Em Direção à Caracterização de Sumários Humanos Multidocumento

Renata Tironi de Camargo[1,2], Ariani Di Felippo[1,2], Thiago A. S. Pardo[2]

[1] Departamento de Letras (DL) – Centro de Educação e Ciências Humanas (CECH)
Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905– São Carlos – SP – Brasil

[2] Núcleo Interinstitucional de Linguística Computacional (NILC)
Inst. de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 - São Carlos, - SP - Brasil

renatatironi@hotmail.com, arianidf@gmail.com, taspardo@icmc.usp.br

Abstract. *We present a proposal for characterizing human multi-document summaries based on corpus analysis. Specifically, we intend to identity content selection strategies and summary production operations commonly used by humans that could guide the automatic multi-document summarization.*

Resumo. *Apresenta-se uma proposta de caracterização linguística de sumários humanos multidocumento com base em análise de corpus. Especificamente, busca-se identificar estratégias humanas de seleção de conteúdo e de produção dos sumários que possam subsidiar a sumarização automática multidocumento.*

1. Introdução

Na subárea do Processamento Automático das Línguas Naturais denominada Sumarização Automática (SA), busca-se produzir automaticamente sumários (ou resumos) de um ou mais textos. No primeiro caso, diz-se que a SA é monodocumento e, no segundo, multidocumento [Mckeown e Radev 1995]. Em termos de formação, sumários podem ser classificados como extratos ou *abstracts* [Mani *et al,* 1999]. Os extratos são sumários compostos por trechos inalterados do texto-fonte. Os *abstracts*, por sua vez, são formados por trechos reescritos dos textos-fonte. A SA pode ser superficial ou profunda em função do nível de conhecimento linguístico que considera [Mani 2001]. Na abordagem superficial, utiliza-se pouco ou nenhum conhecimento linguístico para produzir sumários. A abordagem profunda de SA caracteriza-se pela utilização de teorias/modelos linguísticos que guiam a sumarização.

De modo geral, dois dos principais desafios da SA são (i) reconhecer a informação principal nos textos-fonte e (ii) produzir o sumário. Independentemente do tipo de sumário, da quantidade de textos-fonte sob processamento e da abordagem empregada, esses dois processos são reconhecidamente centrais na SA.

A SA monodocumento é uma tarefa consolidada e, por isso, inúmeros métodos superficiais e profundos têm sido propostos, principalmente com base no entendimento e na sistematização das tarefas humanas de seleção de conteúdo e produção de sumários [Mani 2001]. O interesse pela sumarização automática multidocumento (SAM) é recente e tem se fortalecido diante do volume de informação similar ou repetida

Anais da II Jornada de Descrição do Português, páginas 47–54, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

disponível na *web* com a qual o consulente precisa lidar. Para o português do Brasil (PB), em especial, foram desenvolvidos dois sumarizadores multidocumento: o GIST SUMMarizer [Pardo 2005] e o CST SUMMarizer [Jorge e Pardo 2010]. O GIST SUMMarizer é um sistema baseado em conhecimento linguístico superficial. Para a realização da sumarização multidocumento, todos os textos são justapostos e o processo de sumarização é guiado pela informação da sentença que contém as palavras mais frequentes dos textos (*gist sentence*). Já o CST SUMMarizer produz sumários a partir da anotação das sentenças provenientes de dois ou mais textos com as relações discursivas (p.ex.: *Identity, Equivalence, Subsumption, Contradiction*, etc.) do modelo linguístico-computacional CST (*Cross-document Structure Theory*) [Radev 2000]. O CST SUMMarizer aplica vários critérios para a seleção das sentenças que comporão sumário, os quais são guiados pela redundância entre as sentenças dos textos-fonte, considerada uma pista importante dos tópicos centrais dos mesmos.

