Métodos para Sumarização Automática Multidocumento Usando Modelos Semântico-Discursivos

Paula C. F. Cardoso, Thiago A. S. Pardo, Maria das Graças V. Nunes

Núcleo Interinstitucional de Linguística Computacional (NILC)

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP)

Caixa Postal: 668 - CEP: 13560-970 - São Carlos - SP

{paulastm, taspardo, gracan}@icmc.usp.br

Abstract. *Multi-document summarization aims at producing a summary from a group of texts on the same topic. In this paper, we propose summarization methods that combine information from each source-text and from the relationships among passages of different texts in order to produce more informative summaries, using the Rhetorical Structure Theory (RST) and Cross-document Structure Theory (CST) discourse models.*

Resumo. *A sumarização automática multidocumento visa produzir um sumário a partir de um conjunto de textos que versam sobre um mesmo assunto. Neste artigo, propõem-se métodos de sumarização que utilizam as informações provenientes de cada texto-fonte em conjunto com as informações sobre o relacionamento entre as partes de vários textos para produzir sumários mais informativos. Para isso, são explorados os modelos semântico-discursivos Rhetorical Structure Theory (RST) e Cross-document Structure Theory (CST).*

1. Introdução

A sumarização automática (SA) multidocumento, que visa à produção de um sumário a partir de um conjunto de textos relacionados, tem recebido muita atenção nos últimos tempos, devido a enorme quantidade de informação disponível e a necessidade das pessoas obterem informações em um curto espaço de tempo (Mani, 2001). O cenário consiste de grupos de textos que descrevem eventos relacionados, são publicados em tempos diferentes e apresentam variados estilos de escrita. Em alguns casos, uma mesma informação pode ser encontrada de diferentes maneiras, sendo parafraseada ou até mesmo apresentando incompatibilidade entre as fontes. Em outras situações, para se ter uma breve visão do evento, o leitor precisa ler textos de várias fontes, devido nenhuma delas apresentarem a informação relevante. Entre os desafios da SA multidocumento estão: reconhecer quais são as informações redundantes, complementares e contraditórias no conjunto de textos relacionados, e manter a coerência do sumário.

Para a SA multidocumento, as informações que são mais repetidas e elaboradas entre as fontes são ditas as mais importantes. Nos últimos anos, vários trabalhos de SA multidocumento propuseram diferentes estratégias para encontrar as informações relevantes que deveriam formar o sumário (Mani e Bloedorn, 1997; Pardo, 2005; Radev

Anais do III Workshop "A RST e os Estudos do Texto", páginas 59–74, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

at al., 2004; Afantenos et al., 2007; Jorge e Pardo, 2010; Jorge et al. 2011; etc). O núcleo dessas propostas é a identificação das similaridades e diferenças entre textos relacionados, mas sem analisar a estrutura textual de cada texto-fonte. Acredita-se que, ao considerar a estrutura textual de cada texto-fonte na SA multidocumento, outras informações poderão surgir para enriquecer o sumário resultante, por exemplo, o foco de cada texto.

Neste artigo, propõem-se métodos de seleção de conteúdo que utilizem as informações provenientes de cada texto-fonte em conjunto com as informações sobre o relacionamento entre as partes de vários textos para produzir sumários multidocumento mais informativos. Para isso, são exploradas as teorias semântico-discursivas *Rhetorical Structure Theory* (RST - Mann e Thompson, 1987) e *Cross-document Structure Theory* (CST - Radev, 2000). A RST é uma teoria linguística descritiva que relaciona os segmentos discursivos de um texto através de relações retóricas. Cada segmento discursivo é classificado como núcleo, que é a informação principal, ou satélite que, por sua vez, é a informação adicional. Este princípio de nuclearidade indica as informações importantes em um texto. A CST é uma teoria semântico-discursiva composta de um conjunto de relações que detectam as similaridades, diferenças, contradições, informações complementares e diversidade de estilos de escrita entre textos relacionados. O conhecimento destas informações permite estudar e tratar melhor os desafios da sumarização multidocumento. As teorias RST e CST foram escolhidas por poderem ser aplicadas a uma variedade de textos e já terem sido utilizadas em trabalhos de modelagem e SA de textos na língua portuguesa (Pardo, 2002; Seno, 2005; Jorge e Pardo, 2010; Jorge et al. 2011).

A proposta faz parte de um projeto maior intitulado "SUCINTO", desenvolvido no Núcleo Interinstitucional de Linguística Computacional (NILC[1]). O NILC tem experiência em diversas áreas do Processamento de Línguas Naturais, sendo uma delas a SA. O artigo encontra-se organizado da seguinte forma: na Seção 2, apresentam-se as fases de um sistema de SA; na Seção 3, descreve-se a teoria RST; na Seção 4, descreve-se a teoria CST; na Seção 5 descreve a proposta de SA multidocumento. Por fim, na Seção 6 apresentam-se as considerações finais, incluindo perspectivas futuras.

2. Fases de um Sistema de Sumarização Automática

Em geral, os sistemas de SA têm uma arquitetura genérica, que está dividida em Análise, Transformação e Síntese (Sparck Jones, 1998), ilustrada na Figura 1. A entrada para o processo de sumarização consiste de um ou vários textos-fonte. A Análise visa interpretar os textos-fonte e extrair uma representação formal que possa ser processada automaticamente. Durante a análise podem ser utilizados analisadores morfológicos, sintáticos, semânticos e/ou discursivos. A Transformação é a principal etapa, que visa gerar uma representação interna do sumário a partir da representação fornecida na etapa anterior. Nesta etapa podem ser utilizados métodos de seleção de conteúdo, agregação, e substituição para compactar o conteúdo dos textos-fonte, produzindo uma mensagem que corresponderá ao sumário, mas ainda não necessariamente textual. A Síntese tem o propósito de gerar em linguagem natural a representação interna condensada em um

[1] http://www.nilc.icmc.usp.br/nilc/index.html

sumário propriamente dito. Durante a Síntese podem ser utilizados métodos de tratamento de correferência, fusão, linearização, justaposição e ordenação de sentenças.

