Estratégias de Seleção de Conteúdo com Base na CST (*Cross-document Structure Theory*) para Sumarização Automática Multidocumento

Maria Lucia del Rosario Castro Jorge, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400 - Centro
Caixa Postal: 668 - CEP: 13560-970 - São Carlos/SP, Brasil

{mluciacj,taspardo}@icmc.usp.br

Resumo

O presente trabalho apresenta a definição, formalização e avaliação de estratégias de seleção de conteúdo para sumarização automática multidocumento com base na teoria discursiva CST (*Cross-document Structure Theory*). A tarefa de seleção de conteúdo foi modelada por meio de operadores que representam possíveis preferências do usuário para a sumarização. Estes operadores são especificados em templates contendo regras e funções que relacionam essas preferências às relações CST. Em particular, definimos operadores para extrair a informação principal, apresentar informação de contexto, identificar autoria, tratar redundâncias e identificar informação contraditória. Nossos experimentos foram feitos usando um córpus jornalístico de textos escritos em português brasileiro e mostram que o uso da CST melhora a qualidade do conteúdo selecionado para os sumários, já que se exploram as relações entre os conteúdos dos diferentes textos.

1. Introdução

O uso e a disponibilidade cada vez maior de tecnologias de comunicação têm provocado um aumento considerável no volume de informação, principalmente on-line. Há muita informação redundante, complementar e contraditória, proveniente de diversas fontes. Conseqüentemente, o processamento dessa informação tem se tornado uma tarefa de difícil execução, tanto por humanos quanto por máquinas. Neste contexto, a sumarização multidocumento pode ser uma tarefa útil.

A sumarização automática multidocumento (SAM) consiste na produção automática de um único sumário (também chamado resumo) a partir de um grupo de textos sobre um mesmo tópico ou sobre tópicos relacionados (Mani, 2001). Imagine, por exemplo, que uma pessoa deseje se interar dos principais acontecimentos da recente crise econômica mundial. Em vez de ter que ler uma infinidade de textos sobre o assunto, o que seria inviável, um sistema de SAM poderia lhe fornecer um único sumário sintetizando os fatos relevantes. A Figura 1 mostra um exemplo de sumário multidocumento produzido manualmente a partir de três textos jornalísticos que reportavam diversos ataques criminosos organizados a várias regiões do estado de São Paulo, no Brasil.

> Uma nova série de ataques criminosos foi registrada na madrugada desta segunda-feira, dia 7, em São Paulo e municípios do interior paulista. Os bandidos atacaram agências bancárias, bases policiais e prédios públicos com bombas e tiros. As ações são atribuídas à facção criminosa Primeiro Comando da Capital (PCC), que já comandou outros ataques em duas ocasiões. Eles tinham prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo. A promessa aparentemente começou a ser cumprida na madrugada de hoje. Cidades do interior, como Jundiaí, foram alvo de ataques. Na região do ABC Paulista, pelo menos dez ônibus foram incendiados - sete em Mauá e três em Santo André. Na capital, houve ataques a outros quatro ônibus. Uma bomba caseira foi jogada contra o prédio do Ministério Público, na capital do estado. A Secretaria da Fazenda também foi atingida por uma bomba. Duas bases da Guarda Civil Metropolitana (GCM), sendo uma no Capão Redondo, Zona Sul de São Paulo, foram alvo dos criminosos.

Figura 1. Exemplo de sumário multidocumento

É interessante notar que um sumário multidocumento pode ser construído tendo-se

diferentes objetivos. Se o leitor deseja apenas uma visão geral do acontecimento, o sumário do exemplo é suficiente. Por outro lado, muitas vezes se quer informação contextual (no caso de um leitor que não sabe nada do assunto) ou se deseja visualizar a evolução de alguns fatos ocorridos em um determinado período de tempo (nesses casos, o histórico da facção PCC e como ela tem agido no estado de São Paulo são elementos importantes para o sumário). Ocasionalmente, pode-se querer confrontar diferentes versões de notícias para se detectar contradições entre elas (por exemplo, quais dos ataques são atribuídos pelas três fontes ao PCC). Portanto, sistemas de SAM devem ser capazes de produzir sumários que satisfaçam as preferências de sumarização do leitor/usuário.

A sumarização multidocumento, como grande parte dos sistemas de processamento multidocumento, tem que lidar, também, com diversos desafios provenientes da multiplicidade de informação. Por exemplo, dentre os fenômenos multidocumento, é necessário que se reconheça informação redundante, complementar e, como já mencionado, contraditória, que as correferências sejam resolvidas, que estilos variados de diferentes autores e fontes sejam uniformizados, e que a informação produzida para o usuário seja organizada/ordenada adequadamente, visando-se sempre a coerência e a coesão do texto produzido.

Mani e Maybury (1999), objetivando modelar a tarefa de sumarização e organizar seus diversos processos, sugerem que a sumarização envolva idealmente três tarefas: a análise dos textos-fonte, produzindo-se uma representação completa de seu conteúdo; a transformação desse conteúdo completo em um conteúdo condensado; e, finalmente, a síntese desse conteúdo condensado na forma de sumário, expresso em uma língua natural. Uma etapa completa de análise requer, por exemplo, o uso de léxicos, gramáticas e interpretadores de língua natural de níveis lingüísticos variados; a etapa de transformação deve realizar a seleção do conteúdo relevante, a agregação/fusão, a generalização e a substituição de informação, dentre outras operações; a etapa de síntese, por fim, deve ter capacidades de geração textual, envolvendo a escolha de expressões de referência, ordenação e organização da informação a ser apresentada, etc. Sistemas de SAM que adotam tal abordagem, privilegiando a manipulação e o uso de conhecimento lingüístico sofisticado, são ditos pertencerem à abordagem profunda, ou fundamental. Sistemas que fazem uso de pouco conhecimento lingüístico são ditos pertencerem à abordagem superficial. Apesar de serem mais custosos e exigirem mais recursos, sistemas da abordagem profunda são capazes de produzir sumários melhores.