Apesar de o desenvolvimento visto nos últimos anos, a SAM não dispõe de subsídios linguísticos específicos, sistematizados e formalizados, sobre a tarefa de sumarização humana multidocumento (SHM), em especial, sobre a seleção de conteúdo e produção de sumários. Assim, propõe-se, com base em análise de *corpus,* caracterizar os sumários humanos multidocumento. Para apresentar tal proposta, divide-se este artigo em 5 Seções. Na Seção 2, apresenta-se uma revisão sobre a caracterização da sumarização humana e sua relação com a SA. Na Seção 3, apresenta-se o *corpus* a ser utilizado. Na Seção 4, apresentam-se os passos para a realização da pesquisa. Na Seção 5, algumas considerações finais são feitas.

2. Revisão Bibliográfica

Vários autores têm buscado compreender e sistematizar o modo como os humanos geram versões condensadas a partir de um único documento. De modo geral, reconhece-se que a sumarização humana envolve três processos: (i) exploração do documento (*document exploration*), (ii) avaliação de relevância (*relevance assessment*) e (iii) produção do sumário (*summary production*) [Cremmins 1996; Endres-Niggemeyer 1998]. Buscando emular a sumarização realizada pelos humanos na máquina, Mani e Maybury (1999) sugerem que a SA envolva idealmente três processos: (i) análise dos textos-fonte, em que se produz uma representação completa de seu conteúdo; (ii) transformação, em que o conteúdo completo do texto-fonte é condensado e (iii) síntese, em que o conteúdo condensado é expresso em língua natural na forma de um sumário.

O processo de sumarização humana denominado "avaliação de relevância" [Endres-Niggemeyer 1998] é comumente definido na arquitetura de um sistema de SA como "seleção de conteúdo" e compõe a etapa de transformação. O ponto central da seleção de conteúdo é reconhecer as unidades de significado do texto-fonte (p.ex.: palavras, sintagmas, orações, sentenças, etc.) que contêm a ideia central do mesmo para compor o sumário [Mani 2001]. Do ponto de vista humano, tal tarefa é bastante problemática e controversa, pois a importância de uma unidade textual depende de vários fatores, p. ex.: (i) a tipologia e o gênero dos textos-fonte; (ii) os objetivos do autor do sumário, (iii) os interesse dos possíveis leitores do sumário, etc. Esses fatores tanto influenciam que sumários distintos podem ser produzidos para um mesmo texto-fonte. Apesar dessa variação, os humanos comumente selecionam as unidades de significado (p.ex.: sentenças) de um texto-fonte (jornalístico) que irão compor o sumário monodocumento com base em algumas estratégias superficiais, a saber: (i)

palavras-chave, estratégia segundo a qual os sumarizadores humanos selecionam as sentenças que contêm as palavras-chave do texto para compor o sumário, (ii) palavras-chave do título, segundo a qual sentenças do texto que contêm as palavras-chave do título são selecionadas para compor o sumário, (iii) localização, segundo a qual informação localizada no início e no final do texto (principalmente, jornalístico) é selecionada para compor o sumário e (iv) expressões sinalizadoras, estratégia segundo a qual os humanos selecionam informação com base na identificação de expressões como "em suma", "o objetivo deste trabalho é", etc.; tais expressões variam em função do tipo/gênero textual [Cremmins 1996; Endres-Niggemeyer 1998].

Na trajetória da SA monodocumento, identificam-se vários métodos superficiais propostos com base em uma ou mais estratégias de seleção de conteúdo sistematizadas por Endres-Niggemeyer (1998) [Uzêda *et al.,* 2010; Wan *et al.*, 2010]. Em métodos profundos que se baseiam na representação do texto-fonte segundo a RST[1] (*Rhetorical Structure Theory*) [Mann e Thompson 1987], a seleção é feita pela correlação das estratégias humanas superficiais à estruturação RST. Por exemplo, a primeira sentença de um texto jornalístico, selecionada com base na estratégia de localização, é também a mais nuclear em uma árvore RST bem construída do mesmo texto.

