Sumarização Automática Multidocumento: Seleção de Conteúdo com Base no Modelo CST (*Cross-document Structure Theory*)

Maria Lucia del Rosario Castro Jorge, **Thiago Alexandre Salgueiro Pardo**

Núcleo Interinstitucional de Linguística 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

{mluciacj,taspardo}@icmc.usp.br

Mestrado defendido em Abril de 2010 (com apoio do CNPq e da FAPESP)

Resumo. *Este artigo apresenta a definição, a formalização e a 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. Nossos experimentos foram feitos usando um córpus jornalístico de textos escritos em português brasileiro e mostram que o uso da CST melhora a informatividade dos sumários. A abordagem mostra-se nova para a sumarização multidocumento em língua portuguesa por ser a primeira abordagem que explora conhecimento linguístico para esta tarefa, e, ao mesmo tempo, avança o estado da arte ao modelar e explorar de maneira diferenciada o conhecimento fornecido pela CST.*

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 e narradas de diferentes formas e perspectivas. Com o surgimento de meios de interação mais sofisticados e atualmente quase onipresentes, como a web 2.0, por exemplo, esse cenário torna-se ainda mais complexo. Consequentemente, o processamento da informação tem se tornado uma tarefa de difícil execução para o humano e para a máquina. 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). Na SAM, além de se recuperar a informação mais relevante, deseja-se ser capaz de se priorizar informação especificada pelo usuário (por exemplo, informação contextual sobre um determinado fato/evento), visualizar a evolução de alguns fatos/eventos no tempo, lidar com diferentes estilos de escrita, ordenar eventos e fatos cronologicamente, e manter a coerência e coesão do sumário, dentre vários outros desafios.

Mani e Maybury (1999) sugerem que a sumarização envolve 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 na forma de sumário, expresso em uma língua natural. Sistemas de SAM que adotam tal abordagem, privilegiando a manipulação e o uso de conhecimento lingüístico sofisticado, são ditos pertencerem à abordagem profunda. Neste trabalho, foca-se especificamente no processo mais importante da etapa de transformação: a seleção de conteúdo dos textos de origem para compor o sumário correspondente. 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, neste trabalho são exploradas estratégias de seleção de 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 hipótese deste trabalho é que o uso de conhecimento semântico-discursivo fornecido pela CST pode produzir sumários multidocumento melhores.

As estratégias propostas são formalizadas na forma de operadores de seleção de conteúdo, os quais contêm regras de manipulação da informação textual para satisfação dos interesses dos usuários. Nossos experimentos foram feitos usando um córpus jornalístico de textos em português brasileiro, e mostram que o uso da CST melhora a informatividade dos sumários, confirmando assim nossa hipótese.

Entre as principais contribuições deste trabalho tem-se: a primeira investigação de SAM profunda para o português do Brasil; os melhores resultados obtidos até o momento para o português, superando, inclusive, um dos melhores sistemas de sumarização disponíveis na área; formalização de estratégias de seleção de conteúdo, mapeando-se preferências de usuários em características do modelo CST; validação e rico refinamento do modelo CST, tornando-o mais consistente; construção de recursos e ferramentas inéditos para o português do Brasil, como córpus e editores especializados.

A seguir, na Seção 2, introduz-se o modelo CST e apresentam-se os principais 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, a Seção 5 faz algumas considerações finais.

2. Trabalhos Relacionados

A CST foi originalmente proposta com um conjunto de 24 relações para explicitar o relacionamento entre as partes de vários textos, por exemplo, relações de equivalência, contradição, subsunção, elaboração, etc. Essas relações representam os fenômenos multidocumento usuais de redundância, complementariedade e contradição entre informações, explorados com detalhes por Castro Jorge (2010). Como exemplos de aplicação destas relações a um grupo de textos, mostram-se, na Figura 1, alguns trechos de textos (de fontes diferentes) relacionados.

Sentença 1	Sentença 2	Relação CST
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.	*Subsumtion* (←) Sentença 2 engloba Sentença 1
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.	*Contradiction* (—) Sentenças 1 e 2 apresentam uma contradição (em negrito)

Figura 1. Exemplos de relações CST

É importante notar que algumas relações têm direcionalidade, enquanto outras não. Nos exemplos, a relação *subsumption* é da Sentença 2 para a Sentença 1 (portanto, da sentença da direita para a da esquerda, representada pela seta ←), pois a 2 engloba a 1; por outro lado, a relação *contradiction* não tem direção (representada por —), pois as sentenças contradizem uma a outra igualmente.