Neste trabalho, foca-se na abordagem profunda, mais especificamente, em um dos processos mais importantes da etapa de transformação da SAM: a seleção de conteúdo. Assume-se que a etapa de análise é realizada previamente e corresponde unicamente à representação dos textos-fonte segundo a teoria/modelo lingüístico-computacional CST (*Cross-document Structure Theory*) (Radev, 2000), de natureza semântico-discursiva. Com base na CST, são exploradas estratégias de seleção de conteúdo que selecionam conteúdo relevante em função de preferências de sumarização do usuário. Por fim, a etapa de síntese realiza simplesmente a justaposição do conteúdo selecionado, produzindo o sumário final. A CST é, portanto, a base de desenvolvimento deste trabalho. Ela modela o relacionamento entre o conteúdo multidocumento, ou seja, estabelece relações semântico-dicursivas entre as partes dos textos sendo processados (por exemplo, relações de seqüência temporal, contradição, elaboração, etc.). A hipótese deste trabalho é que esse tipo de conhecimento é importante e, se manipulado adequadamente, pode produzir sumários multidocumento satisfatórios.

As estratégias de seleção de conteúdo propostas neste trabalho visam mapear as preferências de sumarização do usuário às relações previstas na CST, de forma que seja possível identificar nos textos-fonte o conteúdo relevante. Em particular, nossas estratégias são formalizadas e codificadas na forma de operadores de seleção de conteúdo, representados como templates contendo regras especificadas em termos de condições, restrições e operações primitivas de manipulação de informação. Neste trabalho, definimos operadores para extrair a informação principal dos textos-fonte, apresentar

informação de contexto, identificar autoria, tratar redundâncias e exibir informação contraditória dos textos-fonte. Nossos experimentos foram feitos usando um córpus jornalístico de textos escritos em português brasileiro e mostram que o uso da CST melhora a qualidade do conteúdo selecionado para os sumários, comprovando, desta forma, nossa hipótese.

Este trabalho dá continuidade a alguns trabalhos prévios na área para a língua portuguesa (Aleixo e Pardo, 2008a; Jorge e Pardo, 2009, 2010). Por se basear em um modelo semântico-discursivo, este trabalho alinha-se, portanto, à abordagem profunda da sumarização.

A seguir, na Seção 2, introduz-se a CST e apresentam-se os trabalhos relacionados. Na Seção 3, definimos e formalizamos nossos operadores de seleção de conteúdo. A avaliação e discussão dos resultados obtidos são apresentadas na Seção 4. Por fim, na Seção 5, fazem-se algumas considerações finais.

2. Trabalhos Relacionados

2.1 Cross-document Structure Theory

Inspirada na *Rhetorical Structure Theory* (RST) (Mann e Thompson, 1987) e nos trabalhos de Trigg (1983) e Trigg e Weiser (1987), a CST (Radev, 2000) é proposta como uma teoria para relacionar múltiplos documentos que versam sobre um mesmo assunto ou tópicos relacionados.

A CST foi originalmente proposta com um conjunto de 24 relações que representam os fenômenos multidocumento. As 24 relações são listadas na Tabela 1. Como exemplo de aplicação destas relações a um grupo de textos, a Figura 2 mostra alguns trechos de textos (de fontes diferentes) relacionados. Na figura, o primeiro par de sentenças está relacionado por meio da relação *Subsumption*, pois a segunda sentença contém toda a informação da primeira e outras informações adicionais. No segundo par de sentenças, as duas sentenças são iguais, portanto há uma relação *Identity*. Finalmente, o terceiro par de sentenças mostra uma contradição entre a distância ao aeroporto, o que caracteriza uma relação *Contradiction*.

Tabela 1. Conjunto original de relações propostas por Radev (2000)

Identity	*Judgment*
Equivalence (paraphrasing)	*Fulfilment*
Translation	*Description*
Subsumption	*Reader profile*
Contradiction	*Contrast*
Historical background	*Parallel*
Modality	*Cross-reference*
Attribution	*Citation*
Summary	*Refinement*
Follow-up	*Agreement*
Elaboration	*Generalization*
Indirect speech	*Change of perspective*

Relação: *Subsumption*

Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

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

Relação: *Identity*

As vítimas do acidente foram 14 passageiros e três membros da tripulação.

As vítimas do acidente foram 14 passageiros e três membros da tripulação.

Relação: *Contradiction*

A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 10 quilômetros de distância da pista do aeroporto.

Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.

Figura 2. Exemplos de relações CST

A CST propõe um modelo geral em que as relações entre diferentes unidades de texto são representadas. Na Figura 3, ilustra-se este modelo, que assume a forma de um grafo. A figura foi reproduzida exatamente como aparece no trabalho de Radev (2000, p. 78) (em inglês, como no original). É importante notar que, em princípio, podem-se considerar diversas unidades textuais para análise, por exemplo, palavras, sintagmas, sentenças,

parágrafos ou, inclusive, todo o documento. As relações CST são estabelecidas em qualquer nível de análise. Nem todas as unidades textuais têm relações CST entre si, pois, em geral, existem partes dos textos que não estão diretamente relacionadas a um mesmo tópico. As relações estabelecidas também podem ter direcionalidade. Por exemplo, na Figura 2, as relações *Identity e Contradiction* não têm direcionalidade; por outro lado, a relação *Subsumption* tem direcionalidade, já que uma unidade textual está englobando outra.

Assim como sua antecessora RST, a CST está sujeita a ambigüidades na análise (Afantenos et al., 2004; Zhang et al., 2002), já que, como em toda análise subjetiva, pode haver mais de uma relação possível entre segmentos textuais. Com o objetivo de reduzir esta ambigüidade, Zhang et al. (2002) propuseram um refinamento das relações originais, em que são consideradas menos relações: 18. Para a língua portuguesa, o conjunto de relações foi ainda mais refinado (Aleixo e Pardo, 2008b), resultando em 14 relações. Esse refinamento foi feito pela eliminação de relações nunca verificadas experimentalmente e pela junção de relações com definições relacionadas. A Tabela 2 mostra as relações resultantes.

