CSTTool: um *parser* multidocumento automático para o Português do Brasil

Priscila Aleixo, 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
{paleixo, taspardo}@icmc.usp.br

Resumo. Descreve-se, neste artigo, a criação de um *parser* discursivo multidocumento automático para textos escritos em português brasileiro, seguindo-se a teoria discursiva *Cross-document Structure Theory* (CST). A teoria prevê a identificação de relações entre partes de textos que versam sobre o mesmo assunto. Tal conhecimento é de grande valia para aplicações multidocumento de processamento de língua natural, por exemplo, perguntas e respostas e sumarização automática. O *parser* descrito, chamado CSTTool, encontra-se parcialmente desenvolvido e, até o momento, representa o estado da arte para a língua portuguesa.

Palavras-Chave: Análise Discursiva Multidocumento Automática, CST, Processamento de Língua Natural.

Nível do estudante: MSc.
Data prevista de conclusão: Fevereiro de 2009.

1. Introdução

Com o grande acúmulo de informação que os meios eletrônicos hoje proporcionam, existem muitas fontes que relatam o mesmo assunto com as mesmas ou diferentes perspectivas. Um exemplo disso são os informativos ou jornais on-line, onde a notícia é relatada no momento em que acontece, produzindo muitos documentos redundantes ou um mesmo relato em vários documentos diferentes. Além desses fatores, os novos fatos que ocorrem no decorrer do tempo influenciam para que a redundância ocorra.

O leitor neste contexto precisa muitas vezes buscar vários documentos e organizar o conteúdo para identificar a informação desejada. Tal processo não é trivial. Faz-se necessário o conhecimento de como as informações presentes em vários textos de assuntos correlatos se relacionam. Por exemplo, é importante saber que informações se complementam, quais são conflitantes, quais fatos antecedem outros, causas e suas conseqüências, etc. Tal nível de análise encontra-se principalmente no que se convencionou chamar de "discurso", sendo que as relações entre as informações são chamadas de relações discursivas. Em PLN (Processamento de Língua Natural), foi proposta a teoria discursiva multidocumento *Cross-document Structure Theory* (CST)

[1] para lidar com essas questões. A CST é uma das poucas teorias multidocumento existentes.

A CST tem sido usada para várias aplicações de PLN. Por exemplo, em sumarização automática [19, 20, 21] multidocumento, a CST é utilizada para produzir sumários melhores, dado que se verificou que sumários cujas sentenças (provenientes de diferentes documentos) apresentam relações CST entre si são mais informativos. Similarmente ao que ocorre na sumarização, o conhecimento das relações CST pode ajudar sistemas automáticos de perguntas e respostas [22] a produzir respostas melhores e mais direcionadas para diferentes perfis de leitor.

Apesar da grande utilidade na área, sistemas de análise discursiva automática, também chamados de *parsers* discursivos, são de difícil desenvolvimento. Em particular, há poucos *parsers* multidocumento disponíveis. Sabe-se de apenas uma abordagem conhecida para a língua inglesa desenvolvida por [2,3]. Neste artigo, descrevemos o desenvolvimento do primeiro *parser* discursivo multidocumento para o português do Brasil com base na teoria CST. O *parser*, chamado CSTTool, tem sido desenvolvido segundo uma metodologia baseada em córpus, isto é, um conjunto de textos anotados segundo a teoria CST. Os resultados obtidos até o momento representam o estado da arte para o português.

A seguir, na Seção 2, a CST e os trabalhos relacionados são brevemente introduzidos. Na Seção 3, o desenvolvimento do *parser* CSTTool é relatado. Na Seção 4, as avaliações realizadas são descritas. Por fim, na Seção 5, algumas considerações finais são feitas.