Como parte deste trabalho, o conjunto de relações CST foi refinado para se obter uma melhor formalização e reduzir a ambiguidade de muitas relações do modelo, resultando em um conjunto de 14 relações representativas para a língua portuguesa.

Algumas pesquisas têm utilizado CST para fins de SAM, incluindo o próprio autor do modelo, que propôs uma metodologia de sumarização de 4 etapas, a saber: agrupamento de textos de conteúdo similar, estruturação interna dos textos (via análise sintática, por exemplo), estabelecimento de relações CST (gerando, assim, um grafo em que os nós representam as sentenças e as arestas as relações CST), e seleção de sentenças para compor o sumário. Essa seleção, em particular, constrói um ranque de sentenças em função da relevância de cada uma e, em seguida, coleta as primeiras colocadas do ranque para formar o sumário. O número de sentenças coletadas depende da taxa de compressão informada pelo usuário, i.e., o tamanho do sumário em função do tamanho dos textos (com medição em número de palavras).

Dentre os principais trabalhos na área, destaca-se o de McKeown e Radev (1995), que, utilizando um sistema de extração de informação textual, modelaram o conteúdo textual em *templates*

(estruturas pré-definidas com campos preenchidos com informações textuais) e, em função das relações (no estilo da CST) entre *templates*, produziram os *templates* finais para subsidiar a construção do sumário correspondente. Zhang et al. (2002), por sua vez, propuseram a alteração, por meio do uso de relações CST, de ranques produzidos por sumarizadores superficiais (ou seja, que utilizam pouco ou nenhum conhecimento linguístico), mostrando a melhora significativa da qualidade dos sumários. Otterbacher et al. (2002) investigaram como o uso de relações CST ajuda a melhorar a coesão em sumários multidocumento, fazendo com que sentenças relacionadas pela CST apareçam próximas no sumário. Afantenos et al. (2004), com base na CST, propuseram uma nova classificação de relações entre textos. Os autores dividiram as relações em duas categorias: sincrônicas (um fato/evento sendo narrado por diferentes fontes em um determinado momento) e diacrônicas (fatos/eventos em evolução no tempo). Em seguida, propuseram uma metodologia de sumarização com base em *templates* pré-definidos e ontologias. Apesar de muito interessante, esse trabalho foi apenas teórico.

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

3. Definição e Formalização de Operadores de Seleção de Conteúdo

Formalmente, definimos um operador de seleção de conteúdo como um artefato computacional que processa uma representação de conteúdo previamente fornecida e produz uma versão mais curta contendo as informações mais relevantes segundo os critérios especificados. Em particular, neste trabalho, a representação de conteúdo consiste no conjunto de textos representados segundo o modelo CST. Portanto, os operadores são aplicados após os textos-fonte terem sido analisados segundo a CST (na etapa de análise). Atualmente, tal análise deve ser feita manualmente para a língua portuguesa, já que o primeiro analisador automático ainda se encontra em desenvolvimento (Maziero et al., 2010).

O dado de entrada dos operadores é um ranque inicial das unidades informativas dos textos (que, neste trabalho, são sentenças). Esse ranque é obtido com base no grafo construído a partir do relacionamento CST entre as unidades do texto (daqui em diante, esse grafo será referenciado por grafo CST). 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 dos operadores é, a partir do ranque inicial e da preferência do usuário, produzir um ranque refinado, de tal forma que as unidades informativas mais relevantes segundo o critério especificado pelo usuário melhorem de posição no ranque e, portanto, ganhem preferência para estar o sumário. Por fim, dada uma taxa de compressão, são selecionadas tantas sentenças do ranque refinado quanto possível para que a taxa seja respeitada.

No ranque inicial, a relevância das unidades informativas depende do número de relações CST que elas apresentam, isto é, unidades com mais relações CST são consideradas mais relevantes. Na Figura 2, mostra-se um exemplo hipotético de um grafo CST e o ranque inicial formado a partir deste.