Tabela 2. Relações de Aleixo e Pardo (2008b)

Identity	*Attribution*
Equivalence	*Summary*
Translation	*Follow-up*
Subsumption	*Elaboration*
Contradiction	*Indirect speech*
Historical background	*Contradiction*
Modality	*Citation*

2.2 *Sumarização Multidocumento e CST*

Algumas pesquisas têm utilizado CST para fins de SAM, incluindo o trabalho do próprio Radev (2000), que, além de propor o modelo, também propôs uma metodologia de sumarização com CST de 4 etapas, as quais são ilustradas na Figura 4.

Na primeira etapa, os documentos são agrupados de acordo com a similaridade do conteúdo entre eles; na segunda etapa, os documentos são estruturados internamente, possivelmente envolvendo estruturas lexicais, sintáticas e semânticas; na terceira etapa, as relações CST são estabelecidas entre as partes dos textos e as unidades textuais relacionadas são organizadas em um grafo (que, deste ponto em diante, será referenciado por grafo CST) em que cada nó representa uma unidade informativa textual e as arestas representam as relações entre eles; finalmente, na quarta etapa, o conteúdo é selecionado de acordo com a informação dada pelas relações, para compor o sumário final. Para esta última etapa, Radev propõe a criação de operadores de preferência que representem possíveis preferências de sumarização para a seleção de conteúdo. Estas preferências estão associadas a certas relações estabelecidas pela CST. Por exemplo, um operador de contradição deveria selecionar informação relevante, além de apresentar principalmente as informações contraditórias entre os textos que estão sendo processados. Neste caso, sentenças relacionadas por meio da relação *Contradiction* terão uma preferência maior ao se selecionar o conteúdo para o sumário final. A proposta de Radev está baseada no trabalho prévio de Radev e McKeown (1998).

Outro trabalho importante baseado na CST foi o de Zhang et al. (2002). Considerando que após a seleção de conteúdo há um ranque de sentenças para compor o sumário (em função da relevância destas de acordo com uma métrica de importância qualquer), os autores propõem a alteração do ranque por meio do uso das relações CST. Sentenças que apresentam relações CST são preferidas em relação às sentenças que não apresentam tais relações e, portanto, obtêm melhores posições no ranque.

Otterbacher et al. (2002) investigam como o uso de relações CST ajuda a melhorar a coesão em sumários multidocumento. Eles propõem a seleção de sentenças de acordo com o conteúdo relevante e assumem que as sentenças relacionadas por meio de relações CST deveriam aparecer próximas no sumário final, podendo ser reorganizadas em função das restrições temporais impostas pelas próprias relações CST.

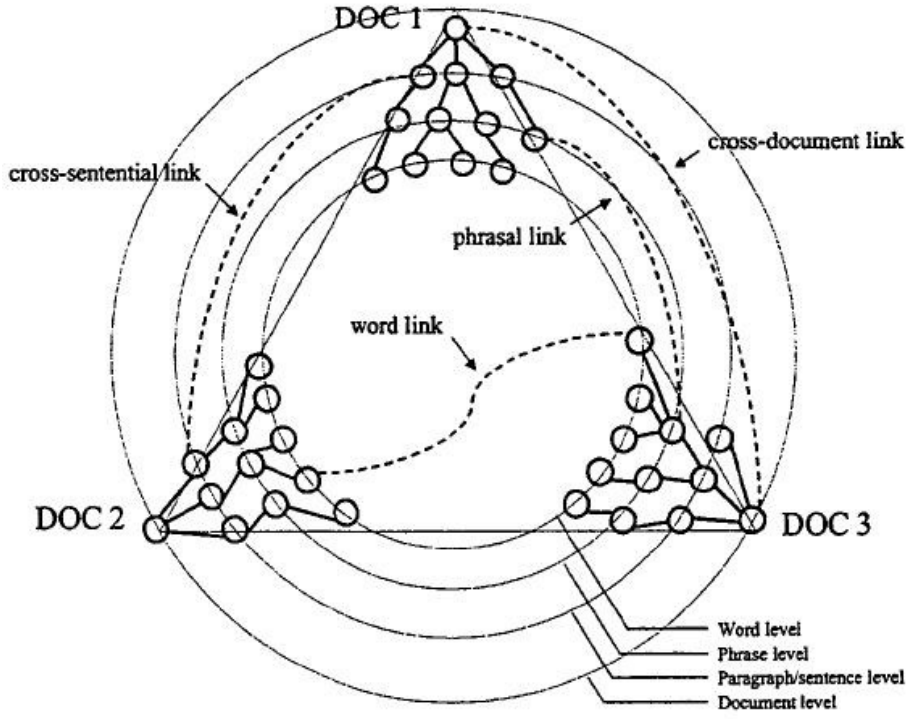

Figura 3. Modelo geral de representação via CST (Radev, 2000, p. 78)

