# Investigação da Identificação da Redundância na Sumarização Multidocumento

**Jackson Wilke da Cruz Souza[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 – Brazil

[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 – Brazil

{jackcruzsouza,arianidf}@gmail.com, taspardo@icmc.usp.br

***Abstract.*** *Considering the necessity of studies on multi-document summarization involving Brazilian Portuguese (PB), we present a research proposal that aims at investigating methods to identify the redundancy between sentences to be included in a multi-document summary in PB and the correspondence between the methods and CST redundancy relations (Cross-document Structure Theory) [Radev, 2000].*

***Resumo.*** *Diante da necessidade de trabalhos sobre sumarização automática multidocumento que envolvem o português do Brasil (PB), apresenta-se uma proposta em que se objetiva investigar (i) métodos de identificação da redundância entre sentenças a serem inseridas em um sumário multidocumento em PB e a (ii) correspondência entre esses métodos e as relações de redundância da CST (Cross-document Structure Theory) [Radev, 2000].*

## 1. Introdução

Na Sumarização Automática Multidocumento (SAM), subárea do Processamento Automático de Língua Natural (PLN), busca-se produzir sumários a partir de uma coleção de textos-fonte que abordam um mesmo tópico [Mckeon e Radev 1995]. As pesquisas em SAM são motivadas pela enorme quantidade de informação disponível, principalmente na *web,* e pelo pouco tempo que as pessoas têm para assimilar tanta informação [Sparck Jones 2007]. Na SAM, o tratamento da redundância é um dos principais tópicos de pesquisa [Jurafsky e Martin 2009], pois um sumário multidocumento deve conter o conjunto de sentenças que melhor representa o tópico ou assunto da coleção sem que haja informação repetida entre elas. Para tanto, é preciso identificar a redundância entre as sentenças provenientes de fontes distintas. Diante da necessidade de trabalhos sobre SAM que envolvem o português do Brasil (PB), propõe-se investigar métodos para identificação da redundância entre sentenças a serem inseridas em um sumário multidocumento em PB. Ademais, pretende-se investigar a correspondência entre esses métodos e as relações de redundância do modelo CST (*Cross-document Structure Theory*) [Radev 2000]. Na próxima Seção, faz-se uma breve revisão sobre a redundância no cenário da SAM. Na Seção 3, descrevem-se as etapas previstas para a investigação dos métodos de identificação da redundância. Na Seção 4, algumas considerações finais sobre a proposta são feitas.

## 2. Breve Revisão Bibliográfica

Uma das principais diferenças entre a SA monodocumento e a SAM diz respeito ao grande volume de informação redundante com o qual se lida ao sumarizar uma coleção de documentos que tratam de um mesmo assunto. As relações de redundância entre os segmentos (sentenças) de textos-fonte distintos podem ser classificadas em parciais ou totais.

**Quadro 1. Exemplo de Redundância Parcial e Total.**

| Tipo de redundância | Textos | Sentenças |
|---|---|---|
| Redundância parcial<br><br>(S1 contém X e Y, S2 contém X e Z) | Texto 1 | S1: A falha no reversor – **mecanismo que ajuda o avião a frear** – foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. |
| | Texto 2 | S2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, **um Airbus A320**, continuou voando, com o reverso direito desligado. |
| Redundância Total | Texto 1 | S1: Na capital, houve ataques a outros quatro ônibus. |
| | Texto 2 | S2: Na capital, houve ataques a outros quatro ônibus. |