2. A Teoria CST e Trabalhos Relacionados

A CST (*Cross-document Structure Theory*), proposta em [1], surgiu frente à necessidade da identificação de relações entre vários textos, estruturando o discurso de forma a conectar sentenças provenientes de diferentes documentos e estabelecendo um ou mais tipos de relações entre elas. É uma das poucas teorias discursivas multidocumento difundidas. Ela se baseia na amplamente conhecida RST (*Rhetorical Structure Theory*) [18], uma das teorias discursivas monodocumento mais utilizada em PLN.

As relações da CST podem acontecer entre palavras, sintagmas, orações, sentenças, parágrafos ou blocos de texto ainda maiores. Apesar de orações e sentenças serem tradicionalmente consideradas as unidades discursivas por excelência, tarefas particulares podem exigir um relacionamento entre unidades menores. Em [1], afirma-se que as relações CST não são mutuamente exclusivas, podendo um mesmo par de segmentos textuais ter mais de uma relação entre si. É interessante notar que, na análise CST, nem todas as sentenças dos textos envolvidos se relacionam. De fato, apenas uma parcela delas apresenta relações entre si.

Após um refinamento das 24 relações propostas inicialmente [1], 18 delas foram mantidas [10], e seguem: *Identity, Equivalence (Paraphrase), Translation, Subsumption, Contradiction, Historical Background, Citation, Modality, Attribution, Summary, Follow-up, Indirect Speech, Fulfillment, Elaboration (Refinement), Description, Reader Profile, Change of Perspective e Overlap*.

Na Figura 1, nota-se a multiplicidade de relações CST: as sentenças S1 e S2 podem ser relacionadas pelas relações *Contradiction* e *Attribution*. No primeiro caso, há informações contraditórias: S1 diz que a colisão foi no 26º andar e S2 diz que foi no 25º andar; no segundo caso, a relação *Attribution* se deve ao fato de que a informação contida em S1 e em S2 está sendo atribuída em S1 a uma jornalista, ou seja, a fonte da informação é identificada.

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

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

Figura 1: Exemplo de relações CST

Os trechos de notícia que estão na Figura 2 relatam um acidente aéreo. Tais trechos foram coletados de fontes distintas. Algumas relações CST são facilmente identificadas, por exemplo, a sentença 2 do texto 2 e a sentença 1 do texto 1 estão relacionadas por *Elaboration*, assim como a sentença 3 do texto 2 e a sentença 2 do texto 1. Na primeira sentença de cada relacionamento a notícia é mais detalhada do que na segunda. É interessante ressaltar que algumas relações possuem direcionalidade, por exemplo, é verdade que a sentença 2 do texto 2 e a sentença 1 do texto 1 possuem uma relação de *Elaboration*, porém o contrário não.

Outros exemplos de relações identificadas nestes trechos são *Overlap* e *Follow-up*. A relação *Overlap* acontece entre a sentença 2 do texto 1 e a sentença 1 do texto 2 e entre a sentença 3 do texto 1 e a sentença 3 do texto 2. Já a relação *Follow-up* acontece entre a sentença 3 do texto 2 e a sentença 1 do texto 1. Nota-se também que algumas sentenças não se relacionam e no final da identificação de todas as relações CST têm-se uma estrutura de grafo em que na maioria das vezes é desconexo.

Nos trabalhos de Zhang (veja, [2], [3], [4]), foi proposto o primeiro e único *parse*r discursivo multidocumento automático para o inglês. Segundo Zhang et al., a análise multidocumento é composta de três principais passos: segmentação sentencial, identificação de segmentos relacionados e escolha das relações.

Em [5], um trabalho paralelo ao de Zhang et al., analisaram-se algumas medidas de similaridade lexical para identificação de segmentos relacionados: medida do cosseno [6], *Word Overlap*, *Longest Common Subsequence* e BLEU [7]. Os autores concluíram que a *Word Overlap* foi a que melhor detectou sentenças relacionadas. Tal medida foi utilizada para selecionar pares de segmentos para serem manualmente relacionados via alguma relação CST. Os segmentos relacionados compõem o córpus que os autores chamam de CSTBank, o qual ainda está em desenvolvimento.