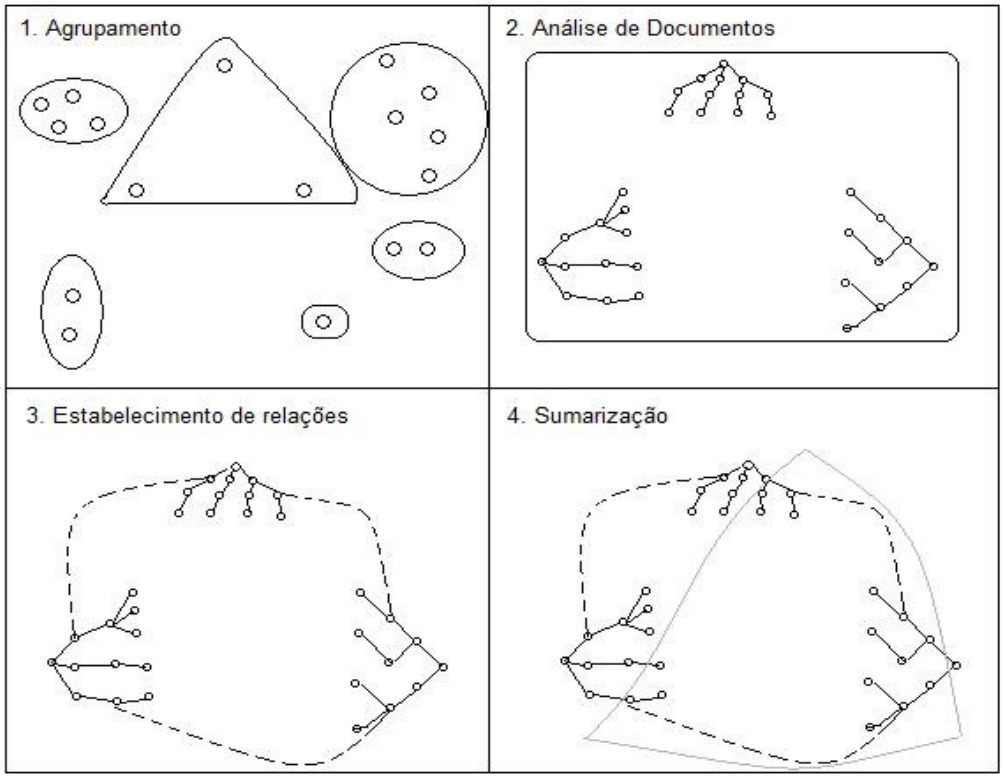

Figura 4. Etapas do processo de sumarização CST (Radev, 2000, p. 81)

Em uma linha um pouco diferente, Afantenos et al. (2004), com base na CST, propuseram uma nova classificação de relações entre textos. Os autores dividem as relações em duas categorias: sincrônicas e diacrônicas. As relações sincrônicas exploram o desenvolvimento de um evento descrito em varias fontes de informação, enquanto as relações diacrônicas exploram o

evento ao longo do tempo em uma mesma fonte de informação. De acordo com esta nova classificação, os autores propõem uma metodologia de sumarização que extrai mensagens dos textos (utilizando ferramentas de extração de informação) e as coloca em formato de *templates*, sendo que as mensagens são relacionadas pelas relações propostas. Com base nesses *templates* relacionados, os autores afirmam que é possível se produzir bons sumários. Os autores apenas apresentam essas idéias iniciais e mostram alguns exemplos para textos do domínio do esporte, mas não formalizam ou avaliam sua proposta.

A seguir, delineamos nossa proposta de seleção de conteúdo com base na CST.

3. Definição e Formalização de Operadores de Seleção de Conteúdo

Como discutido anteriormente, o objetivo deste trabalho é explorar estratégias de seleção de conteúdo para SAM, relacionando possíveis preferências de sumarização do usuário às relações da CST, modelo utilizado para representar os textos-fonte. Seguindo a proposta de Radev (2000), após definir cada estratégia de seleção de conteúdo, elas são representadas na forma de operadores.

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 condensada contendo as informações mais relevantes segundo os critérios especificados. Em particular, neste trabalho, a representação de conteúdo consiste no conjunto de textos representados segundo a teoria CST. Portanto, os operadores são aplicados após os textos-fonte terem sido analisados segundo essa teoria (na etapa de análise). Atualmente, tal análise deve ser feita manualmente para a língua portuguesa, já que o primeiro analisador automático ainda está em desenvolvimento. Para a língua inglesa, já há um analisador disponível (Zhang et al., 2003), o qual poderia automatizar o processo para essa língua, apesar de ainda não ter grande precisão.

De fato, o dado de entrada para nossos operadores não é o grafo CST produzido na etapa de análise, mas um ranque inicial das unidades informativas contidas nele. Esse 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 de um operador é, a partir do ranque inicial, produzir um ranque refinado, de tal forma que as unidades informativas mais relevantes segundo o critério especificado pelo usuário melhorem de posição 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, em número de palavras), são selecionadas tantas sentenças do ranque quanto possível (a partir das sentenças mais bem posicionadas) 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 SAM (Mani, 2001) e, de fato, pode ser facilmente verificada.

Na Figura 5, 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.

No momento, quando algumas unidades apresentam o mesmo numero de relações, elas são ranqueadas na ordem em que são lidas do grafo.

Neste trabalho, consideramos as sentenças como unidades informativas, pois em geral são bem formadas e autocontidas.

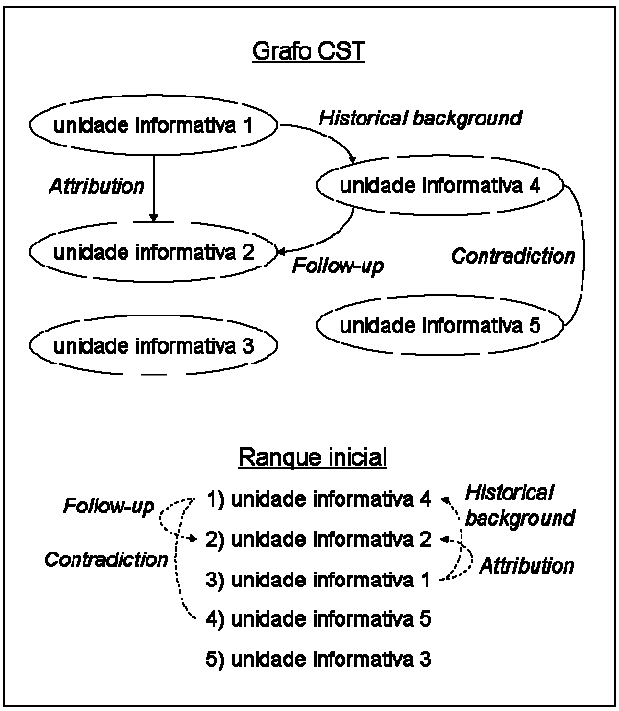

Figura 5. Exemplo de ranque inicial a partir de um grafo CST

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 seguinte:

