constrói um ranque de sentenças pontuadas com base em três características principais: a posição da sentença no texto, a distância lexical entre a sentença e o centróide do texto (sentença mais representativa do texto), e finalmente, o tamanho da sentença. Estes três elementos são combinados linearmente e produzem uma pontuação para cada sentença. De acordo com essa pontuação, as sentenças são ranqueadas, sendo que as sentenças colocadas nas primeiras posições são consideradas as mais relevantes. Finalmente o sumarizador seleciona as sentenças melhor pontuadas que irão compor o sumário final.

A proposta de Zhang et al. (2002) adiciona o conhecimento da CST ao processo de ranqueamento do sumarizador MEAD, adicionando ao valor da pontuação original o numero de relações CST da respectiva sentença.

pontuação anterior = pontuação anterior + número de relações CST

Assim, depois de ter re-calculado a pontuação de cada sentença, estas são re-ranqueadas. Para adaptar esta proposta ao cenário do português brasileiro, neste trabalho foi usada a versão 3.12 do sumarizador MEAD, disponível online[2]. Nesta versão do MEAD são incluídos alguns pacotes adicionais contendo 12 re-ranqueadores com base na CST. Cada re-ranqueador utiliza um critério diferente para pontuar as sentenças no ranque. No presente trabalho, utilizamos o re-ranqueador mais simples, que utiliza como critério de re-ranqueamento o número de relações CST de uma determinada sentença, sem considerar nenhuma preferência de sumarização. O desempenho do MEAD é avaliado antes de aplicar o re-ranqueamento e depois de aplicá-lo.

O sistema trabalha sobre um grupo que contém todas as informações multidocumento necessárias. O grupo contém textos sobre um mesmo assunto. Além desses textos, o grupo também contém dois arquivos principais (“nome.clust” e

[2] http://www.summarization.com/mead

"nome.sentjudge"). O arquivo do tipo *.clust*, contém os nomes ou referências aos textos de entrada originais. O arquivo do tipo *.sentjudge* contém as anotações CST dos textos, e, essas anotações estão indicadas em formato xml, sendo que este formato é o mesmo que utilizamos para anotar as relações CST no córpus CSTNews. As Figuras 39 e 40 mostram exemplos de arquivos do tipo *.clust* e arquivos do tipo *.sentjudge* respectivamente.

```
<CLUSTER LANG="ENG">
     <D DID="D2_C8.txt" />
     <D DID="D1_C8.txt" />
     <D DID="D3_C8.txt" />
</CLUSTER>
```

Figura 38: Exemplo de arquivo tipo .clust

```
<TABLE>
<R SDID="D3_C8.txt" SSENT="10" TDID="D2_C8.txt" TSENT="1">
<RELATION TYPE="Elaboration" JUDGE="thiago"/>
</R>
<R SDID="D2_C8.txt" SSENT="2" TDID="D3_C8.txt" TSENT="3">
<RELATION TYPE="Overlap" JUDGE="thiago"/>
</R>
<R SDID="D2_C8.txt" SSENT="3" TDID="D3_C8.txt" TSENT="2">
<RELATION TYPE="Overlap" JUDGE="thiago"/>
</R>
<R SDID="D2_C8.txt" SSENT="4" TDID="D3_C8.txt" TSENT="1">
<RELATION TYPE="Follow-up" JUDGE="thiago"/>
</R>
</TABLE>
```

Figura 39: Exemplo de arquivo tipo .sentjudge

O código XML mostrado na Figura 39 indica os nomes dos arquivos dos três textos que compõem o cluster. Em particular, cada nome de cada texto indica o número do documento e o número do cluster ao qual o texto pertence, assim, o número do documento está escrito depois da letra "D" e número do Cluster depois da letra "C". O MEAD lê primeiro este arquivo para saber quais serão os textos a sumarizar.

Os textos de entrada devem ser convertidos ao formato de arquivo tipo.*clust* do MEAD. Isto é feito usando o comando "./text2cluster.pl DIRECTORY", onde "DIRECTORY" é a pasta contendo os textos a serem sumarizados. Cada pasta representa um cluster contendo textos sobre um mesmo assunto.

Na Figura 40**,** o código mostra as anotações das relações CST entre pares de sentenças. O código XML usado é o mesmo código usado para a anotação do córpus CSTNews.