Figura 1: Etapas de um sistema de sumarização (Sparck Jones, 1998)

Este artigo aborda a seleção de conteúdo que é uma das etapas da fase de Transformação.

3. RST

A RST foi proposta por Mann e Thompson (1987) como uma teoria descritiva dos principais aspectos da organização de um texto. A ideia principal é que um texto coerente é formado por unidades mínimas de discurso (*Elementary Discourse Units* ou proposições) que estão ligadas umas às outras, por meio de relações retóricas (relações de coerência ou discurso). Cada proposição tem uma importância e é classificada como **núcleo** (informação principal) ou **satélite** (informação adicional).

Mann e Thompson (1987) estabeleceram um conjunto de 23 relações retóricas que podem ser aplicadas a uma variedade de textos. Nesse conjunto, cada relação é classificada em **semântica** (*subject-matter*) ou **intencional** (*presentational*). As relações semânticas são aquelas que informam o leitor sobre algo. As relações intencionais alteram a inclinação do leitor para algo. Outros pesquisadores modificaram e/ou complementaram o conjunto de relações buscando maior clareza das relações, por exemplo, Marcu (1997) e Pardo (2002). Marcu adicionou as relações **estruturais** que auxiliam na estruturação dos textos. Pardo, por sua vez, definiu um conjunto composto das relações de Mann e Thompson e algumas de Marcu, totalizando 32 relações.

Uma análise RST é mapeada na forma de árvore, cujos nós-folha representam as proposições e os nós internos simbolizam relações retóricas entre as proposições. As relações que se estabelecem entre um núcleo e um satélite são chamadas de **mononucleares**. As relações entre proposições de mesmo grau de importância (entre núcleos) são ditas **multinucleares**. A Tabela 1 apresenta o conjunto utilizado por Pardo e a classificação de cada relação. Nessa tabela, as relações marcadas com um asterisco são multinucleares.

A teoria já foi bastante explorada, o que contribuiu para construção de córpus anotados, analisadores discursivos, métodos e sistemas de sumarização. Para a língua portuguesa foram construídos vários córpus: CórpusTCC (Pardo e Nunes, 2004)[2], Rhetalho (Pardo e Seno, 2005)[3], Summ-it (Collovini et al., 2007) e o CSTNews (Aleixo

[2] http://www.icmc.usp.br/~taspardo/CorpusTCC.zip
[3] http://www.icmc.usp.br/~taspardo/rhetalho.zip

e Pardo, 2008b)[4]. Destes, o CSTNews é o maior córpus multidocumento disponível anotado com RST que se tem conhecimento, composto de 50 grupos de textos jornalísticos de domínios variados. Os textos do córpus estão organizados por assunto, sendo que cada grupo de textos possui em média 3 textos, o sumário manual e a anotação RST de cada texto. O CSTNews será utilizado nesta proposta de SA multidocumento.

Tabela 1: Conjunto de relações definido por Pardo (2005)

Relação	Tipo de relação	Relação	Tipo de relação
Antithesis	Intencional	*Motivation*	Intencional
Attribution	Estrutural	*Non-volitional cause*	Semântica
Background	Intencional	*Non-volitional result*	Semântica
Circumstance	Semântica	*Otherwise*	Semântica
Comparison	Semântica	*Parenthetical*	Estrutural
Concession	Intencional	*Purpose*	Semântica
Conclusion	Semântica	*Restatement*	Semântica
Condition	Semântica	*Solutionhood*	Semântica
Elaboration	Semântica	*Summary*	Semântica
Enablement	Intencional	*Volitional cause*	Semântica
Evaluation	Semântica	*Volitional result*	Semântica
Evidence	Intencional	*Contrast* *	Semântica
Explanation	Semântica	*Joint* *	Semântica
Interpretation	Semântica	*List* *	Semântica
Justify	Intencional	*Same-Unit* *	Estrutural
Means	Semântica	*Sequence* *	Semântica

A Figura 2 apresenta um trecho de texto e a sua árvore retórica, extraídos do córpus CSTNews. A marcação entre colchetes no texto corresponde ao número identificador de cada proposição e é utilizada na árvore. Na árvore, as proposições marcadas com N são núcleos e com S são satélites. Nesse exemplo, as proposições foram relacionadas por *Sequence*, que relaciona eventos que apresentam sucessão temporal; *Non-Volitional Cause*, na qual o S apresenta uma situação que pode ter causado o N; e *Attribution*, na qual o N representa uma fala e o S apresenta a fonte da fala.

[1] Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. **[2]** Segundo uma porta-voz da ONU, **[3]** o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.

[4] A aeronave se chocou com uma montanha **[5]** e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.