Para o desenvolvimento do *parser* discursivo para a língua inglesa, mais especificamente, a detecção de relação entre pares de sentenças, Zhang usou aprendizado de máquina. Alguns atributos considerados são: número de palavras na primeira sentença, número de palavras na segunda sentença, número de palavras em comum, número de palavras em comum após o *stemming*. Com a ajuda da Wordnet [8], os autores também identificaram sinônimos e antônimos entre verbos, substantivos, adjetivos e advérbios para, assim, então, melhorar a seleção de atributos.

Texto 1
(S1) Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. **(S2) Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.** **(S3) A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.**
Texto 2
(S1) 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. **(S2) As vítimas do acidente foram 14 passageiros e três membros da tripulação.** **(S3) Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.**

Figura 2: Trecho de uma mesma notícia de fontes distintas

Segundo Zhang, o algoritmo AdaBoost [9] é o mais eficiente na detecção de sentenças relacionadas, pois além de ser considerado um dos melhores algoritmos de aprendizado, ele também suporta dados multi-rotulados. O algoritmo foi testado para as tarefas de identificação de segmentos relacionados e identificação de relações. Para a primeira tarefa, foi obtida uma precisão de 71,99%, cobertura de 63,73% e F-*mesuare* de 67,56%; para a segunda, a precisão média foi de 85,40%.

Na próxima seção é apresentado o *parser* CSTTool, o primeiro sistema para o português do Brasil baseado na CST.

3. CSTTool: um *parser* discursivo multidocumento para o Português do Brasil

A arquitetura proposta para o *parser* discursivo multidocumento para o português é ilustrada na Figura 3. Como se pode notar, ela segue o processo dos três passos proposto por Zhang et al.

Os dados de entrada são os textos-fonte presentes nas coleções de documentos que versam sobre assuntos relacionados. Inicialmente, a segmentação textual delimita as sentenças dos textos. A seguir, o processo de identificação de segmentos relacionados aplica medidas de similaridade para identificar possíveis pares de segmentos de textos diferentes que estejam relacionados. Por fim, para os pares detectados, relações discursivas da CST são propostas.

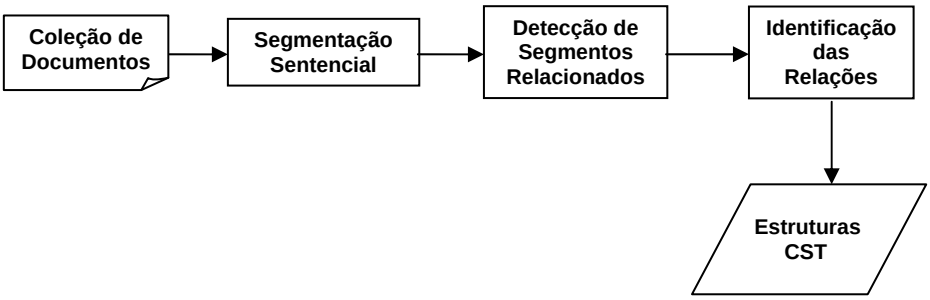

Figura 3: Arquitetura da CSTTool

Como base para o desenvolvimento do *parser*, fez-se necessário a criação de um córpus anotado segundo a CST, o que implicou em também determinar quais relações discursivas utilizar. A seguir, na Subseção 3.1, o processo de refinamento do conjunto de relações e a anotação do córpus são relatados. Na Seção 3.2, as etapas do *parser* multidocumento já desenvolvidas são descritas.

3.1. Refinamento das relações CST e anotação de córpus

Após o estudo das relações CST, foi concluído que, para este trabalho, algumas relações definidas por [10] seriam redundantes e, portanto, foram excluídas deste trabalho. Das 18 relações definidas por Zhang, 14 delas foram mantidas na anotação do córpus. As relações excluídas foram: *Description*, *Fulfillment*, *Reader Profile* e *Change of Perpective*. Além da exclusão dessas relações, a *Historical Background* que foi definida originalmente como uma relação que traz um fato histórico, neste estudo, é apenas chamada de *Background*, retirando assim essa restrição e podendo, então, ocorrer entre outros fatos que não sejam apenas históricos.