Na Figura 41, mostra-se um exemplo de um sumário gerado a partir do MEAD. Na Figura 42, mostra-se um sumário automático usando MEAD e CST.

SÃO PAULO - 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. Com isso, a reforma emergencial, que começaria em breve, foi descartada.
Com isso, as outras duas partes ficam disponíveis para pousos e decolagens. Enquanto durarem as obras, os vôos serão transferidos para o Aeroporto de Viracopos, em Campinas, a 95 km da Capital, segundo informações dadas pelo ministro Nelson Jobim pro meio de nota. O Ministério da Defesa não soube informar por quanto tempo a pista permanecer fechada.

Figura 40: Exemplo de sumário MEAD sem usar CST

SÃO PAULO - 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. Com isso, a reforma emergencial, que começaria em breve, foi descartada.
O ministro da Defesa, Nelson Jobim, anunciou a reforma que, segundo estudos da Empresa Brasileira de Infra-Estrutura Aeroportuária Infraero , poderá ser feita sem que a pista seja interditada. Apesar da definição, o cronograma da obra não foi divulgado.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.

Figura 41: Exemplo de sumário MEAD usando CST

Pode-se notar que, na Figura 42, as sentenças 2 e 3 não são as mesmas da Figura 41, sendo que as sentenças da Figura 42 têm seu foco nas informações referentes ao assunto principal do tópico, fornecendo mais detalhes.

Seguindo o critério do experimento anterior, aplicamos uma função de re-ranqueamento ao ranque gerado pelo sumarizador GistSumm. Inicialmente o GistSumm da uma pontuação a cada sentença utilizando métodos estatísticos como *Keywords* (Luhn, 1958) ou a medida TF-ISF *(Term frequency Inverse Sentence frequency)* (Laroca et al., 2000). A primeira sentença considera-se a mais relevante. A pontuação do resto das sentenças depende da similaridade lexical entre a sentença sendo analisada e a sentença principal. Depois, as sentenças que tiverem pontuação maior que a média das pontuações das outras sentenças, serão colocadas nas primeiras posições do ranque. Assim, é gerado para cada grupo o ranque correspondente e este é salvo num arquivo de texto. Cada arquivo contém dados sobre o número do documento, o número da sentença, a própria sentença e a pontuação dela. Estas informações são carregadas numa

estrutura de dados e logo as pontuações são re-calculadas usando as informações contidas no grafo CST do grupo correspondente. Depois de re-calcular as pontuações, o ranque é reorganizado, deixando no topo as sentenças melhor pontuadas, sendo que as primeiras são incluídas no sumário final. O número de sentenças incluídas depende da taxa de compressão escolhida.

Nas Figuras 43 e 44, são mostrados dois sumários usando GistSumm sem CST e com CST, respectivamente.

A aviação de Israel realizou durante a madrugada desta segunda-feira, dia 7, ataques a 150 alvos no Líbano.
Enquanto isso, soldados israelenses mataram 10 integrantes da milícia do Hezbollah.
Durante este domingo, dia 6, foram travadas lutas sangrentas.
A aviação israelense atacou 150 alvos na madrugada de hoje no Líbano, enquanto soldados do Estado judeu mataram 10 milicianos do Hezbollah nas aldeias libanesas de Bint Djebeil e Kafr Hula, segundo informações de fontes militares.
Os combates se intensificaram hoje após a sangrenta batalha deste domingo, quando a guerrilha xiita do Hezbollah matou 15 pessoas e deixou mais de 200 feridas, entre militares e civis.

Figura 42: Exemplo de sumário gerado pelo GistSumm sem usar CST

A aviação de Israel realizou durante a madrugada desta segunda-feira, dia 7, ataques a 150 alvos no Líbano.
Enquanto isso, soldados israelenses mataram 10 integrantes da milícia do Hezbollah.
Durante este domingo, dia 6, foram travadas lutas sangrentas.
Os foguetes e ataques do Hezbollah causaram a morte de 15 pessoas e deixaram mais de 200 feridas.
Já o Exército de Israel provocaram a morte de 30 militantes do Hezbollah.
Comandos israelenses mataram outros três guerrilheiros libaneses na cidade de Tiro, onde destruíram sete plataformas de lançamento de foguetes, informaram as fontes israelenses.
A aviação israelense atacou 150 alvos na madrugada de hoje no Líbano, enquanto soldados do Estado judeu mataram 10 milicianos do Hisbol nas aldeias libanesas de Bint Djebeil e Kafr Hula, segundo informações de fontes militares.