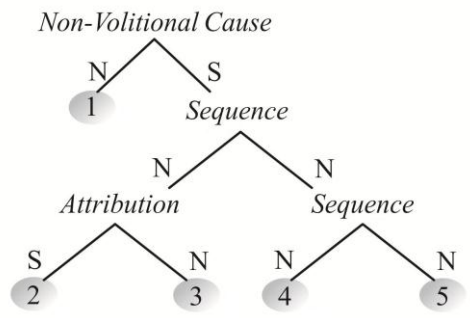

Figura 2: Exemplo de árvore RST (CSTNews)

[4] http://www.icmc.usp.br/~taspardo/sucinto/cstnews.html

A anotação RST do CSTNews foi concluída recentemente para viabilizar recursos para a execução desta proposta. O processo de anotação manual contou com 8 anotadores que utilizaram a ferramenta da edição RSTTool (O' Donnel, 2000). Apesar de existirem analisadores discursivos automáticos (ver Pardo e Nunes, 2008), a escolha pela anotação manual foi para garantir a consistência da tarefa e evitar que erros na anotação automática deturpem os resultados dos métodos que serão propostos. A concordância da anotação foi calculada sobre os critérios de segmentação do texto, níveis internos da árvore, a nuclearidade de cada segmento e a relação entre proposições. As métricas utilizadas foram precisão (P), cobertura (C) e *F-measure* (F). A Tabela 2 apresenta a média da anotação de cada critério para o córpus CSTNews.

Tabela 2: Concordância da anotação RST do córpus CSTNews

Critérios Avaliados	P	C	F
Segmentos	0,91	0,91	0,90
Níveis internos da árvore	0,78	0,78	0,78
Nuclearidade	0,78	0,78	0,78
Relações	0,66	0,66	0,66

A RST oferece vantagens para SA monodocumento por identificar o núcleo como a informação mais saliente quando comparado com o satélite, que, por sua vez, em algumas situações, pode ser omitido sem prejuízo para a interpretação do texto. Essas informações destacam quais proposições devem compor o sumário. Com base nisso, surgiram diversos métodos de sumarização monodocumento, por exemplo, Ono et al. (1994), O'Donnell (1997), Marcu (1997), Uzêda et al. (2010), etc. Uzêda et al. realizaram vários experimentos comparativos entre esses métodos e concluíram que tais métodos apresentavam performance comparável. Neste trabalho, optou-se pelo método de Marcu por servir de base para várias pesquisas de SA e já ter sido aplicado em textos da língua portuguesa (Pardo, 2002; Seno, 2005).

Marcu propõe o uso de um conjunto promocional que é formado pelas unidades mais salientes de cada nó interno da árvore. O conjunto promocional para cada nó da árvore é construído de maneira *bottom-up*, de forma que: o conjunto promocional de uma folha é composto dela mesma; e cada nó interno da árvore inclui em seu conjunto promocional a união dos conjuntos promocionais de seus filhos nucleares. Após conhecer os conjuntos promocionais de uma árvore, Marcu contabiliza o número de níveis da árvore e atribui esse valor como uma nota para a raiz. Em seguida, percorre-se a árvore em direção ao segmento sob avaliação e cada vez que o segmento não está no conjunto promocional de um nó durante o percurso, o segmento tem a pontuação decrementada de 1. A ideia dessa abordagem é que as unidades textuais mais nucleares (bem pontuadas) são as informações mais relevantes para compor o sumário. A Figura 3 apresenta a árvore retórica da Figura 2, com a pontuação em negrito de cada unidade textual, adquirida pelo método de Marcu. Uma vez obtida a importância das proposições, o sistema pode gerar sumários de diferentes tamanhos, respeitando a pontuação e a taxa de compressão. Conforme a Figura 3, a ordem de importância parcial de cada proposição é $1 > \{3, 4, 5\} > 2$, onde > indica prioridade a esquerda, ou seja, a proposição 1 é mais importante do que as demais proposições. Esta ordenação é

chamada de parcial devido a presença de proposições com a mesma pontuação no conjunto de proposições.

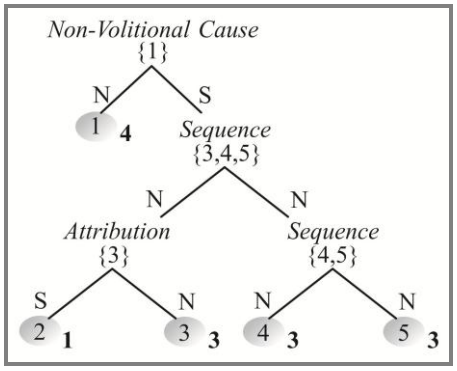

Figura 2: Árvore RST com pontuação pelo método de Marcu (1997)

4. CST

A CST (Radev, 2000) é um modelo semântico-discursivo multidocumento, formado por um conjunto de relações que permitem identificar similaridades, contradições, variações de estilos de escrita e informações complementares entre textos que descrevem o mesmo assunto. As relações são de naturezas diversas e podem ocorrer entre palavras, sintagmas, sentenças e textos.