CONDIÇÃO(S_i, S_j, Direcionalidade, Relação)

Uma dada condição é satisfeita se existem a relação e a direcionalidade (de S_i até S_j: →; o caso oposto: ←; ou nenhuma direcionalidade: —) especificadas entre duas sentenças S_i e S_j, sendo que S_i aparece antes de S_j no texto. As restrições são opcionais, pois representam possíveis requisitos extras para que o operador seja aplicado. Atualmente, só usamos a restrição 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(S_i,S_j): 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(S_i,S_j): trocam-se as posições das sentenças i e j no ranque;
- ELIMINA(S_j): elimina-se a sentença j do ranque.

Para o presente trabalho, definimos e formalizamos 5 operadores que representam possíveis estratégias de seleção de conteúdo. São elas: 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. Chamamos 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 6, mostra-se o operador para apresentação de informação contextual.

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 6. Operador de apresentação de informação de contexto

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.

A aplicação deste operador ao ranque inicial da Figura 5 irá produzir o ranque refinado da Figura 7, 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.

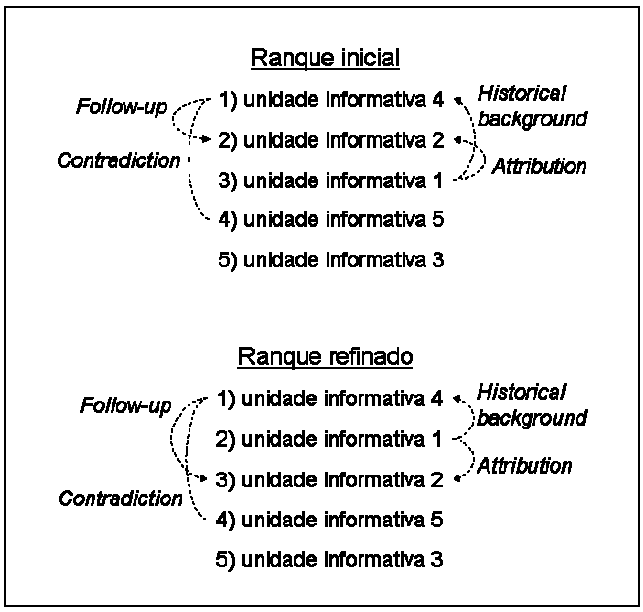

Figura 7. Ranque refinado

Na Figura 8 a seguir, é ilustrado um exemplo de sumário multidocumento usando o operador de apresentação de informação contextual. Como podemos ver na figura, a segunda e a terceira sentença (grifadas) contêm informação contextual e histórica, respectivamente, em relação a primeira sentença.

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 8. Exemplo de sumário produzido pelo operador de apresentação de informação contextual

De fato, pode-se notar que a segunda sentença é redundante em relação a 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 neste artigo.

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 9 mostra o operador correspondente. A forma de interpretação deste operador é a mesma do operador anterior. É interessante notar que, como a direcionalidade não importa neste caso, repetem-se regras 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, ←, *Historical background*)
$\Rightarrow$ SOBE(S_i, S_j)
CONDIÇÃO(S_i, S_j, →, *Historical background*)
$\Rightarrow$ SOBE(S_i, S_j)
CONDIÇÃO(S_i, S_j, ←, *Follow-up*)
$\Rightarrow$ SOBE(S_i, S_j)
CONDIÇÃO(S_i, S_j, →, *Follow-up*)
$\Rightarrow$ SOBE(S_i, S_j)

Figura 9. Operador de apresentação de eventos que evoluem no tempo

A Figura 10 mostra um sumário produzido pelo uso desse operador. Pode-se notar que a segunda sentença (grifada) contém informação sobre um 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 10. Exemplo de sumário produzido pelo operador de apresentação de eventos que evoluem no tempo

A Figura 11 mostra o operador para exibir informações contraditórias, as quais são expressas por meio da relação *Contradiction*, enquanto a Figura 12 mostra o operador para

identificação de fonte/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*.

Nome
Exibição de informações contraditórias

Descrição
Preferência por informações contraditórias

Regra
CONDIÇÃO(S_i, S_j, —, *Contradiction*)
$\Rightarrow$ SOBE(S_i, S_i)

Figura 11. Operador de exibição de informações contraditórias

Nome
Identificação de fonte/autoria

Descrição
Preferência por informações atribuídas a uma fonte

Regras
CONDIÇÃO(S_i, S_j, ←, *Attribution*),
CONDIÇÃO(S_i, S_j, ←, *Subsumption*)
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)
CONDIÇÃO(S_i, S_j, ←, *Citation*),
CONDIÇÃO(S_i, S_j, ←, *Subsumption*)
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)

Figura 12. Operador de identificação de fonte/autoria

As Figuras 13 e 14 mostram sumários produzidos por esses operadores, com a informação privilegiada grifada. Pode-se notar no sumário da Figura 13 que as duas 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. No sumário da Figura 14, 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 15. Em particular, neste operador, também são definidas algumas restrições em relação ao comprimento 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.