Figura 43: Exemplo de sumário gerado pelo GistSumm usando CST

Observando-se as Figuras 43 e 44, pode-se notar que as sentenças 4 e 5 da Figura 43 foram re-ranqueadas e novas sentenças foram introduzidas no sumário da Figura 44, sendo que no sumário correspondente desta figura tem-se maior ênfase nas informações que fornecem detalhes contextuais sobre o tópico, como por exemplo números de mortos e pessoas feridas. Também é possível notar que em ambos os sumários não houve tratamento de redundância.

4.6.1. Resultados da ROUGE para os experimentos com o MEAD e GistSumm

Os resultados obtidos a partir da avaliação do sumarizador MEAD com a medida ROUGE são mostrados na Figura 44 a seguir.

	Precisão	Cobertura	F-measure
MEAD sem CST	0.5242	0.4602	0.4869
MEAD com CST	0.5599	0.4988	0.5230

Figura 44: Resultados da ROUGE para o sumarizador MEAD simples e MEAD usando CST

É possível observar a partir dos resultados que o uso da CST também ajuda a melhorar o desempenho do sumarizador MEAD, sem considerar nenhuma preferência em particular, apenas o critério geral do número total de relações CST de cada sentença do ranque. Também é importante notar que no caso dos operadores de preferência e do MEAD (com o uso da CST e sem o uso da CST) a precisão sempre é mais alta do que a cobertura. Isto pode ser conseqüência da taxa de compressão utilizada, já que nos primeiros lugares do ranque estão as principais sentenças que deveriam ser incluídas no sumário final, mas, dependendo da taxa de compressão, podem-se incluir poucas ou muitas destas sentenças importantes. Quanto mais alta é a taxa de compressão, menos sentenças serão incluídas e, portanto há uma menor probabilidade de que sejam cobertas a maioria das sentenças do sumário de referência. No nosso caso, a taxa de compressão é de 70% sobre o texto de maior tamanho. Finalmente, na Figura 45 mostra-se os resultados para o sumarizador GistSumm usando as duas estratégia descritas anteriormente: com o uso da CST e sem o uso da CST.

	Precisão	Cobertura	F-measure
GistSumm sem CST	0.3599	0.6643	0.4599
GistSumm com CST	0.4945	0.5089	0.4994

Figura 45: Resultados da ROUGE para o sumarizador GistSumm simples e GistSumm usando CST

Em termos gerais pode-se dizer que os métodos com base na CST têm melhor desempenho que os métodos estatísticos. No caso do GistSumm, apesar dele apresentar pior desempenho em relação aos outros métodos explorados neste trabalho, observa-se que com o uso da CST melhorou-se a precisão e a medida-f. Isto indica que o uso da CST em sumarizadores superficiais como o GistSumm e o MEAD ajuda a melhorar a informatividade dos sumários produzidos.

Finalmente, foi aplicado o Teste-T (Zar,1999) para determinar se os resultados são estatisticamente significativos. Os resultados do teste estatístico mostram que os resultados da ROUGE para a medida-f, cobertura e precisão são estatisticamente significativos com 95% de confiança.

Os resultados obtidos para os sumarizadores GistSumm e MEAD também confirmam a hipótese de que o uso do modelo CST ajuda a melhorar a informatividade dos sumários produzidos.

O próximo capítulo apresenta as considerações finais deste trabalho.

Capítulo

5

5. Considerações finais

A partir dos experimentos realizados, neste Capítulo são discutidas as contribuições e limitações do trabalho. Também são propostos alguns trabalhos futuros que seguem na linha desta dissertação.

5.1. Contribuições

Neste trabalho foi realizada a primeira investigação de abordagem profunda para sumarização multidocumento para o português do Brasil. A partir disto, foram dadas varias contribuições. Primeiramente, foi realizado um novo refinamento do modelo CST a partir do trabalho de Aleixo e Pardo (2008a). Este refinamento propõe uma nova tipologia das relações. Além desta tipologia, as definições das relações foram melhor formalizadas. Os resultados obtidos, de acordo com a medida Kappa, mostram que a concordância na anotação do córpus aumentou com o uso das relações refinadas. Esta tipologia proposta pode ser avaliada em qualquer língua em que o modelo CST seja aplicado.