Assim como aconteceu com a RST, pesquisadores experimentaram a CST e fizeram modificações no conjunto de relações. Inicialmente, Radev havia sugerido 24 relações, mas Zhang et al. (2003) realizaram um experimento com textos da língua inglesa, perceberam que algumas relações eram ambíguas e propuseram um refinamento para 18 relações. Aleixo e Pardo (2008b) aplicaram o conjunto de Zhang et al. para textos da língua portuguesa e fizeram um novo refinamento para 14 relações. A partir desse refinamento, Maziero et al. (2010) dividiram as relações em dois grupos principais e determinaram uma tipologia das relações. O primeiro grupo inclui as relações cuja principal finalidade é relacionar o conteúdo de segmentos e o segundo grupo contêm as relações de apresentação e forma. Cada grupo foi dividido em mais categorias. No grupo de conteúdo, as relações são classificadas em redundância, complemento ou contradição. O subgrupo redundância é dividido em dois subgrupos: de redundância total e redundância parcial. O subgrupo complemento se refere a fatos temporais ou não. O grupo de apresentação e forma é dividido em dois subgrupos: de fonte/autoria e estilo. Algumas relações CST possuem direcionalidade (relações assimétricas) e outras não (simétricas). Por exemplo, a relação *Contradiction* é simétrica, pois não importa a ordem que as sentenças são lidas: uma sempre contradiz a outra. Já a relação *Attribution* é assimétrica, pois as sentenças compartilham a mesma informação, mas somente uma delas possui a fonte. A Figura 4 apresenta a tipologia de relações CST e as relações simétricas são identificadas por um asterisco.

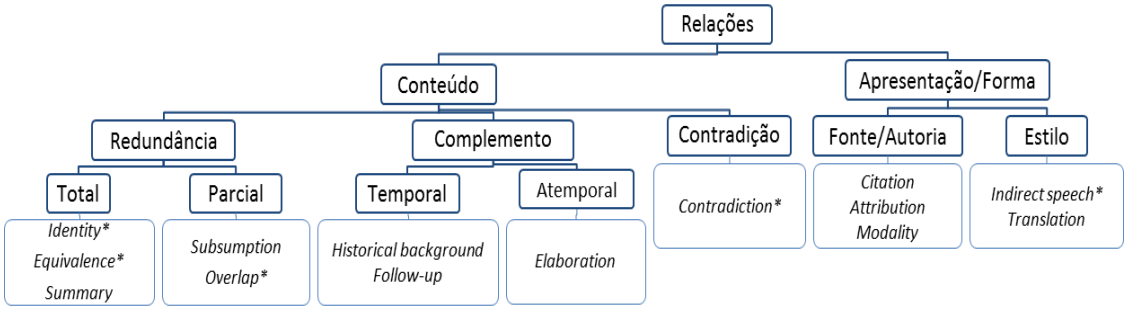

Figura 3: Tipologia das relações CST (Maziero et al., 2010)

A Figura 5 apresenta um exemplo da relação *Contradiction*. Neste exemplo, o trecho de texto 1 diverge do trecho de texto 2 quanto ao número de mortos e feridos e quanto a fonte da informação.

Trecho de texto 1	***Trecho de texto 2***
Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.	Por causa da explosão, morreram dez pessoas e 31 foram hospitalizadas - declarou à agência "Interfax" o alto funcionário de Moscou, Vladimir Resin.

Figura 4: Exemplo da relação *Contradiction* (CSTNews)

Dos recursos para língua portuguesa que utilizam CST, há o córpus CSTNews (Aleixo e Pardo, 2008b), o mesmo que foi descrito na Seção 3. Além das informações monodocumento, o córpus possui anotação CST, sumários multidocumento automáticos e manuais para cada grupo. A anotação CST contou com a participação de 4 anotadores que empregaram as relações da Figura 4.

Para facilitar a anotação do córpus CTNews, foi desenvolvida a ferramenta CSTTool (Aleixo e Pardo, 2008a). A CSTTool recebe na entrada dois textos, segmenta-os em sentenças, calcula a semelhança lexical entre todos os pares de sentença (as sentenças são de diferentes textos) e indica os melhores pares para o anotador, que, por sua vez, seleciona manualmente a relação mais adequada(s) entre as sentenças. O cálculo da semelhança lexical é necessário para diminuir o espaço de busca por sentenças relacionadas, pois Zhang et al. (2003) verificaram que as relações CST ocorrem entre pares de sentenças lexicalmente similares. Zhang et al. investigaram várias medidas de similaridade e concluíram que a medida *Word Overlap* é eficiente nessa tarefa. Por isso, a medida *Word Overlap* foi inserida na ferramenta CSTTool, facilitando assim o trabalho do anotador. A concordância das relações indicadas pelos anotadores, a direcionalidade das relações e o tipo das relações selecionadas, foi calculada com a medida kappa (Carletta, 1996). A Tabela 4 mostra os resultados dessa avaliação, os quais foram considerados bons por se tratar de uma tarefa subjetiva.

Tabela 4: Concordância da anotação CST do CSTNews

Parâmetros da concordância	***Valor de concordância***
Relações	0.50
Direcionalidade	0.44
Tipo da relação	0.61

Como dito anteriormente, a CST fornece as bases para sumarização multidocumento ao identificar quais são as informações relevantes entre as fontes. Pesquisadores que utilizaram a CST para SA alcançaram bons resultados quanto à informatividade e qualidade dos sumários (Otterbacher et al., 2002; Zhang et al., 2002; Jorge e Pardo, 2010; Jorge et al., 2011). Destes, Jorge e Pardo (2010) utilizaram a CST para sumarizar os textos do córpus CSTNews. Os autores utilizaram estratégias de seleção de conteúdo para mapear as preferências de sumarização do usuário às relações previstas na CST. As estratégias de sumarização foram formalizadas e codificadas na forma de operadores de seleção de conteúdo, representados como *templates* contendo regras específicas em termos de condições, restrições e operações primitivas de manipulação de informação. O sistema produz um sumário multidocumento pelo ranqueamento de sentenças com maior número de relações CST. Em outra linha, Jorge et al. (2011) combinaram os relacionamentos CST com características superficiais, como o tamanho e a posição da sentença, em uma abordagem de aprendizado de máquina para SA multidocumento, representando a tarefa de seleção de conteúdo como um problema de classificação supervisionada. O experimento foi conduzido sobre o córpus CSTNews e os resultados mostraram que as características linguísticas ajudam a produzir um modelo de classificação melhor.