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 13. Exemplo de sumário automático produzido pelo operador de apresentação 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 14. Exemplo de sumário automático produzido pelo operador de identificação de fonte/autoria

Nome
Tratamento de redundâncias

Descrição
Preferência por informações não redundantes

Regras
CONDIÇÃO(S_i, S_j, —, *Identity*)
$\Rightarrow$ ELIMINA(S_j)
CONDIÇÃO(S_i, S_j, —, *Equivalence*), $|S_i| \leq |S_j|$
$\Rightarrow$ ELIMINA(S_j)
CONDIÇÃO(S_i, S_j, —, *Equivalence*), $|S_i| > |S_j|$
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)
CONDIÇÃO(S_i, S_j, ←, *Subsumption*)
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)
CONDIÇÃO(S_i, S_j, →, *Subsumption*)
$\Rightarrow$ ELIMINA(S_j)

Figura 15. Operador de tratamento de redundâncias

É desejável que o operador de tratamento de redundâncias seja aplicado antes de qualquer outro operador (excetuando-se o operador genérico, logicamente, já que ele constrói o

ranque inicial), pois ele evita que conteúdo redundante seja incluído nos sumários. Ele evitaria, por exemplo, a sentença redundante (a segunda sentença) do sumário da Figura 8.

Na Figura 16, mostra-se o algoritmo geral para o procedimento de aplicação de operadores de seleção de conteúdo.

O procedimento tem como entrada o grafo CST e, como saída, o ranque refinado. Inicialmente, a partir do grafo CST, é construído o ranque inicial. Em seguida, lê-se a preferência de sumarização do usuário e, então, seleciona-se o operador correspondente, o qual é aplicado para todo par possível de sentenças no ranque, produzindo o ranque refinado.

Após esse processo, devem-se selecionar as sentenças mais bem ranqueadas que irão compor o sumário final, respeitando-se a taxa de compressão especificada pelo usuário. A etapa de síntese realiza a justaposição das sentenças selecionadas (não impondo nenhuma ordem em específico entre elas), exibindo o sumário final para o usuário.

De acordo com a forma que o método de seleção de conteúdo foi projetado, a partir do ranque inicial e da aplicação opcional do operador de tratamento de redundância, só se permite a aplicação de um dos demais operadores de seleção de conteúdo, a saber, de apresentação de informação de contexto, exibição de informação contraditória, identificação de autoria, e de apresentação de eventos que evoluem com o tempo. Ao permitir a aplicação de mais de um destes operadores, o ranqueamento feito pelo operador anterior pode ser alterado pelo novo operador. De fato, o último operador a ser aplicado vai fazer sua ordenação no ranque prevalecer.

Uma possibilidade para tornar possível ler mais de uma preferência de sumarização do usuário é ordenar as preferências em função de suas prioridades (que podem ser definidas pelo próprio usuário). Conseqüentemente, a aplicação dos operadores selecionados seria na ordem inversa, ou seja, deixando-se para o fim a aplicação dos operadores cujas preferências correspondentes têm maior prioridade, pois seriam essas que iriam prevalecer.

Outra possibilidade para lidar com várias preferências seria compor operadores mistos, considerando conjuntos maiores de relações em cada operador. Logicamente, ainda se teria que priorizar alguma informação, de forma que o operador possa produzir um ranque de informações que supra as expectativas do usuário. Para tal encaminhamento, acredita-se que estudos de caso com usuários sejam desejáveis, o que embasaria e tornaria possível o projeto de operadores mistos.

Nesse ponto, é interessante que se diga que a seleção de relações para a composição dos operadores atuais foi baseada nas bases teóricas da CST e na semântica de cada relação. Teoricamente, é possível compor novos operadores com relações diferentes, que poderiam, inclusive, incorporar outras preferências dos usuários que não são utilizadas nesse trabalho. Nesse trabalho, lidamos apenas com as preferências mais diretas e facilmente mapeadas para as relações da CST. Há relações não utilizadas nos operadores e que poderiam eventualmente produzir novos operadores ou serem incorporadas em alguns dos existentes.

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 genérico/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 16. Algoritmo de aplicação dos operadores de seleção de conteúdo

É interessante notar que nenhum dos operadores atuais lida com a relação *Overlap*. Esta relação indica que duas unidades informativas possuem informação em comum, além de informações particulares a cada uma. Veja, por exemplo, as duas sentenças abaixo de textos diferentes:

> Brasil e Finlândia se enfrentarão novamente neste sábado, às 12h30 (horário de Brasília), com transmissão ao vivo do canal de TV a cabo SporTV.
>
> Os dois times voltam a se enfrentar às 12h30 deste sábado, no mesmo ginásio, que normalmente é utilizado para competições de hóquei no gelo.

Há uma relação de *Overlap* entre elas, pois têm informação em comum (grifada), mas também têm informações extras: a primeira sentença informa por onde será feita a transmissão, enquanto a segunda dá mais detalhes do ginásio onde ocorrerá o evento. Há elementos redundantes que devem ser tratados no processo de seleção de conteúdo. Para tratar a redundância nesse caso, não se pode excluir uma das sentenças, como fizemos com as relações *Identity*, *Equivalence* e *Subsumption*, pois se estaria excluindo informações novas e que poderiam ser importantes. O que se precisa, de fato, é fundir as sentenças que apresentam relações *Overlap*, produzindo-se uma única sentença como a abaixo (dentre várias possibilidades):

> Com transmissão ao vivo do canal de TV a cabo SporTV, Brasil e Finlândia voltam a se enfrentar às 12h30 deste sábado (horário de Brasília), no mesmo ginásio, que normalmente é utilizado para competições de hóquei no gelo.

Para a língua portuguesa, poderia ser utilizado o sistema de fusão de Seno e Nunes (2009). Tal opção ainda não foi incorporada no estágio atual do método de seleção de conteúdo, pois implicaria em outros fatores a serem considerados, por exemplo, a gramaticalidade e o foco das sentenças fundidas, e a questão de se deixar de se produzir extratos (sumários formados pela justaposição de segmentos inalterados dos textos-fonte, os quais temos explorado aqui) para se produzir *abstracts* (em que há operações de reescrita textual).

A seguir apresentamos a avaliação das estratégias de seleção de conteúdo propostas.

4. Experimentos e Resultados

Para avaliar nossos operadores de seleção de conteúdo, construímos um protótipo de um sumarizador multidocumento, ao qual chamamos CSTSumm (*CST SUMMarizer*). Esse protótipo aplica o algoritmo da Figura 16 e realiza a síntese do sumário como explicado anteriormente. Os operadores propostos 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 conteúdo desse arquivo XML é carregado pelo protótipo no início de sua execução.