A partir da investigação realizada foi possível mostrar que os textos em português brasileiro podem ser representados com o modelo CST, o que faz possível explorar as informações dos textos, detectando similaridades, diferenças e informações complementares. O conhecimento destas informações permite estudar e tratar melhor os desafios da sumarização multidocumento.

Com foco na seleção de conteúdo, foram definidos, formalizados e avaliados um conjunto de operadores para sumarização multidocumento com base no modelo CST. Estes operadores podem ser aplicados em textos escritos em qualquer língua, desde que sejam modelados de acordo com o modelo CST. As regras definidas permitem explorar as informações fornecidas pelo modelo CST e assim tratar os principais desafios da

sumarização multidocumento: informações redundantes, contraditórias e complementares.

A partir da definição e formalização teórica dos operadores foi desenvolvido um protótipo de sumarizador. Este protótipo é o primeiro para sumarização multidocumento de abordagem profunda para o português do Brasil, mas também pode ser aplicado em textos escritos em outras línguas.

Foram explorados dois métodos de sumarização de abordagem superficial: MEAD e GistSumm. Nos dois métodos foi incorporado o conhecimento fornecido pela CST (considerando a anotação a partir do novo refinamento proposto para este trabalho) ao processo correspondente de seleção de conteúdo. Mostra-se que o uso do conhecimento do modelo CST ajuda a melhorar o desempenho destes dois sumarizadores.

Finalmente, foram fornecidos os resultados da avaliação de cada um dos métodos investigados nesta dissertação, sendo que estes resultados foram dados em termos de informatividade e qualidade.

Em geral pode-se dizer que os resultados dos métodos que usam o conhecimento do modelo CST mostram um melhor desempenho que aqueles que não usam. Isso valida a hipótese principal deste trabalho: o uso do conhecimento profundo que o modelo CST fornece permite melhorar a informatividade e qualidade dos sumários automáticos.

5.2. Limitações do trabalho

Apesar dos bons resultados, o trabalho realizado ainda possui algumas limitações. Por exemplo, ainda não contamos com um *parser* automático para anotação CST. Isto faz com que o processo de sumarização seja ainda dependente do humano, o que resulta em um processo custoso, trabalhoso e subjetivo. Para resolver esta limitação, atualmente há uma pesquisa em desenvolvimento no NILC que visa construir um *parser* automático para anotação CST.

Uma outra limitação do trabalho é que o tratamento de redundância nas sentenças relacionadas pela relação de *Overlap* não faz parte do processo de seleção de conteúdo. No entanto, este tratamento é feito na etapa de síntese. O sistema de fusão de sentenças de Seno e Nunes (2009) ainda está em desenvolvimento, fazendo com que a integração do sistema de fusão no protótipo de sumarização ainda não seja possível.

Outro ponto a considerar é que, neste trabalho, por enquanto, só é permitida a aplicação de um operador de seleção de conteúdo por vez. Ao permitir a aplicação de mais de um operador num mesmo processo de sumarização, o ranqueamento feito pelo

operador anterior pode ser alterado pelo novo operador e, portanto a preferência do operador anterior se perderia. De fato, a preferência do último operador aplicado é a que vai prevalecer. Em particular, o único operador que é combinado com o resto dos operadores é o operador de redundância, pois este operador é o único que elimina informações redundantes, sem perder informações de outros operadores. Assim, o operador de redundância é aplicado após ter aplicado qualquer um dos outros operadores propostos neste trabalho, segundo a preferência de sumarização escolhida.

Pode-se notar também que neste trabalho não foi tratada a ordenação de sentenças no sumário final, já que o foco do trabalho foi a etapa de transformação, em particular na tarefa de Seleção de Conteúdo. A pesar de não ter sido estudada em profundidade a tarefa de ordenação de sentenças, foi considerado um critério simples de ordenação baseado na posição da sentença no documento original. Particularmente isto pode causar a colisão de mais de uma sentença já que há vários textos e as sentenças podem ter a mesma posição, mas em textos diferentes, neste caso, utiliza-se um segundo critério de ordenação que é o número do documento ao qual pertence aquela sentença. Este método pode não ser o mais efetivo o que é propósito de pesquisas futuras.