Além de córpus e estratégias de SA com CST, está em desenvolvimento junto ao grupo de pesquisa em que se insere este trabalho, um projeto que utiliza técnicas de aprendizado de máquina para identificar as relações em textos em português (Maziero et al., 2010). Essa iniciativa facilitará a criação de outros córpus e experimentos com base em CST, e ainda, poderá contribuir para aumentar o número de projetos de SA para o português do Brasil.

5. Métodos de Sumarização Multidocumento com RST e CST

Neste artigo são feitas duas propostas de SA multidocumento, com o foco na etapa de Transformação, especificamente a tarefa de seleção de conteúdo.

5.1 Método 1: sentenças com mais relações CST e poda de satélites

Uma das propostas, referenciada por Método 1, parte das sentenças com mais relações CST para depois aplicar a poda dos satélites. A CST indica as sentenças mais relevantes do conjunto de textos e a RST aponta as proposições mais salientes de cada texto. A Tabela 5 apresenta o algoritmo do Método 1, com seus passos já divididos nas etapas de um sistema de sumarização.

A fase de Análise do Método 1 inicia-se com a anotação CST de cada grupo de textos do CSTNews, seguida da anotação individual de cada texto com RST. Vale ressaltar que esta fase foi previamente concluída. Na fase de Transformação, especificamente no passo 3, calcula-se a taxa de compressão, que é a razão entre o tamanho do sumário e o tamanho do texto-fonte (Mani, 2001). Neste trabalho, a taxa de compressão é de 70% sobre o tamanho do maior texto do grupo, medido em número de palavras. Adotou-se esse valor devido os sumários manuais do córpus seguirem essa mesma taxa. Por exemplo, em um grupo com o texto-fonte A com 180 palavras e o texto-fonte B com 125 palavras, a taxa de compressão é calculada sobre o tamanho de A. Neste caso, espera-se um sumário com aproximadamente 54 palavras.

Tabela 5: Método 1 de sumarização

ANÁLISE	*1.* Analisar cada grupo de texto usando a CST *2.* Analisar cada texto usando a RST
TRANSFORMAÇÃO	*3.* Calcular taxa de compressão *4.* Selecionar a sentença com mais relações CST *5.* Eliminar os satélites da estrutura RST da sentença selecionada *6.* Se ainda houver espaço (se a taxa de compressão não tiver sido atingida), faça: *6.a* Selecionar a próxima sentença com mais relações CST *6.b* Verificar a redundância da sentença podada com as sentenças que já foram selecionadas antes: *6.b.1* Se a relação for do grupo de redundância *total*: as sentenças devem ser tratadas, para evitar a redundância no sumário. *6.b.2* Se a relação for do grupo de redundância *parcial*: Se a relação for *Overlap*, manter a sentença. Se a relação for *Subsumption*, manter a sentença que engloba a outra e esta, será excluída. *6.b.3* Se for qualquer outra relação, manter a sentença *6.c* Podar a estrutura RST dessa sentença *7.* Repetir passo 6 até satisfazer a taxa de compressão especificada
SÍNTESE	*8.* Realizar fusão de segmentos *9.* Justapor segmentos para formar o sumário *10.* Ordenar segmentos

No passo 4, calcula-se o número de relações CST que cada sentença recebeu, organiza-se um *ranking* ordenado, iniciando com as sentenças que tem mais relações CST. Depois, seleciona-se a primeira sentença do *ranking* e no passo 5, aplica-se a poda dos satélites desta sentença e a parte restante passa para o grupo de sentenças do sumário.

No passo 6, verifica-se se a taxa de compressão foi preenchida. Se sim, encerra-se a busca de sentenças. Se a taxa de compressão ainda não foi preenchida, (passo 6a) seleciona-se a próxima sentença do ranque CST. Esta nova sentença pode pertencer a outro texto-fonte do grupo, ou seja, a origem da sentença atual pode não ser a mesma da sentença escolhida no passo 4. Neste caso, (passo 6b) deve-se verificar qual a relação existente entre as sentenças já selecionadas para o sumário com a sentença candidata. Se a redundância for total (passo 6.b.1), deve-se aplicar o tratamento descrito na Figura 6. Por outro lado, se a redundância for parcial (passo 6.b.2), a relação será de *Overlap* ou *Subsumption*. Se a relação for *Overlap*, a sentença candidata é aprovada para o grupo de sentenças do sumário. Se for *Subsumption*, a sentença que engloba a outra sentença é aprovada para o sumário e a outra sentença será eliminada. Neste caso, pode haver troca com a sentença anteriormente selecionada para o sumário. Mas, se for qualquer outra relação CST, a sentença candidata torna-se aprovada para o grupo de sentenças do sumário (passo 6.b.3). Após isso, ocorre a poda dos satélites desta sentença (passo 6.c). O passo 6 se repete até que a taxa de compressão seja atingida.

Tratamento para os relacionamentos de redundância total
a) se a relação entre duas sentenças for *Identity*, fica a sentença já selecionada, e descarta-se a candidata. b) se a relação entre duas sentenças for *Equivalence* ou *Summary*, fica a sentença com o menor número de palavras. Neste caso, pode haver troca com a sentença que havia sido anteriormente selecionada para o sumário.