A "produção de sumários" pelos humanos, por sua vez, é guiada por várias operações de "recorta e cola" do texto-fonte: (i) deleção (*truncation*), (ii) inserção (*insertion*); (iii) substituição (*replacement*); (iv) reordenação (em inglês, *reordering*); (v) amálgama (*mergging* ou *aggregation*); (vi) generalização/ especialização (*generalization/ specialization*); e (vii) paráfrase lexical (*lexical paraphrasing*) [Jing e Mckeown, 1999, 2000]. Com base na primeira sentença do texto-fonte da Figura 1 e no sumário manual S1, vê-se várias operações desse tipo. Por exemplo, identificam-se as seguintes deleções: (i) do determinante "a", que antecede o nome "empresa"; (ii) do sintagma nominal "Produtos Pirata Indústria e Comércio Ltda.", que tem função de especificador; (iii) do sintagma preposicional "de Contagem" e (iv) da informação parentética. Além das deleções, identifica-se a inserção do adjetivo "mineira" (especificador de "empresa"), inferido da informação de localização da empresa.

Texto-fonte
[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem [2] (região metropolitana de Belo Horizonte), [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente.

Sumário
S1: Empresa mineira deverá registrar este ano um aumento de produtividade nas áreas comercial e industrial de 11% e 17%, respectivamente.

Figura 1: Exemplo de sumário manual.

Assim como as estratégias de seleção, as operações de produção têm fundamentado o processo de síntese, em vários métodos superficiais e profundos [Jing e Mckeown 2000]. Para exemplificar a aplicação das operações na etapa de síntese, supõe-se que um sumarizador superficial e extrativo tenha selecionado a primeira

[1] De acordo com a RST, as unidades de um texto (p.ex.: sentenças) podem ser relacionadas por relações discursivas como causa-efeito, contraste, etc. A RST é amplamente utilizada para a SA monodocumento principalmente por distinguir, dada uma relação, o segmento nuclear (N) e o satélite (S).

sentença do texto-fonte da Figura 1 com base no método da localização e que outro sumarizador, profundo e extrativo, tenha selecionado a mesma sentença com base na nuclearidade da mesma na árvore RST do texto-fonte. Munidos, por exemplo, da operação de "deleção de parênteses", esses sistemas são capazes de produzir sumários sem a informação contida entre.

Quanto à SHM, não há trabalhos que buscam compreender o modo como os humanos produzem um resumo a partir de dois ou mais textos relacionados. Segundo Mani (2001), a sumarização multidocumento é pouco intuitiva para os humanos. Contudo, McKeown *et al*. (2005) demonstraram que sumários multidocumentos automáticos ou manuais (humanos) são úteis em experimentos que simulavam a apreensão de informação por humanos. Na literatura, é possível identificar alguns indícios de como um sumário multidocumento é elaborado pelos humanos.

Quanto à seleção de conteúdo, evidências empíricas demonstram que o humano seleciona um texto de sua preferência como base para selecionar as informações principais para compor o sumário e, na sequência, recorre aos demais textos da coleção para complementar as informações do sumário [Mani 2001]. A informação principal selecionada a partir do texto-base é provavelmente a informação mais redundante, pois é usual que a mídia destaque o foco das notícias; tal fato, empiricamente observado, mas relatado de forma dispersa e implícita na literatura, tem guiado a SAM.

Nos métodos superficiais mais simples e eficientes de SAM, as unidades de significado (p.ex.: sentenças) com mais sobreposição de conteúdo (ou seja, itens lexicais comuns) entre si são selecionadas para compor o sumário [McKeown e Radev 1999]. No GIST SUMMarizer, por exemplo, as sentenças são selecionadas para compor o sumário em função do número de palavras em comum com a *gist sentence*.