Como não se tem conhecimento de um córpus para o português do Brasil que satisfaça os requisitos deste trabalho, foi construído o primeiro córpus de notícias jornalísticas anotadas segundo a CST, as quais foram coletadas de fontes distintas. Esse córpus chamado de CSTNews possui atualmente 50 coleções de textos jornalísticos de domínios variados e cada coleção possui em média 4 documentos de diferentes fontes que versam sobre um mesmo assunto. O número total de documentos é de 195, sendo que há 3.574 sentenças e 72.148 palavras. O CSTNews foi construído nos moldes do CSTBank [5], um córpus de textos jornalísticos de assuntos relacionados coletados de fontes distintas na língua inglesa. Tal córpus possui 41 documentos com, em média, 29 palavras por sentença. Os números totais de sentenças e palavras não são informados.

Os textos foram coletados manualmente de jornais on-line por um período de 2 meses, entre Agosto e Setembro de 2007. As fontes dos textos foram os jornais on-line Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Essas fontes foram escolhidas devido a grande popularidade na web e também por trazerem as principais notícias do dia, que é o que importa para o córpus, ou seja, uma mesma

notícia publicada em fontes diferentes. Os textos jornalísticos foram escolhidos por possuírem uma linguagem clara e do dia a dia, além da facilidade de serem encontrados na web. O processo de anotação do CSTNews ocorreu durante 3 meses, de Fevereiro à Abril de 2008.

Para a anotação do córpus, dois anotadores foram devidamente treinados para tal tarefa. Uma ferramenta de suporte à anotação foi desenvolvida. A ferramenta delimita automaticamente as sentenças dos textos que estão sendo anotados e aplica a medida de similaridade lexical *Word Overlap*, como definida por [5], para indicar aos anotadores possíveis pares relacionados. Restringir via alguma medida de similaridade lexical o número de pares relacionados mostrou-se essencial para a anotação do córpus, pois, caso contrário, o número de possibilidade é tão grande que inviabiliza a anotação.

Observou-se com a anotação que as relações excluídas não fizeram diferença na identificação das mesmas. Além disso, os anotadores não sentiram a necessidade de inclusão de novas relações. De todas as possibilidades de encontrar uma relação entre duas sentenças, 46,83% dessas possibilidades, na verdade, não existiam, pois não havia relação, e 53,17% das possibilidades de fato possuíam algum tipo de relação. A porcentagem de freqüência de cada relação identificada no córpus é exibida na Tabela 1. Como esperado, algumas relações raramente aparecem, como *Citation* e *Indirect Speech*, e a relação *Translation* não foi identificada nenhuma vez no córpus.

Tabela 1: Freqüência das relações no CSTNews

Relação	Freqüência no córpus
Elaboration	23,98%
Overlap	19,85%
Subsumption	15,24%
Background	6,49%
Atributtion	5,68%
Equivalence	5,09%
Follow-up	4,72%
Contradiction	4,35%
Summary	4,35%
Identity	3,69%
Modality	3,54%
Indirect Speech	2,73%
Citation	0,29%
Translation	0,00%

Para medir a concordância entre os anotadores, foi calculada a medida Kappa [11] para um conjunto de textos com 3 documentos, 60 sentenças, 1.139 palavras e 1.043 possíveis pares de sentenças relacionadas. Os resultados obtidos da medida Kappa no geral e para as relações identificadas, são mostrados na Tabela 2.