Figura 6: Decisões para eliminar a redundância

Na fase de Síntese, poderá haver fusão (8), justaposição (9) e ordenação dos segmentos (10) que foram selecionados para o sumário. A fusão de sentenças consiste em produzir, a partir de conjunto de sentenças relacionadas, uma nova sentença que resume as informações comuns apresentadas no conjunto. Nessa etapa, poderá ser utilizado o sistema de fusão para o português de Seno e Nunes (2009) para tratar as sentenças relacionadas por *Overlap* que foram inseridas no sumário. A justaposição e ordenação tratam da organização das sentenças usando uma sequência lógica. A fase de Síntese será realizada parcialmente.

Uma observação é quanto à poda de proposições, realizada sempre que uma sentença selecionada contem satélites. Nos primeiros experimentos, observou-se alguns casos que, após a poda dos satélites de sentenças que tinham alguma relação CST entre si, aquela relação CST era eliminada ou novas surgiam. Isso pode ser observado na Figura 7, que contêm duas sentenças, chamadas de D1S1 e D2S3, sendo que D representa o texto-fonte e S a sentença. As sentenças foram segmentadas em proposições e são identificadas por números entre colchetes. A Figura 7 mostra a árvore RST das sentenças e o relacionamento CST, este representado pelas linhas entre as duas árvores. Aplicando-se a poda de satélites nessas sentenças, marcado pela linha tracejada, a fonte da informação que está em D1S1 será perdida, eliminando a relação de *Attribution* que ocorre entre os dois textos. Entre as informações nucleares que sobram, a relação de *Overlap* também é eliminada e neste caso, o analista poderia achar que não surge uma nova relação CST ou que surge certa sequência, por exemplo.

D1S1:	**D2S3:**
[1] Segundo uma porta-voz da ONU, **[2]** o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.	**[6]** Todos morreram **[7]** quando o avião, **[8]** prejudicado pelo mau tempo, **[9]** não conseguiu chegar à pista de aterrissagem e **[10]** caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.

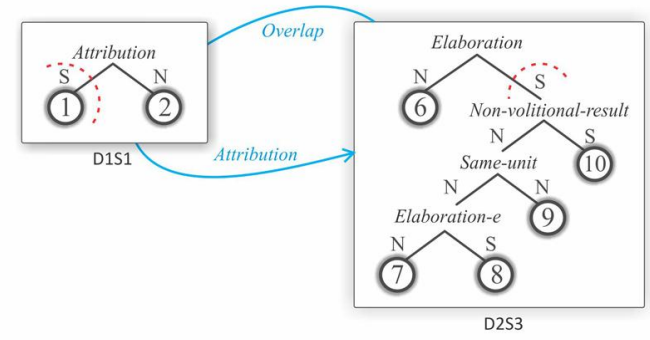

Figura 7: Exemplo de relacionamento RST e CST entre sentenças (CSTNews)

Outro exemplo de mudança de relação após a poda é dado na Figura 8, na qual as duas sentenças D2S2 e D1S3 são relacionadas por *Subsumption.* Com a poda dos satélites, a relação *Subsumption* é eliminada, e observa-se que uma das possíveis relações CST que surge é *Overlap*, mas o analista também poderia apontar outras relações.

D2S2:	**D1S3:**
[S] Internado em um hospital em Buenos Aires, **[N]** ele teve uma recaída e voltou a sentir dores **[S]** devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.	**[N]** Maradona teve uma recaída de hepatite aguda.

Figura 8: Exemplo de mudança de relação CST entre sentenças (CSTNews)

Sabe-se que o ideal seria realizar nova anotação RST e CST após a poda, mas no estágio atual isso não é possível. Os fatores que contribuem para que não seja feita uma nova anotação são: a) a hipótese de que, ao perder algumas relações e surgirem outras, isso não irá afetará significativamente o processo de sumarização multidocumento; e b) a anotação depende de esforço humano, o que torna a tarefa cara. Nesta situação, decidiu-se que nenhuma alteração será feita nas anotações RST e CST. Se ocorrer algo similar aos exemplos das Figuras 7 e 8, devem-se manter as anotações, independente de alguma relação ser eliminada após a poda de satélites.

Um dos problemas em sumarização multidocumento está relacionado à redundância que, por sua vez, tem influência direta na manutenção da coerência do sumário. A forma definida para tratar isso foi descrita no Método 1, especificamente no passo 6. Para ilustrar o uso daquelas decisões, observe as sentenças da Figura 9. As duas sentenças que são de textos diferentes, relatam um acidente em Moscou e entre elas há uma relação de *Subsumption* (redundância parcial). Suponha que D3S5 já havia sido selecionada para o sumário, e que a próxima sentença candidata pelo passo 6 é D2S1. Observa-se que D2S1 engloba a sentença do sumário. Neste caso, a decisão tomada será de trocar a sentença do sumário pela sentença candidata, que manterá a mesma informação de D3S5 e ainda acrescentará mais detalhes.

Sumário	***Sentença candidata***
[D3S5] Anteriormente, a Polícia havia informado sobre nove mortos, sendo três deles crianças, e 25 feridos.	**[D2S1]** Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.

Figura 9: Exemplo de relacionamento (CSTNews)

5.2 Método 2: conjunto promocional e sentenças com mais relações CST

A segunda proposta, referenciada por Método 2, inicia com aplicação do método de Marcu (descrito na Seção 3) sob a árvore RST de cada texto. Após a poda, as sentenças são selecionadas de acordo com o número de relacionamentos CST. A Tabela 6 apresenta o algoritmo do Método 2, com seus passos já divididos nas etapas de um sistema de sumarização.