5.3. Trabalhos futuros

Visando explorar mais profundamente a sumarização multidocumento e considerando os resultados obtidos neste trabalho, sugerem-se os seguintes trabalhos futuros:

- Explorar estratégias que permitam integrar mais de uma preferência de sumarização num mesmo processo de seleção de conteúdo. Isto pode significar a aplicação de mais de um operador de seleção de conteúdo ou a definição e formalização de um operador que integre mais de uma preferência de sumarização. Esta tarefa pode ser explorada a partir do modelo CST ou também incluir ou investigar novas estratégias que permitam atingir o objetivo.
- Explorar novos métodos de Seleção de Conteúdo para sumarização multidocumento, que permitam combinar estratégias tanto da abordagem profunda quanto da superficial.
- Visa-se também contruir novas estratégias para sumarização multidocumento aprendidas a partir de córpus em português. Isto inclui estratégias estatísticas de aprendizado, assim como o uso de conhecimento

lingüístico sobre a estruturação do texto para extrair padrões de sumarização no córpus.

- Explorar estratégias com base na CST para melhorar a coerência e coesão de sumários multidocumento. Isto pode incluir, por exemplo, explorar o modelo CST para o problema da ordenação de sentenças ou a resolução de cadeias de correferência. Também pode-se considerar abordagens mistas para tratar estes problemas, o que pode resultar na criação de novas estratégias para abordar estes problemas.

Finalmente, outros trabalhos futuros a considerar podem incluir a investigação e exploração de outras tarefas correspondentes às diferentes etapas da sumarização multidocumento, visando assim melhorar às estratégias na etapa de análise ou na etapa de síntese.

Referências Bibliográficas

Afantenos, S.D.; Doura, I.; Kapellou, E.; Karkaletsis, V. (2004). Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In the Proceedings of SETN, pp. 410-419.

Afantenos, S.D. (2007). Reflections on the Task of Content Determination in the Context of Multi-Document Summarization of Evolving Events. *In Recent Advances on Natural Language Processing 2007*. Borovets-Bulgaria.

Aleixo, P. e Pardo, T.A.S. (2008a). *CSTNews:* Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva CST (Cross-Document Structure Theory). Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP.

Aleixo, P. e Pardo, T.A.S. (2008b). *CSTTool:* um parser multidocumento automático para o Português do Brasil. *In the* P*roceedings of the IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence* – WTDIA. Salvador-Bahia.

Aleixo, P. and Pardo, T.A.S. (2008c). Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL, pp. 298-303. Vila Velha, Espírito Santo. October, 26-28

Allan, J. (1996). Automatic Hipertext Linking Type. *In Proceedings of Hypertext 1996. Washington D.C.*

Barzilay, R.; Elhadad, N.; McKeown. K. (2001). Sentence Ordering in Multidocument Summarization. In the *Proceedings of the 1st Human Language Technology Conference*. San Diego, California.

Barzilay, R.; McKeown, K.; Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In the *Proceedings of the ACL*. Maryland, USA.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In the *Proceedings of ACM-SIGIR*. Melbourne, Australia.

Farzindar, A.; Rozon, F.; Lapalme, G. (2005). CATS A Topic-Oriented Multi-Document Summarization System at DUC 2005.

Fleiss, JL. (1971). Measuring nominal scale agreement among many raters. Psychol Bull Vol 76, nº 5, 378-382

Goldstein, J.; Mittal, V.O.; Carbonell, J. G.; Kantrowitz, M. (2000). Multi-Document Summarization by Sentence Extraction. *In ACL Workshop On Automatic Summarization.*

Larocca Neto, J.; Santos, A.D.; Kaestner, A.A.; Freitas, A.A. (2000). Generating Text Summaries through the Relative Importance of Topics. *In M.C. Monard and J.S. Sichman (eds.), Lecture Notes in Artificial Intelligence, No. 1952, pp.300-309.* Srpinger- Verlag.

Leuski, A.; Lin, C-Y.; Hovy, E. (2003). iNeATS: Interactive Multi-Document Summarization. In the *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan.

Lin, C-Y. and Hovy, E. (2002). From single to multi-document summarization: a prototype system and its evaluation. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, USA.