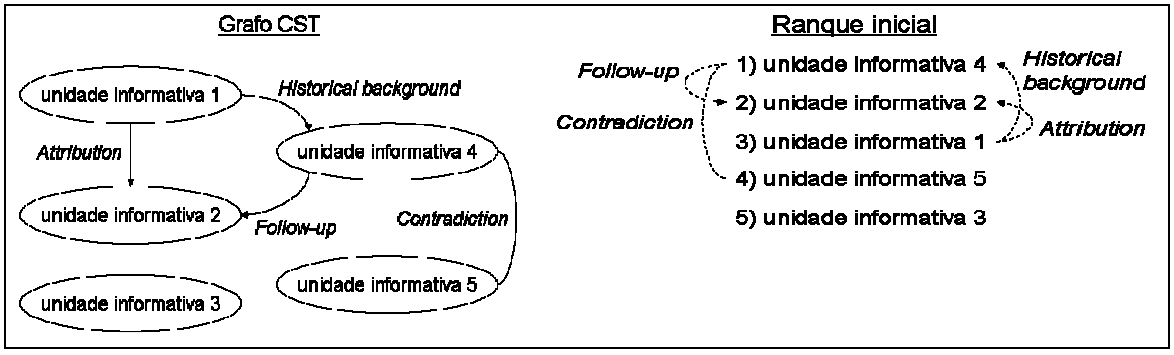

Figura 2. Exemplo de ranque inicial a partir de um grafo CST hipotético

Os operadores de seleção de conteúdo com base na CST estão definidos na forma 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 $\Rightarrow$ AÇÕES. Cada condição tem o formato CONDIÇÃO(S_i, S_j, Direcionalidade, Relação) e uma dada condição é

satisfeita se existem a relação e a direcionalidade (de S_i até S_j: $\rightarrow$; o caso oposto: $\leftarrow$; ou nenhuma direcionalidade: —) especificadas entre duas sentenças S_i e S_j. As restrições são opcionais, pois representam possíveis requisitos extras para que o operador seja aplicado.

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 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 3, mostra-se como exemplo o operador para apresentação de informação contextual. Nesse operador, procuram-se por pares de sentenças (ao longo do ranque) que apresentem relações CST do tipo *historical background* e *elaboration*, já que essas relações são as que fornecem informação contextual.

Nome	Operador para apresentação de informação contextual
Descrição	Preferência por informações históricas e complementares
Regras	CONDIÇÃO(S_i, S_j, $\leftarrow$, *Elaboration*) $\Rightarrow$ SOBE(S_i, S_j) CONDIÇÃO(S_i, S_i, $\leftarrow$, *Historical background*) $\Rightarrow$ SOBE(S_i, S_i)

Figura 3. Operador de apresentação de informação de contexto

A aplicação deste operador no ranque inicial da Figura 2 irá produzir o ranque refinado em que a unidade 1 sobe de posição no ranque, sendo realocada logo após a unidade 4 e acima da unidade 2, pois apresenta a relação de *historical background* com a unidade 4.

Após o refinamento do ranque, podem existir sentenças redundantes, por isso, para resolver o problema, faz-se necessário aplicar o operador de tratamento de redundância, o qual elimina sentenças com conteúdo repetido do ranque (indicado por relações de equivalência e subsunção, por exemplo). O ranque refinado, com base na taxa de compressão especificada, indicará as sentenças para compor o sumário final. No texto do exemplo, supondo que 2 sentenças sejam permitidas no sumário, as unidades informativas 4 e 1 seriam escolhidas.

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 procedimento explicado na seção anterior.

Para nossos experimentos, usamos um córpus composto de 50 coleções de textos jornalísticos escritos em Português Brasileiro (Aleixo e Pardo, 2008), sendo que cada coleção tem 2 ou 3 textos sobre o mesmo tópico. Esse córpus contém a análise CST (realizada manualmente) de cada coleção de textos e o resumo humano (genérico, sem preferências) correspondente, cujo tamanho corresponde a 30% do tamanho do maior texto da coleção (em número de palavras).

Os sumários automáticos foram produzidos considerando a mesma taxa de compressão dos sumários humanos. Neste trabalho, consideramos dois métodos de avaliação: o automático, que é usado para medir a informatividade dos sumários genéricos**,** e o humano, que é usado para avaliar fatores como a informatividade, coerência, coesão, gramaticalidade e redundância dos sumários com preferências do usuário.

Para a avaliação automática, foi usada a medida ROUGE (Lin e Hovy, 2003). Esta medida produz valores de precisão, cobertura e medida-f (que combina precisão e cobertura), tradicionais na área de pesquisa em questão. Os resultados da ROUGE foram comparados com os resultados obtidos pelo único sumarizador multidocumento superficial conhecido para a língua portuguesa, o GistSumm (Pardo, 2005), e o sumarizador MEAD (Radev et al., 2001), que é um sumarizador multilíngue superficial bastante utilizado na área. Além disso, também estendemos os experimentos aplicando a metodologia proposta por Zhang et al. (2002) aos sumarizadores superficiais anteriores. Assim, as sentenças ranqueadas por esses sumarizadores foram re-ranqueados por meio das relações CST.

Na Tabela 1, são mostrados os resultados da avaliação. Em geral podemos observar que todos os sumários produzidos pelos operadores têm melhores resultados do que o GistSumm e o MEAD em termos da medida-f. Também podemos observar que os desempenhos dos sumarizadores GistSumm e MEAD melhoram com o uso das relações CST, conforme predito por Zhang et al. (2002).