Tabela 6: Método 2 de sumarização

ANÁLISE	*1.* Analisar cada texto usando a RST *2.* Podar as estruturas RST dos textos pelo método de Marcu *3.* Analisar cada grupo de texto usando a CST
TRANSFORMAÇÃO	*4.* Calcular taxa de compressão *5.* Selecionar a sentença com mais relações CST *6.* Se ainda houver espaço (se a taxa de compressão não tiver sido atingida), faça: *6.a* Selecionar a próxima sentença com mais relações CST *6.b* Verificar a redundância da sentença podada com as sentenças que já foram selecionadas antes: *6.b.1* Se a relação for do grupo de redundância *total*: as sentenças devem ser tratadas, evitando a redundância no sumário. *6.b.2* Se a relação for do grupo de redundância *parcial*: Se a relação for *Overlap*, manter a sentença. Se a relação for *Subsumption*, manter a sentença que engloba a outra *6.b.3* Se for qualquer outra relação, manter a sentença *7.* Repetir passo 6 até satisfazer a taxa de compressão especificada
SÍNTESE	*8.* Realizar fusão de segmentos *9.* Justapor segmentos para formar o sumário *10.* Ordenar segmentos

Na etapa de Análise, os passos 1 e 3 estão concluídos, pois utilizam a anotação disponível no córpus CSTNews, da mesma forma como no Método 1. No passo 2, aplica-se o método de Marcu (1997) em todas as árvores retóricas do conjunto para conhecer a pontuação de cada proposição. Após, organiza-se as proposições ordenadas pela pontuação e elimina-se o complemento da taxa de compressão em cada texto. Como a taxa de compressão é a mesma do Método 1, neste caso o complemento corresponde a 30% sobre o número de proposições menos importantes. Acredita-se que, com os segmentos restantes ainda, teremos informação suficiente para formar o sumário. Por exemplo, um texto segmentado em 5 proposições, após eliminar 30% das proposições, ainda terá 4 proposições candidatas ao sumário. A Figura 10 mostra a mesma árvore Figura 3, mas indicando pela linha tracejada qual proposição será eliminada pelo método de Marcu devido a sua pontuação ser muito baixa.

Na fase de Transformação, no passo 4, é calculada a taxa de compressão da mesma forma que no Método 1, considerando o total de palavras do maior texto. No passo 5 é selecionada a primeira sentença candidata ao sumário. Como no Método 1, o critério de selecionar as sentenças com mais relações CST permanece. Nesse ponto do algoritmo pode acontecer que após a poda RST (passo 2) a sentença candidata tenha sido totalmente eliminada na poda RST ou que tenha restado somente parte da sentença. Se a sentença tiver sido totalmente eliminada, busca-se a próxima sentença do *ranking* CST, senão, insere-se no sumário o que sobrou da sentença.

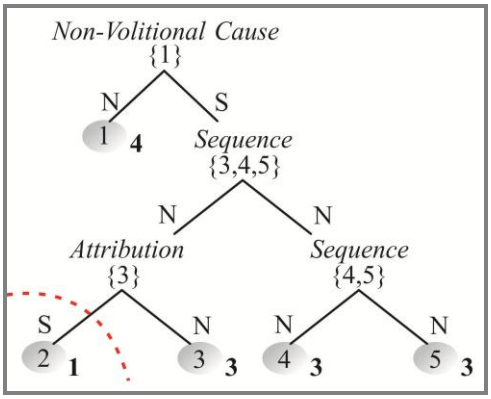

Figura 10: Árvore RST com a poda da proposição 2

Após a seleção da primeira sentença, no passo 6, verifica-se se ainda há espaço para buscar mais sentenças. Se ainda houver, (a) busca-se a próxima sentença com mais relações CST e (b) verifica se há alguma redundância com a(s) sentença(s) candidata(s). Neste ponto, a fase de Transformação é similar aquela apresentada no Método, com a diferença de que a poda já foi realizada na fase de Análise.

A fase de Síntese é similar a fase do Método 1.

6. Considerações Finais

Este artigo apresentou dois métodos de SA multidocumento baseados nas teorias RST e CST que estão sendo investigados. O foco de ambos os métodos é a seleção de conteúdo.

Além dos métodos apresentados, pretende-se adotar um critério de peso para as relações CST. Os pesos serão definidos em função da redundância: quanto mais redundante, mais peso receberá a relação. A Figura 11 apresenta as relações CST e seus respectivos pesos. Inicialmente os pesos foram definidos da seguinte forma:

a) As relações de redundância total ocorrem entre as informações que são muito repetidas entre as fontes, e, portanto, são consideradas como informações relevantes. As relações deste grupo recebem peso máximo de 1.0.

b) As relações de redundância parcial ocorrem entre informações que contêm informações em comum e também alguma informação nova. Como é difícil detectar estas diferenças, atribuiu-se o peso médio 0.5.

c) A relação *Contradiction* ocorre entre sentenças que, apesar de contraditórias, apresentam alguma informação em comum e, neste caso, recebem uma pontuação intermediária de 0.8.

d) As relações de complemento conectam sentenças que trazem informações complementares, uma sobre a outra, e neste caso, as relações deste grupo recebem a pontuação 0.5.

e) As relações de apresentação e forma não atuam diretamente sobre o conteúdo e, por isso, recebem a pontuação 0.1.