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

Luhn, H.P. (1958). The automatic creation of literature abstracts. *In IBM Journal of Research and Development. Vol. 2, pp. 159-165.*

Mani, I. and Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In the *Proceedings of the 14th National Conference on Artificial Intelligence* (AAAI), pp. 622-628. American Association for Artificial Intelligence.

Mani, I. and Maybury, M. T. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA.

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

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

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

McKeown, K. and Radev, D.R. (1995). Generating summaries of multiple news articles. In the *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 74-82, Seattle, WA.

McKeown, K.; Barzilay, R.; Evans, D.; Hatzivassiloglou, V.; Klavans, J.L.; Nenkova, A.; Sable, C.; Schiffman, B.; Sigelman, S. (2002). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In the *Proceedings of the Human Language Technology Conference*.

McKeown, K.; Passonneau, R.; Elson, D.; Nenkova, A.; Hirschberg J. (2005). Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization. In the *Proceedings of the 28th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brazil.

Mihalcea, R. and Tarau, P. (2005). An Algorithm for Language Independent Single and Multiple Document Summarization. In the *Proceedings of the International Joint Conference on Natural Language Processing* (IJCNLP). Korea.

Nenkova, A. and Louis, A. (2008). Can you Summarize this? Identifying correlates of Input difficulty for generic multi-document summarization. In the *Proceedings of the ACL.*

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

Otterbacher, J.C.; Radev, D.R.; Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In the *Proceedings of the Workshop on Automatic Summarization*, pp 27-36. Philadelphia.

Papineni, K.A.; Roukos, S.; Ward, T.; Zhu, W.J. (2001). Bleu: a method for automatic evaluation of machine translation. *Tech. Rep. RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center*. Yorktown Heights, NY.

Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

Pardo, T.A.S. (2005). *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP/Brasil.

Pardo, T.A.S (2006). SENTER: Um Segmentador Sentencial Automático para o Português do Brasil. *Série de Relatórios do NILC. NILC-TR-06-01.* São Carlos-SP/Brasil.

Pardo, T.A.S.; Filho, P.P.B.; Uzêda, V.R.; Nunes, M.G.V. (2007). Experiments on Applying a Text Summarization System for Question Answering. *In Evaluation of*

Multilingual and Multi-Modal Information Retrieval. Lecture Notes in Computer Science. Springer Berlin/Heidelberg.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong.

Radev, D.R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, N. 3, pp. 469-500.

Radev, D.R.; Jing, H.; Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In the *Proceedings of the ANLP/NAACL Workshop*, pp. 21-29.

Radev, D.R.; Blair-Goldensohn, S.; Zhang, Z. (2001a). Experiments in single and multi-document summarization using MEAD. In the *Proceedings of the First Document Understanding Conference*. New Orleans, LA.

Radev, D.R.; Blair-Goldensohn, S.; Zhang, Z.; Raghavan, R.S. (2001b). Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In the *Proceedings of Human Language Technology Conference*. San Diego, CA.

Radev, D.R.; Weigu, R.; Zhang, Z. (2001d). WebInEssence: A Personalized Web-Based multi-document Summarization and Recomendation System.

Radev, D.R.; Otterbacher, J.; Zhang, Z. (2003). CSTBank: Cross-document Structure Theory Bank. University of Michigan, Department of Electrical Engineering and Computer Science.

Radev, D., Otterbacher, J., Zhang, Z. (2008). Cross-document Relationship Classification for Text Summarization.

Radev, D.R. and Qazvinian, V. (2008). Scientific Paper Summarization Using Citation Summary Networks.

Reiter, E. and Dale, R. (2000). Building Natural Language Generation Systems. Cambridge University Press.

Salton, G. (1989). The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley. *Automatic Text Processing*.

Seno, E.R.M. and Nunes, M.G.V. (2009). Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português. *Linguamática*, Vol. 1, pp. 71-87.

Siegel, S.; Castellan, N.J. (1988). Nonparametric statistics for the Behavioral Sciences. McGraw Hill

Trigg, R. (1983). A Network-Based Approach to Text Handling for the Online Scientific Community. *PhD. Thesis. University of Maryland Technical Report, TR-1346.* College Park MD

Trigg, R.; Weiser, M. (1986). TEXTNET: A Network-Based Approach to Text Handling. *In ACM Transactions on Office Information Systems, Volume 6.*

Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2009). A comprehensive summary informativeness evaluation for RST-based summarization methods.*International Journal of Computer Information Systems and Industrial Management Applications* - IJCISIM, Vol. 1, pp. 188-196.