Nos métodos profundos baseados na representação dos textos-fonte segundo a teoria/modelo linguístico-computacional CST, de natureza semântico-discursiva (*Cross-document Structure Theory*) [Radev 2000], as unidades de significado redundantes são comumente identificadas pelo tipo de relação entre as sentenças, por exemplo, *Identity* e *Equivalence* (redundância total)*, Subsumption* (redundância parcial), etc., ou pelo número de relações entre elas, ou seja, quanto mais relações CST houver entre as sentenças de textos distintos, maior a chance de haver sobreposição de conteúdo entre elas, o que indica a importância das mesmas para a composição do sumário. Os trechos de notícia da Figura 3, coletados de fontes distintas, relatam um mesmo acidente aéreo. Neles, algumas relações CST são facilmente identificadas. Por exemplo, a sentença [1] do texto 1 e a sentença [1] do texto 2 estão ligadas pela relação *Attribution*, pois tais sentenças apresentam informação em comum, sendo que a sentença [1] do texto 2 atribui essa informação a uma fonte/autoria. Outra relação entre as mesmas unidades também pode ser identificada. No caso, a relação é a *Subsumption*, já que a sentença [1] do texto 2 apresenta, além do mesmo conteúdo da sentença [1] do texto 1, informações adicionais. Na Figura 2, a sentença [1] do texto 1 e a sentença [1] do texto [2] estão relacionadas por duas relações CST, o que indica sobreposição de conteúdo e, consequentemente, maior relevância dos segmentos em detrimento dos demais com menor número de relações. Diferentemente dos métodos superficiais, essa sobreposição de conteúdo não é baseada unicamente nos itens lexicais das sentenças, mas na relação identificada entre elas.

Texto-fonte 1
[1] Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
[2] Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
[3] A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.

Texto-fonte 2
[1] 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.
[2] As vítimas do acidente foram 14 passageiros e 3 membros da tripulação.
[3] 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: Textos sobre um mesmo assunto provenientes de fontes distintas.

Quanto à produção dos sumários multidocumento, desconhecem-se trabalhos que tenham investigado e sistematizado operações do tipo *cut and paste* na SHM. No entanto, algumas dessas operações têm sido empregadas no cenário multidocumento, de forma similar à que é feita na SA monodocumento [Otterbacher e Radev 2004].

Assim, apesar dos indícios de como a SHM é feita, não se sabe de tentativas de caracterizações linguísticas dos sumários multidocumento. Em outras palavras, não se sabe de trabalhos que tenham investigado, p.ex.: (i) a origem das informações do sumário multidocumento, (ii) a influência do tipo das relações CST na seleção da informação que compõe o sumário, (iii) o nível mínimo de redundância (expressa também pelas relações da CST) esperada para que uma sentença seja considerada relevante, etc. Ademais, desconhecem-se trabalhos que tenham investigado as operações do tipo *cut and paste* no cenário da SHM.

Assim, diante dos indícios sobre a SHM, formulam-se duas hipóteses centrais ao trabalho: (i) há estratégias recorrentes de seleção de conteúdo na SHM, além da escolha pela informação mais redundante, e (ii) há correlação entre tais estratégias e a modelagem CST. Para verificar as hipóteses formuladas, objetiva-se: (i) investigar a SHM com vistas à identificação e sistematização de estratégias de seleção de conteúdo, (ii) verificar se há correlação entre as estratégias de seleção e a modelagem CST e (iii) formalizar as estratégias de seleção (i-ii) por meio de regras explícitas que auxiliem a SAM. Com base na literatura, formula-se ainda uma terceira hipótese, no caso, a de que há operações *cut and paste* recorrentes na produção dos sumários. Assim, buscar-se-á caracterizar os sumários multidocumento em função das operações de *cut and paste*. Para tanto, partir-se-á de um *corpus* composto por conjuntos de textos relacionados via CST e seus respectivos sumários humanos.

3. O *corpus*

Dentre os *corpora* multidocumento disponíveis, destacam-se o CSTBank [Radev e Otterbacher 2003] e o CSTNews [Aleixo e Pardo 2008; Maziero *et al.* 2010]. O CSTBank, primeiro *corpus* multidocumento da literatura, é composto por 6 coleções de textos jornalísticos em língua inglesa, os quais foram anotados manualmente via as relações do modelo CST.