As relações que obtiveram o maior grau de concordância entre os anotadores foram *Overlap* (0.562) e *Elaboration* (0.321). A medida Kappa calculada sobre todos os possíveis pares de sentenças foi de 0.258. Segundo [12], um resultado abaixo de 0.67 é considerado não confiável, porém, há de se considerar que na anotação CST são

possíveis 19 rótulos diferentes e não mutuamente exclusivos, o que torna o trabalho muito mais complexo.

Tabela 2: Concordância entre os anotadores e as relações

Relação	Kappa
Elaboration	0.321
Overlap	0.562
Subsumption	0.006
Follow-up	0.009
Summary	0.003
Indirect Speech	0.013
Não há relação	0.279
Total	0.258

Nos experimentos de [5], a medida Kappa foi calculada em um conjunto de textos com 7.579 possíveis pares de sentenças relacionadas e o resultado obtido foi de 0.4021, não muito diferente do experimento mostrado neste trabalho, apesar de possuir um número maior de possíveis pares de sentenças relacionadas.

3.2. Desenvolvimento do *parser* discursivo

O primeiro passo do *parser* discursivo consiste na segmentação sentencial. Para tal segmentação, foi utilizado o segmentador SENTER [13] publicamente disponível. É uma ferramenta robusta e de ampla utilização.

A segunda etapa, de identificação de segmentos relacionados, um grande experimento foi realizado. Foram implementadas e testadas as melhores medidas de similaridade lexical propostas por [5] com o objetivo de se determinar a melhor medida para a língua portuguesa. Além disso, variações das medidas foram produzidas pelo uso de recursos lingüístico-computacionais. Tais variações foram aplicadas para computar quais dessas melhorariam a identificação das relações CST aumentando a precisão e principalmente a cobertura, assumindo que uma boa cobertura traz todos os pares de sentenças que realmente possuem uma relação CST, assim como foi feito no trabalho de [10].

Além da *Word Overlap*, a medida do cosseno também foi implementada e ambas com as variações: as medidas (Cosseno e *Word Overlap*) com o uso de um *thesaurus* [14], lematização [15] e *stoplist* . Foram feitas todas as combinações possíveis com diferentes limites (*thresholds*) que variam de 0.1 a 0.5 para a *Word Overlap* e de 0.1 a 1.0 para a medida do cosseno.

Acreditava-se até então que, um *thesaurus* que traria verbos, adjetivos, substantivos e advérbios com seus antônimos e sinônimos poderia melhorar os resultados na descoberta das relações, assim como uma *stoplist*, que remove palavras pouco significativas como, preposições, artigos, entre outras.

Porém, o experimento identificou que a medida do cosseno combinada com a lematização, que reduz todas as ocorrências de uma mesma palavra sob uma única forma, sem variações de gênero, número e grau, é a melhor variante para a

identificação de segmentos relacionados. Levando em consideração que a cobertura é o fator mais importante, a medida mostrou melhor desempenho que a *Word Overlap*, com uma cobertura de 93-100% contra 53-93% com *thresholds* de 0.1 e 0.2 respectivamente. Interessante notar que o gênero dos textos não interferiu para tal resultado, mas sim a língua em foco. Um estudo mais específico será necessário para descobrir os fatores que elegem diferentes medidas para diferentes idiomas.

Tal experimento foi realizado com duas coleções de textos contendo 3 documentos cada coleção, totalizando assim, 314 sentenças, 2.440 palavras e 2.658 possíveis combinações. Nos experimentos de [10] foram utilizados 9 documentos, 269 sentenças com 18.023 possíveis combinações. O número de possíveis combinações é maior devido ao fato do experimento de [10] possuir mais documentos.

As medidas BLEU e *Longest Common Subsequence* não foram utilizadas, pois não se mostraram adequadas nos experimentos para o inglês. A razão é que tais medidas utilizam combinações de *n-grams*, o que para este experimento não é de grande valia devido a pouca freqüência de *n-grams* em comum nos textos envolvidos na anotação.