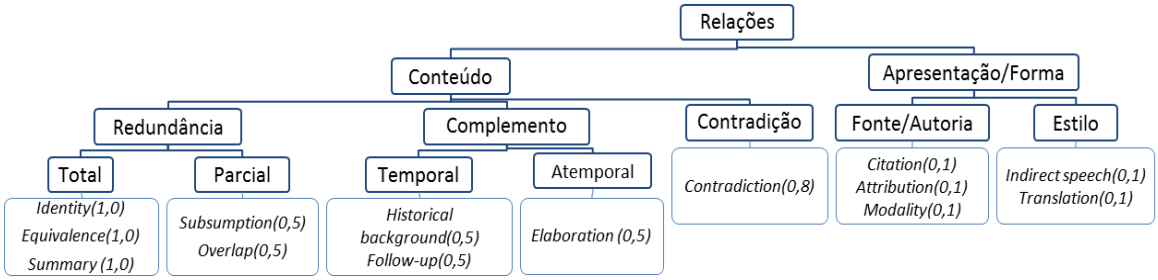

Figura 11: Proposta de pesos para as relações CST

Como trabalhos futuros, será feita a implementação dos métodos descritos na Seção 5 e os sumários automáticos serão avaliados por métricas clássicas da área. Com base nos resultados, poderão ainda serem exploradas outras formas de combinar as teorias RST e CST para sumarização multidocumento.

Por meio deste trabalho, será possível mostrar o impacto que o uso de conhecimento semântico-discursivo tem sobre a tarefa de sumarização ao combinar as teorias semântico-discursivas. Espera-se concluir a pesquisa apresentando alguns métodos de SA multidocumento que gerem sumários informativos.

Agradecimentos

Agradecemos ao CNPq, CAPES e FAPESP pelo suporte financeiro.

Referências

Afantenos, S.D.; Karkaletsis, V; Stamatopoulos, P; Halatsis, C. (2007). Using synchronic and diachronic relations for summarization multiple documents describing evolving events. *Journal of Intelligent Information Systems*, 30(3):183–226.

Aleixo, P. e Pardo, T.A.S. (2008a). *CSTTool: Uma Ferramenta Semi-automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 321. São Carlos/SP, 14p.

Aleixo, P. e Pardo, T.A.S. (2008b). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos/SP, 15p.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistic.* Vol. 22 (2), pp. 249-254.

Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Rino, L.; Vieira, R. (2007). Summ-it: um córpus anotado com informações discursivas visando à sumarização automática. In *Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL'2007. Rio de Janeiro/RJ.

Jorge, M.L.C.; Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-*

based Methods for Natural Language Processing, pp. 74-82. July 16, Uppsala/Sweden.

Jorge, M.L.C.; Agostini, V.; Pardo, T.A.S. (2011). Multi-Document Summarization Using Complex and Rich Features. *VII Encontro Nacional de Inteligência Artificial* – ENIA. XXXI Congresso da Sociedade Brasileira de Computação. Natal-RN, Brasil.

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

Mann, W.C.; Thompson, S.A. (1987). *Rhetorical Structure Theory: Toward a functional theory of text organization.* Vol. 8, N. 3,243-281p

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD Thesis, Department of Computer Science, University of Toronto.

Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010). Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* - NLPCS, pp. 60-69. Funchal/Madeira, Portugal.

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.

O'Donnell, M. (2000). RSTTool 2.4 - A Markup Tool for Rhetorical Structure Theory. In the *Proceedings of the International Natural Language Generation Conference* (INLG'2000), 13-16 June 2000, Mitzpe Ramon, Israel.

Ono, K.; Sumita, K.; Miike, S. (1994). Abstract Generation Based on Rhetorical Structure Extraction. In the *Proceedings of the International Conference on Computational Linguistic* – Coling-94, pp. 344-348, Japan.

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.

Pardo, T.A.S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos/SP.

Pardo, T.A.S. e Nunes, M.G.V. (2004). *Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 231. São Carlos/SP, Abril, 73p.

Pardo, T.A.S. (2005). *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP, Fevereiro, 8p.

Pardo, T.A.S. e Seno, E.R.M. (2005). Rhetalho: um corpus de referência anotado retoricamente. In *Anais do V Encontro de Corpora*. São Carlos/SP, Brasil. 25 a 26 de Novembro.

Pardo, T.A.S. and Nunes, M.G.V. (2008). On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, Vol. 15, N. 2, pp. 43-64.

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.; Otterbacher, J.; Zhang, Z. (2004). CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In *the Proceedings of Fourth International Conference on Language Resources and Evaluation*.

Seno, E.R.M. (2005). *Especificação de Heurísticas de Sumarização de Estruturas RST com Base na Preservação dos Elos Co-Referenciais*. 2005. Dissertação (Mestrado) – Departamento de Computação, Universidade Federal de São Carlos, São Carlos.

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.

Sparck Jones, K. (1998). Automatic summarizing: factors and directions. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in automatic text summarization*, chapter 1, pp. 1-12.

Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2010). A Comprehensive Comparative Evaluation of RST-Based Summarization Methods. *ACM Transactions on Speech and Language Processing*, Vol. 6, N. 4, pp. 1-20.

Zhang, Z.S.; Blair-Goldensohn, S.; Radev, D.R. (2002). Towards CST-Enhanced Summarization. In the *Proceedings of AAAI 2002 Conference*. Edmonton, Alberta.

Zhang, Z.; Otterbacher, J.; Radev, D. (2003) Learning cross-document structural relationships using boosting. In *Proccedings of the Twelfth International Conference on Information and Knowledge Management CIKM 2003*, pages 124–130, New Orleans, Louisiana, USA.