Tabela 1. Resultados das avaliações

	PRECISÃO	COBERTURA	MEDIDA-F
Informação Principal	0.57218	0.52359	0.54384
Tratamento de Redundância	0.55137	0.54539	0.54299
Exibição de Informações Contraditórias	0.57108	0.51974	0.54114
Identificação de Autoria	0.56518	0.52368	0.53994
Apresentação de Eventos que Evoluem no Tempo	0.55136	0.49869	0.52110
Apresentação de Informação de Contexto	0.52079	0.48962	0.50171
GistSumm	0.66435	0.35997	0.45998
GistSumm enriquecido com CST	0.49450	0.50890	0.49940
MEAD	0.52420	0.46020	0.48690
MEAD enriquecido com CST	0.55990	0.49890	0.52300

Já os resultados da avaliação humana mostraram um bom desempenho nos critérios avaliados, sendo que a informatividade é um dos critérios em que melhor desempenho se obteve. Também foi observada uma baixa presença de sentenças redundantes nos sumários finais. A seguir, na Figura 4, mostra-se um exemplo de sumário multidocumento genérico e sem redundância produzido automaticamente pelo CSTSumm a partir de 3 textos.

O Brasil arrasou a Finlândia no primeiro confronto entre as seleções, nesta sexta-feira, na cidade de Tampere, pela Liga Mundial de vôlei 2006, por 3 sets a 0, com parciais de 25/17, 25/22 e 25/21. O resultado de hoje deixou o Brasil perto de conquistar a única vaga do Grupo B da Liga Mundial, que classifica o melhor de cada uma das quatro chaves, a Rússia (país-sede) e mais um time convidado pela Federação Internacional de Vôlei, para a fase final, de 23 a 27 de agosto, em Moscou (Rússia). Brasil e Finlândia se enfrentarão novamente neste sábado, às 12h30 (horário de Brasília), com transmissão ao vivo do canal de TV a cabo SporTV.

Figura 4. Exemplo de sumário multidocumento genérico produzido automaticamente

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.

Por enquanto, além do operador de tratamento de redundância, só permitimos a aplicação de um operador de seleção de conteúdo em um dado momento. Como trabalho futuro, pretende-se estudar a possibilidade do uso de mais operadores por vez. Outros trabalhos futuros incluem a elaboração de novas estratégias de seleção de conteúdo com base na CST.

Na medida do possível, todos os resultados deste trabalho (tanto teóricos como práticos) estão públicos e disponíveis na página do projeto maior do qual este trabalho fez parte: o projeto "sucinto" (www.icmc.usp.br/~taspardo/sucinto).

Ao longo do mestrado foram produzidos 5 artigos. É importante destacar que o artigo publicado no evento STIL 2009 (principal evento de Processamento de Linguagem Natural no Brasil, promovido oficialmente pela SBC) ganhou o prêmio de segundo melhor trabalho do evento. Outro

artigo importante foi o publicado na revista Linguamática, uma das principais revistas da área na comunidade ibero-americana. Destacam-se também os artigos publicados no PROPOR e no workshop da ACL, conferências internacionais importantes na área. A seguir são listados os artigos publicados (também disponíveis na página do projeto citado anteriormente):

- Castro Jorge, M.L.R. e 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. Uppsala/Suécia.
- Maziero, E.G.; Castro Jorge, M.L.R.; 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.
- Jorge, M.L.C. e Pardo, T.A.S. (2010). Formalizing CST-based Content Selection Operations. In the *Proceedings of the 9th International Conference on Computational Processing of Portuguese Language* - PROPOR (Lecture Notes in Artificial Intelligence 6001), pp. 25-29. Porto Alegre/RS, Brasil.
- Castro Jorge, M.L.R. e Pardo, T.A.S. (2010). Estratégias de Seleção de Conteúdo com Base na CST (Cross-document Structure Theory) para Sumarização Automática Multidocumento. *LinguaMÁTICA*, Vol. 2, N. 1, pp. 95-109.
- Castro Jorge, M.L.R. e Pardo, T.A.S. (2009). Content Selection Operators for Multidocument Summarization based on Cross-document Structure Theory. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL, pp. 1-8. São Carlos/SP, Brasil.

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. (2008). *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.

Castro Jorge, M.L.R. (2010). *Sumarização automática multidocumento: seleção de conteúdo com base no Modelo CST (Cross-document Structure Theory)*. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Abril, 86p.

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.

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.

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

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

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*.

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

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