A última etapa do *parser* discursivo, de detecção das relações CST entre as sentenças relacionadas, ainda se encontra em desenvolvimento. Pretende-se utilizar duas abordagens: (a) heurísticas de detecção de relações conforme os padrões verificados no córpus anotado e (b) aprendizado de máquina. Deverão ser utilizadas ferramentas lingüístico-computacionais para suporte à análise. Prevê-se, no momento, o uso de um analisador sintático-semântico, mais especificamente, o PALAVRAS [16] e um *thesaurus* para a língua portuguesa [14]. Para a abordagem de aprendizado de máquina, deverão ser avaliados os algoritmos utilizados na literatura relacionada e outros novos.

4. Considerações finais

Acredita-se que há várias contribuições provenientes deste trabalho. Essa é a primeira pesquisa nessa linha para o português do Brasil, língua tão carente em pesquisas em análise discursiva, em geral. Córpus e recursos lingüísticos inéditos vêm sendo construídos e disponibilizados no decorrer do projeto. Metodologias de análise diferentes estão sendo avaliadas e sua potencialidade de aplicação para a questão de pesquisa em foco está sendo averiguada, como a variação da medida de similaridade lexical entre o inglês e o português descrita.

O próximo e último passo para a automatização de todo o processo de identificação das relações CST, é implementar e avaliar as abordagens para detecção de relações. A validação do *parser* CSTTool poderá ser feita em um sumarizador multidocumento já construído, por exemplo, o GistSumm [17].

Agradecimentos

Os autores agradecem à FAPESP, à CAPES e ao CNPq pelo suporte ao projeto.

Referências

1. 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.
2. Zhang, Z.; Otterbacher, J.; Radev, D.R. (2003). Learning Cross-document Structural Relationships using Boosting. In the *Proceedings of ACM CIKM 2003*. New Orleans, LA.
3. Zhang, Z. and Radev, D.R. (2004). Combining Labeled and Unlabeled Data for Learning Cross-Document Structural Relationships. In the *Proceedings of IJCNLP 2004*, pp. 32-41.
4. Zhang, Z. (2005). Natural Language Relations: Classification and Aplicattion. PhD Thesis, University of Michigan.
5. 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*.
6. Salton, G. and Lesk, M.E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, Vol. 15, N. 1, pp. 8-36.
7. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318. Philadelphia, PA.
8. Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.
9. Freund, Y. and Schapire, R. E. (1997). A decision theoric generalization of on-line learning and an application on boosting. *Journal of Computer and System Sciences, 119-139.*
10. Zhang, Z.; Blair-Goldensohn, S.; Radev, D.R. (2002). Towards CST-enhanced summarization. In the Proceedings of the AAAI 2002 Conference. Edmonton, Alberta.
11. Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. 22(2):249–254.
12. Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.
13. Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.
14. Dias da Silva, B.C.; Oliveira, M.F.; Moraes, H.R.; Hasegawa, R.; Amorim, D.; Paschoalino, C.; Nascimento, A.C. (2000). A Construção de um Thesaurus Eletrônico para o Português do Brasil. In *Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*.
15. Nunes, M.G.V. et al. (1996). The Design of a Lexicon for Brazilian Portuguese: Lessons Learned and Perspectives. In the Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese, pp. 61-70.
16. Bick, E. (2000). The Parsing System PALAVRAS: Automatic Gramatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press.
17. Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, pp. 210-218. Springer-Verlag, Germany.
18. Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.
19. 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.
20. Radev, D.R.; Blair-Goldensohn, S.; Zhang, Z.; Raghavan, R.S. (2001b). Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In the *Proceedings of Human Language Technology Conference.* San Diego, CA.

21. Radev, D.R.; Blair-Goldensohn, S.; Zhang, Z. (2001a). Experiments in single and multidocument summarization using MEAD. In the *Proceedings of the First Document Understanding Conference*. New Orleans, LA.
22. Otterbacher, J.C. (2006). *Short-term Event Tracking in Dynamic Online News*. P.H.D Thesis. University of Michigan. 223 p.