Para nossos experimentos, usamos um córpus composto de 50 coleções de textos jornalísticos escritos em Português Brasileiro (Aleixo e Pardo, 2008b), sendo que cada coleção tem 2 ou 3 textos sobre o mesmo tópico, e cada texto tem em média 20 sentenças. Esse córpus, chamado CSTNews, também contém a análise CST de cada coleção de textos e o sumário humano correspondente (genérico, com as informações mais importantes dos textos, sem preferências particulares), cujo tamanho corresponde a 30% do tamanho do maior texto da coleção (em número de palavras). Os textos do córpus foram coletados de vários jornais on-line brasileiros, como Folha de São Paulo, Estadão e Jornal do Brasil. O córpus foi analisado segundo a CST por 4 lingüistas computacionais previamente treinados nesse tipo de anotação, obtendo resultados de concordância satisfatórios.

A Figura 17 mostra a freqüência de ocorrência das relações CST no córpus. Pode-se notar que algumas relações ocorrem pouco (por exemplo, *Modality*, *Translation* e *Summary*), uma nunca ocorre (*Citation*) e outras ocorrem muito (por exemplo, *Elaboration* e *Overlap*).

Os sumários automáticos foram gerados para todos os operadores propostos neste trabalho, considerando a mesma taxa de compressão dos sumários humanos. Com exceção do operador genérico, o operador de tratamento de redundâncias foi aplicado antes dos demais operadores serem aplicados.

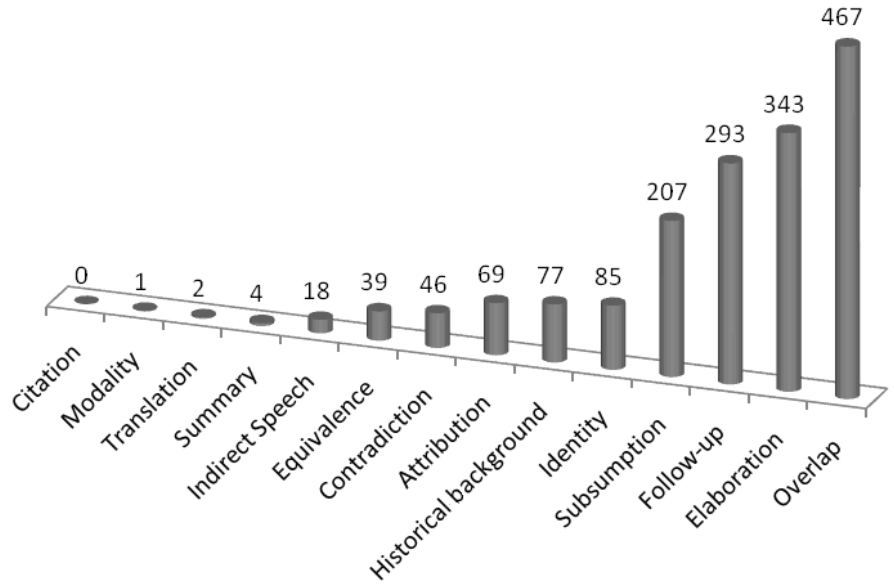

Figura 17. Relações CST no córpus

Neste trabalho consideramos dois métodos de avaliação: o automático, que é usado para medir a informatividade dos sumários, e o humano, que é usado para avaliar a coerência do sumário.

Para a avaliação automática, foi usada a medida ROUGE (Lin e Hovy, 2003), que é uma medida automática que computa o quão similar um sumário automático é em relação ao sumário humano correspondente. Basicamente, a similaridade é computada em função do número de n-gramas em comum entre os sumários, produzindo-se valores de precisão, cobertura e medida-f, tradicionais na área de pesquisa em questão. A precisão indica o quanto do sumário automático é, de fato, relevante; a cobertura indica o quanto do sumário humano é reproduzido no automático; a medida-f é uma medida única de desempenho, combinando precisão e cobertura. Apesar da comparação de n-gramas parecer simples demais para ser confiável, os autores da medida demonstraram que ela é tão boa quanto humanos em ranquear sumários em função de sua informatividade. De fato, tal medida foi amplamente aceita na comunidade de pesquisa e é usada até mesmo nas avaliações em larga escala organizadas anualmente (veja, por exemplo, as TACs – *Text Analysis Conferences* – principais competições mundiais na área de sumarização). Neste trabalho, utilizamos a ROUGE-1, ou seja, fazemos somente a comparação de unigramas, que, como os autores da medida mostraram, já basta para que se tenham resultados confiáveis.

Na avaliação humana, por enquanto, avaliamos somente o aspecto da redundância. O número de sentenças redundantes foi calculado para uma pequena amostra de sumários produzidos pelos diferentes operadores. O fato de se utilizar apenas uma amostra advém do custo e do tempo necessários para a avaliação humana.

Os resultados foram comparados com os resultados obtidos pelo único sumarizador multidocumento conhecido para a língua portuguesa, o GistSumm (Pardo et al., 2003, 2005). Este sumarizador concatena todos os textos de uma mesma coleção em um único arquivo e, posteriormente, sumariza-o utilizando um método de sumarização baseado nas palavras mais freqüentes. Esse método é muito simples, mas ainda assim robusto, correspondendo, portanto, a um ótimo *baseline*.

Na Tabela 3 são mostrados os resultados da avaliação para todos os operadores e para o GistSumm. Note que o operador de tratamento de redundância também foi avaliado de forma isolada, sem ser combinado com os demais.