No Quadro 1, cujos exemplos foram retirados de Maziero *et al.* (2010), a redundância parcial entre as sentenças S1 e S2, provenientes de textos distintos, caracteriza-se pelo fato de que S1 e S2 apresentam informações em comum (trechos sublinhados) e ambas apresentam informações adicionais distintas entre si (trechos em negrito). Diz-se que S1 contém X e Y, S2 contém X e Z. A redundância total entre S1 e S2, por sua vez, caracteriza-se pelo fato de que tais sentenças apresentam identidade total de forma e conteúdo. De um modo geral, a redundância pode ser identificada por meio de métodos superficiais e profundos. Dentre os superficiais mais simples, citam-se os que se baseiam no número de palavras em comum entre as sentenças [Jurafsky e Martin 2009]. Tais métodos comumente utilizam medidas como: *cosseno*, *word overlap* e outras. Outros métodos superficiais, linguisticamente mais ricos, identificam a redundância com base em: (i) sobreposição de sujeito, de sequência de etiquetas morfossintáticas e de entidades nomeadas; (ii) tamanho das sentenças e (iii) localização das sentenças no texto-fonte. Dentre os métodos profundos, citam-se os que identificam a redundância com base no tipo de relação CST detectada entre as sentenças [Mani 2001]. Na próxima Seção, apresentam-se as etapas previstas para a investigação ora proposta.

## 4. Metodologia

Além da revisão da literatura sobre SAM e métodos de identificação de redundância, as seguintes etapas são propostas para a investigação de métodos de detecção de redundância e da possível correspondência entre estes e as relações de redundância da CST:

(a) Seleção do *corpus*: consiste em selecionar do *corpus* para o estudo os métodos. No caso, o *corpus* é o CSTNews [Aleixo e Pardo 2008], posto que este é o único em PB alinhado no nível retórico via CST. O CSTNews é composto por 50 coleções de textos jornalísticos de domínios variados. Com base na tipologia de relações CST de Maziero *et al.* (2010), serão selecionados pares de sentenças: (i) completamente redundantes,

relacionadas pelas relações *Identity, Equivalence* e *Summary*, (ii) parcialmente redundantes, relacionados pelas relações *Subsumption* e *Overlap,* e não redundantes (p.ex.: sentenças de coleções distintas).

(b) Teste dos métodos de identificação da redundância: consiste em aplicar os métodos às sentenças selecionadas em (a). Objetiva-se aplicar os métodos superficiais mais simples, baseados no número de palavras em comum entre as sentenças, e os linguisticamente mais ricos, baseados, por exemplo, na (i) sobreposição de sujeito, de sequência de etiquetas morfossintáticas e de entidades nomeadas, no (ii) tamanho das sentenças e na (iii) localização das sentenças no texto-fonte. Após a aplicação dos testes, obter-se-á um ranqueamento das sentenças do *corpus* quanto à redundância.

(c) Estudo da correlação entre os métodos e as relações CST: consiste em verificar a correlação entre os métodos aplicados em (b) (isolados ou em conjunto) e as relações CST anotadas entre as sentenças dos pares selecionados. Em outras palavras, busca-se verificar quais métodos expressam adequadamente as diferenças de redundância.

(d) Avaliação: consiste em identificar manualmente os métodos que melhor descrevem o *corpus* e seus tipos de redundância.

## Agradecimento

À FAPESP, pelo apoio financeiro (Proc. 2011/07637-9) concedido a este trabalho.

## 5. Considerações Finais

No estágio da pesquisa, em paralelo ao levantamento bibliográfico, está sendo realizada a tarefa de seleção das sentenças do *corpus*. Na sequência, os métodos de identificação da redundância serão delimitados e aplicados aos pares de sentenças extraídos do CSTNews. Por fim, salienta-se que este trabalho busca identificar métodos que mais adequadamente discriminam sentenças total e parcialmente redundantes, gerando, assim, subsídios para o desenvolvimento da SAM para o PB.

## Referências Bibliográficas

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.

Jurafsky, D. e Martin, J.H (2009) Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall, New Jersey.

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

Mckeown, K. e Radev, D.R. (1995) "Generating summaries of multiple news articles". In the Proceedings of the 18th ACM-SIGIR, Seattle, p. 74-82.

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

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 Signal Workshop on Discourse and Dialogue, Hong Kong, p. 74-83.

Sparck Jones, K (1993) Discourse modeling for Automatic Summarisation. Tech. Report No. 290. University of Cambridge, UK.