O CSTNews, por sua vez, é um *corpus* em PB anotado via CST, composto por 50 coleções de textos jornalísticos de domínios variados (p.ex.: mundo, política, ciência, esporte, cotidiano, etc.). Os textos foram coletados manualmente de jornais online, a saber: Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Cada coleção do CSTNews é composta por em média 3 textos que tratam do mesmo assunto e seu respectivo sumário humano. No caso, optou-se por utilizar o CSTNews tendo em vista o foco do projeto na língua portuguesa e o fato de este recurso englobar os sumários humanos multidocumento, os quais se pretende caracterizar neste trabalho.

4. Etapas metodológicas

A partir da seleção de uma ou mais coleções do CSTNews, foram especificadas 6 tarefas com o objetivo de verificar as hipóteses formuladas na Seção 2, a saber: (i) alinhamento dos textos-fonte de uma mesma coleção a seu sumário, (ii) caracterização das sentenças dos textos-fonte quanto às estratégias de seleção de conteúdo, (iii) caracterização das sentenças dos textos-fonte quanto às relações CST, (iv) identificação de estratégias de SHM com base nas caracterizações realizadas em (iii) e (iv), (v) representação formal das estratégias identificadas em (iv) e avaliação das mesmas e (vi) caracterização dos sumários em função das operações *cut and paste* e

a) Alinhamento dos textos-fonte a seu sumário humano: Atualmente está sendo realizada a tarefa de alinhamento manual dos textos-fonte ao seu respectivo sumário humano. O alinhamento está sendo feito em nível sentencial com base na similaridade ou sobreposição de conteúdo. No exemplo da Figura 3, a sentença [1] do texto-fonte 1 e a sentença [1] do texto-fonte 2 foram alinhada à sentença [1] do sumário humano e a sentença [2] do texto-fonte 2, à sentença [2] do sumário. Ao final desse processo, ter-se-á, dado um *cluster,* o conjunto de sentenças de cada texto-fonte cujo conteúdo foi transposto para o sumário e o conjunto de sentenças cujo conteúdo não foi transposto para o sumário humano.