Tabela 3. Resultados das avaliações

	Cobertura	Precisão	Medida-f	Sentenças redundantes
Informação Principal (operador genérico)	0.57218	0.52359	0.54384	3
Tratamento de Redundância	0.55137	0.54539	0.54299	0
Exibição de Informações Contraditórias	0.57108	0.51974	0.54114	1
Identificação de Autoria	0.56518	0.52368	0.53994	2
Apresentação de Eventos que Evoluem no Tempo	0.55136	0.49869	0.52110	3
Apresentação de Informação de Contexto	0.52079	0.48962	0.50171	4
GistSumm	0.66435	0.35997	0.45998	5

As primeiras 3 colunas da tabela reportam os resultados médios da ROUGE (que variam de 0 a 1 e, quanto maiores, melhores). Em geral podemos observar que todos os sumários produzidos pelos operadores têm melhores resultados do que o GistSumm em termos da medida-f. Logicamente, o operador genérico tem a maior medida-f dentre os operadores, já que os sumários de referência são genéricos também. Comparamos os sumários com preferências com os sumários genéricos para poder verificar seu nível de informatividade, independentemente do fato de terem priorizado outras informações. Em termos de precisão, o operador de tratamento de redundâncias é o melhor, pois elimina informações repetidas e pode, assim, incluir outras informações relevantes no sumário. É interessante notar também a alta cobertura do GistSumm e sua baixíssima precisão.

É importante notar que muitos operadores produziram resultados próximos do operador genérico e do de tratamento de redundância. Isso se deve ao fato de que alguns operadores têm poucas relações correspondentes disponíveis no ranque inicial, não alterando significativamente o sumário produzido. Por exemplo, há poucas relações *Contradiction* no córpus, de forma que há grandes chances de o sumário automático não ser muito alterado pelo operador de exibição de informações contraditórias.

O teste estatístico anova mostrou que os resultados da ROUGE obtidos são significantes com 95% de confiança.

A última coluna da tabela exibe o número de sentenças redundantes encontradas nos sumários. Pode-se observar que todos os operadores geraram sumários menos redundantes e, portanto, mais coerentes do que os sumários gerados pelo GistSumm. Como esperado, o operador de tratamento de redundâncias produziu sumários sem redundância alguma. Por outro lado, mesmo com a aplicação prévia do operador de tratamento de redundância, os operadores de preferência produziram redundâncias. Essas redundâncias são explicadas principalmente pela presença das relações *Contradiction* e *Overlap*: a primeira sempre traz alguma redundância consigo, enquanto a segunda não foi devidamente tratada neste trabalho (via fusão das sentenças envolvidas, por exemplo). Outra possibilidade é que existam sentenças nos textos que não tenham sido anotadas com relações CST, mas que de fato tenham relação entre si e contenham redundância.

A seguir, fazemos algumas considerações finais.

5. Considerações Finais

Neste trabalho, foram definidos, formalizados e avaliados um conjunto de operadores de seleção de conteúdo para SAM com base na CST. Mostramos que o uso da CST permite explorar o conhecimento entre vários textos que versam sobre um mesmo assunto, o que ajuda na seleção de conteúdo, melhorando a informatividade e coerência nos sumários finais.

Trabalhos futuros incluem a elaboração de novas estratégias de seleção de conteúdo com base na CST, incluindo possivelmente a criação de novos operadores de seleção de conteúdo. A avaliação do impacto da preferência do usuário, em particular, merece uma atenção maior. Neste artigo, tratou-se apenas da questão da informatividade, mas certamente alguma avaliação humana deverá ser conduzida, de tal forma que se possa mensurar a satisfação do usuário frente aos sumários gerados de acordo com suas preferências. Alternativamente,

sumários humanos com preferências específicas podem ser produzidos para serem considerados sumários de referência para avaliação automática dos sumários com preferências.

Acreditamos que a CST pode auxiliar em outros processos da sumarização, como ordenação das sentenças do sumário e resolução das correferências. A ordenação das sentenças, em especial, pode ter um grande efeito na coerência final do sumário, e deve ser foco de próximas pesquisas. Além disso, cremos também que a taxa de compressão utilizada interfere nos resultados obtidos, desde que, quanto maior a taxa, menos informação o sumário pode conter. Tal influência deve ser investigada em trabalhos futuros.

Por fim, é interessante notar que, em princípio, o trabalho apresentado é independente de língua e de gênero e domínio textual, já que a CST e, portanto, os operadores derivados dela são independentes de língua e genéricos o suficiente para serem aplicados a outros tipos de textos.

6. Agradecimentos

Os autores agradecem à FAPESP e ao CNPq pelo suporte a este trabalho.

7. Referências

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.

Aleixo, P. and Pardo, T.A.S. (2008a). 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.

Aleixo, P. and Pardo, T.A.S. (2008b). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 326.

Jorge, M.L.C and Pardo, T.A.S. (2009). Content Selection Operators for Multidocument Summarization based on Cross-document Structure Theory. In the *Brazilian Symposium in Information and Human Language Technology.* São Carlos, Brazil.

Jorge, M.L.C. and Pardo, T.A.S. (2010). Formalizing CST-based Content Selection Operations. In the *Proceedings of the International Conference on Computational Processing of Portuguese Language* - PROPOR. April 27-30, Porto Alegre/RS, Brazil.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In the *Proceedings of 2003 Language Technology Conference*. Edmonton, Canada.

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

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

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

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.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal.

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.

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. (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*.

Seno, E.R.M. e Nunes, M.G.V. (2009). *Fusão Automática de Sentenças Similares em Português*. *Linguamática*, Vol. 1, pp. 71-87.

Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community*. Ph.D. Thesis. Department of Computer Science, University of Maryland.

Trigg, R. and Weiser, M. (1987). TEXTNET: A network-based approach to text handling. *ACM Transactions on Office Information Systems*, Vol. 4, N. 1, pp. 1-23.

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

Zhang, Z.; Otterbacher, J.C.; Radev, D.R. (2003). Learning Cross-document Structural Relationships Using Boosting. In the *Proceedings of Conference on Information and Knowledge Management*, pp. 124-130.