Wan, X. and Yang, J. (2006). Improved affinity graph based multi-document summarization. In Proceedings of HLT-NAACL2006.

Zar, J.H. (1941). Bioestatistical Analysis . Prentice Hall.

Zhang, Z.; Goldenshon, S.B.; Radev, D.R. (2002). Towards CST-Enhanced Sumarization. In Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002). Edmonton.

APÊNDICE A - Relações CST refinadas para a anotação do córpus CSTNews

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←S2
Restrições: S2 apresenta o mesmo conteúdo que S1, mas de forma mais compacta.
Comentários: Summary é um tipo de equivalence, mas summary deve haver diferença significativa de tamanho entre as sentenças.

Nome da Relação: *Subsumption*
Direcionalidade: S1→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←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←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←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 senteças.
Comentários:

Nome da Relação: *Citation*
Direcionalidade: S1←S2
Restrições: S2 cita explicitamente informação proveniente de S1.
Comentários: Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.

Nome da Relação: *Attribution*
Direcionalidade: S1←S2
Restrições: S1 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 B - Exemplos de relações CST entre dois textos

Texto 1: O Globo

S<1.1>: TOQUIO- Um terremoto de 6.8 graus na escala Richter, com epicentro a 17 quilômetros de profundidade, atingiu a costa noroeste do Japão às 10h13m desta segunda-feira (22h13m de domingo em Brasília).
S<2.1>: O terremoto, 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).
S<3.1>: Duas mulheres de cerca de 80 anos morreram quando suas casas desmoronaram após o tremor.
S<4.1>: Os detalhes das outras duas mortes, informadas pela emissora pública NHK, não estão disponíveis.
S<5.1>: Um pequeno incêndio aconteceu em um transformador elétrico da usina nuclear de Kashiwazaki Kariwa, a maior do mundo, localizada perto do epicentro, mas o fogo já foi controlado.
S<6.1>: Os reatores nucleares foram desligados e não houve liberação de radiação.

Texto 2: Gazeta do Povo

S<1.2>: Um forte terremoto matou ao menos cinco pessoas no noroeste do Japão nesta segunda –feira.
S<2.2>: Os prédios chegaram a tremer em Tóquio, e os reatores de usinas nucleares em Niigata desligaram-se automaticamente para checagens, embora não haja relatos de vazamento de radiação.
S<3.2>: Duas mulheres na faixa dos 80 anos morreram quando suas casas ruíram durante o tremor de magnitude 6.8, na área de Niigata, cerca de 250 km noroeste de Tóquio, informou a imprensa japonesa.
S<4.2>: "Prateleiras altas caíram e as coisas voaram por toda parte", contou Harumi Mikami, 55, uma professora que estava em sua escola na Cidade de Kashiwazaki, perto do epicentro do terremoto.
S<5.2>: Cerca de 1700 pessoas fugiram de suas casas para quase 100 centros de resgate, segundo a rede NHK e a Prefeitura de Niigata.
S<6.2>: Um incêndio em um transformador elétrico na usina nuclear de Kashiwazaki Kariwa foi rapidamente extinto, mas ainda não está claro quando a companhia elétrica de Tóquio vai religar três unidades no complexo, disse Yoshinobu Kamijima, porta-voz da empresa.
S<7.2>: O Japão é um dos países do mundo mais suscetíveis a terremotos, com um tremor ocorrendo ao menos cada cinco minutos.

Relações entre os textos

Par de sentenças	Relações	Direcionalidade
S<1.1>, S<2.1>	*Overlap*	Sem direcionalidade
S<1.1>, S<2.3>	*Overlap*	Sem direcionalidade
S<1.1>, S<2.4>	*Elaboration*	S<2.4> → S<1.1>
S<1.1>, S<2.7>	*Historical Background*	S<2.7> → S<1.1>
S<1.2>, S<2.7>	*Historical Background*	S<2.7> → S<1.2>
S<1.3>, S<2.3>	*Subsumption*	S<2.3> → S<1.3>
S<1.5>, S<2.2>	*Overlap*	Sem direcionalidade
S<1.5>, S<2.6>	*Overlap*	Sem direcionalidade
S<1.6>, S<2.2>	*Subsumption*	S<2.2> → S<1.6>