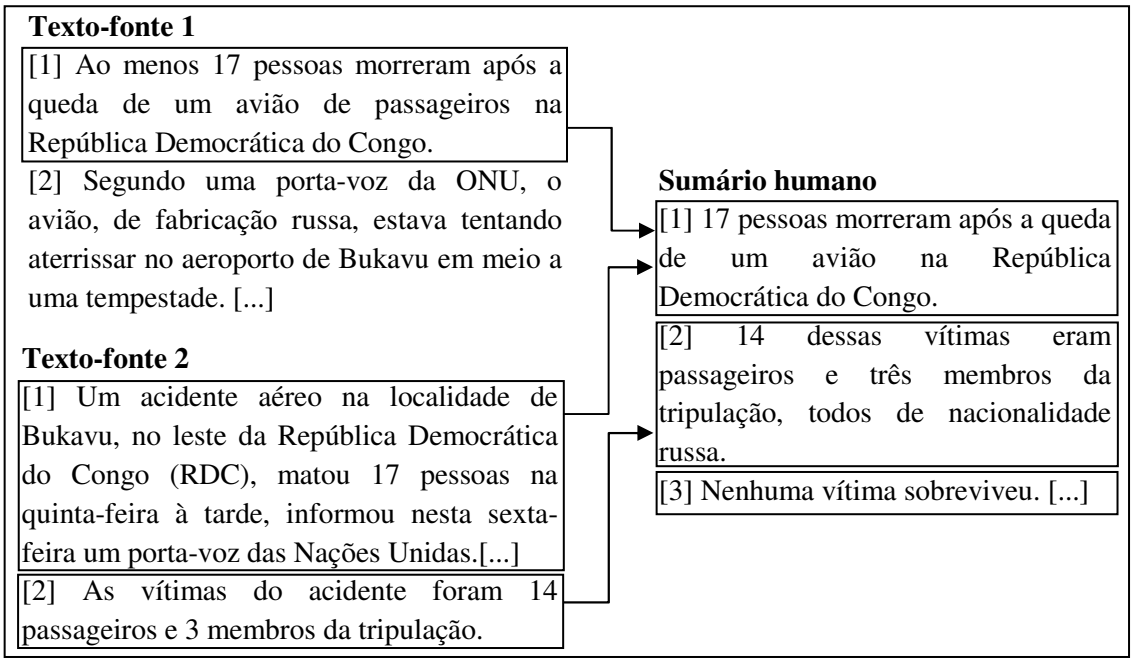

Figura 3: Exemplo de alinhamento de textos-fonte de um *cluster* ao sumário humano.

b) Caracterização das sentenças dos textos-fonte quanto às estratégias superficiais de seleção de conteúdo: Após o alinhamento manual (a), todas as sentenças dos textos-fonte serão caracterizadas em função das principais estratégias superficiais de seleção de conteúdo, a saber: (i) frequência das palavras da sentença, (ii) posição da sentença no texto-fonte, (iii) número de palavras do título presentes na sentença e (iv) número de palavras-chave presentes na sentença. Para tanto, as *stopwords* serão removidas dos textos-fonte e as palavras dos mesmos serão lematizadas. Para a caracterização de uma sentença S em função da frequência de suas palavras constitutivas, por exemplo, *w1, w2* e *w3,* calcular-se-á a frequência de cada *w* de S em todo o *cluster* e, na sequência, a frequência de *w1, w2* e *w3* em todo o *cluster* será somada, caracterizando S quanto à estratégia "frequência". Sobre a caracterização de S quanto à posição, ressalta-se que o primeiro parágrafo do texto-fonte será considerado "início", o último será considerado "fim" e os demais parágrafos intermediários, "meio". Para a caracterização de S quanto às palavras-chave, considerar-se-á como palavras-chave as unidades lexicais que compõem o conjunto das 10% mais frequentes de todo o *cluster.*

c) Caracterização das sentenças dos textos-fonte quanto às relações CST: Além da caracterização descrita em (b), buscar-se-á nesta etapa responder: (i) As sentenças dos textos-fonte selecionadas pelos humanos para compor o sumário estão sempre anotadas por relações CST? (ii) Se sim, tais sentenças indicam relações CST predominantes na seleção do conteúdo para compor o sumário? (iii) Se não, há outros fatores interferindo na seleção (p. ex.: a posição da sentença, data de publicação do texto-fonte, etc.)? Para tanto, a partir do alinhamento descrito em (a), caracterizar-se-ão as sentenças dos textos-fonte quanto à presença dos tipos de relações estabelecidos por Maziero *et al.* (2010): relações de conteúdo (*redundância, complemento* e *contradição*) e de forma (*fonte/autoria* e *estilo*). Ao final, espera-se diferenciar, em função da CST, as sentenças cujo conteúdo foi selecionado para o sumário das sentenças cujo conteúdo não foi selecionado.

d) Identificação de estratégias de seleção na SHM: Consiste em identificar, com base nas caracterizações realizadas em (b) e (c), regras gerais de sumarização (ou seja, de seleção de conteúdo) empregadas pelos humanos. Além disso, pretende-se comparar tais regras manuais com os resultados gerados pelo aprendizado de máquina. Para tanto, as caracterizações realizadas em (b) e (c) serão transformadas em um *frame,* no qual cada estratégia superficial e cada relação CST serão caracterizadas como *features* (traços) das sentenças a partir dos quais a máquina infere regras de sumarização.

e) Formalização e avaliação: Consiste em especificar regras formais de seleção de conteúdo dos textos-fonte baseadas nas estratégias identificadas nas tarefas já explicitadas acima. Por regras "formais", entendem-se regras suficientemente explícitas e não-ambíguas de modo a serem computacionalmente aplicáveis à tarefa de SAM. Na sequência, buscar-se-á avaliar tais estratégias por meio da utilização das mesmas em um método de SAM.

f) Caracterização dos sumários em função das operações *cut and paste*: Consiste na identificação de operações de "recorta e cola" utilizadas por humanos para se produzir o sumário, ou seja, cada sentença será analisada para que seja possível identificar quantos tipos de operações foram encontradas e quais são elas. Esta etapa, no entanto, se baseia em uma proposta apenas, pois será completada apenas se possível.

5. Considerações Finais

O trabalho está em estágio inicial, concentrando-se nas tarefas de revisão bibliográfica e alinhamento dos textos-fonte do CSTNews a seus respectivos sumários humanos. Ao final, espera-se identificar estratégias que possam efetivamente subsidiar a sumarização automática multidocumento, especificamente em PB.

6. Referências

Aleixo, P. e Pardo, T.A.S. (2008) CSTNews: um *corpus* de textos jornalísticos anotados segundo a Teoria Discursiva Multidocumento CST (*Cross-document Structure Theory*). *Série de Relatórios Técnicos do ICMC*, São Carlos-SP, n. 326, 12p.

Cremmins, E.T. (1996) *The art of abstracting*. Arlington, Virginia: Information Resources Press.

Endres-Niggemeyer, B. (1998) *Summarization Information*. Springer, Berlin.

Jing, H. and McKeown, K. R. (1999) The decomposition of human-written summary sentence. In the *Proceedings of the 22th International ACM-SIGIR*, New York, p. 129-136.

____. (2000) Cut and paste based text summarization. In the *Proceedings of the NAACL Conference*, San Francisco, p. 178-185.

Jorge, M.L.C. and Pardo, T.A.S. (2010) Experiments with CST-based Multidocument Summarization. In the *Proceedings of the 5th ACL Workshop TextGraphs* 2010, Uppsala, Sweden, p. 74-82.

McKeown, K. and Radev, D. R. (1995) Generating summaries of multiple news articles. In the *Proceedings of the 18th International ACM-SIGIR*, Seattle, p. 74-82.

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

Mani, I. *et al.* (1999) Improving summaries by revising them. In the *Proceedings of the 37th Annual Meeting of the ACL*, New Brunswick, New Jersey, p. 558-565.

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

Maziero, E. G., Jorge, M. L. C. and Pardo, T. A. S. (2010) Identifying Multidocument Relations. In the *Proceedings of the 7th NLPCS*, Funchal, Portugal, p. 60-69.

Otterbacher, J. and Radev, D. R (2004) A resource for revision-based multi-document summarization and evaluation. In the *Proceedings of the 4th LREC*, Lisbon, Portugal.

Pardo, T.A.S. *GIST SUMMarizer*: extensões e novas funcionalidades. *Série de Relatórios do NILC*. NILC-TR-05-05. São Carlos-SP, 8p., 2005.

Radev, D. R. (2000) A common theory of information fusion from multiple text sources, step one: cross-document structure. In the *Proceedings of the ACL SIGDIAL Workshop on Discourse and Dialogue*, Hong Kong, p. 74-86.

Sparck Jones, K. (1995) Discourse modeling for Automatic Summarisation. *Tech. Report No. 290.* University of Cambridge. UK, February.

Uzêda, V.R.; Pardo, T.A.S. and Nunes, M.G.V. (2010). A comprehensive comparative evaluation of RST-based summarization methods. *ACM Transactions on Speech and Language Processing*, vol. 6, n. 4, p. 1-20.

Wan, X.; LI, H. and XIAO, J. (2010) Cross-language document summarization based on machine translation quality prediction. In the *Proceedings of the 48th Annual Meeting of the ACL*, Uppsala, Sweden, p. 917–926.