# Proposta de Sumarização Automática Multidocumento usando modelos semântico-discursivos

**Paula C. F. Cardoso, Thiago A. S. 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

{pcardoso,taspardo}@icmc.usp.br

***Abstract.*** *Automatic text summarization are computational systems that are designed to select the most important information in a text to produce a shorter version called a summary. This article proposes a multi-document summarization based on semantic-discursive models to produce summaries more informative and consistent. Summarization strategies will be applied to texts in Portuguese journalistic in nature.*

***Resumo.*** *Sumarizadores automáticos de textos são sistemas computacionais que têm o objetivo de selecionar as informações mais importantes de um texto para produzir uma versão mais curta chamada de sumário. Este artigo apresenta uma proposta de sumarização automática multidocumento com base em modelos semântico-discursivos para produção de sumários mais informativos e coerentes. As estratégias de sumarização serão aplicadas a textos em português de caráter jornalístico.*

## 1. Introdução

Desde o inicio da computação, o homem procurou maneiras de tornar a comunicação mais natural entre usuário e máquina, de modo que isso aconteça por meios de línguas naturais, e não por meio de instruções e comandos. Uma das áreas que se preocupa em solucionar problemas dessa natureza é o Processamento de Línguas Naturais (PLN). O PLN é visto tradicionalmente como subárea da Inteligência Artificial (IA), e lida com diversas tarefas, como tradução automática, perguntas e respostas, revisão ortográfica e gramatical, e sumarização automática (SA, foco desta pesquisa), dentre várias outras, visando tornar o computador apto a manipular adequadamente a língua humana, quer seja em sua geração e/ou interpretação.

Com o advento da internet e a enorme quantidade de informação disponível, principalmente on-line, e o tempo cada vez mais escasso que as pessoas têm para absorver toda essa informação, a SA mostra-se como uma atividade muito importante (Radev e Mckeown, 1998). Para um usuário é impossível ler a quantidade de textos de um assunto de seu interesse em pouco tempo. Um sumarizador automático de textos é um sistema computacional que seleciona as informações mais importantes de um texto para produzir uma versão mais curta, geralmente chamada de resumo ou sumário. Neste artigo, apresenta-se uma introdução à

sumarização automática de textos na Seção 2, aborda-se a teoria CST (*Cross-document Structure Theory*) na Seção 3, e, no final, descreve-se a proposta de pesquisa.

## 2. Sumarização automática de textos

A tarefa de sumarização automática de textos tem real importância no cenário atual em que a web se tornou um mundo caótico de informações e as pessoas têm pouco tempo de ler textos na íntegra. Com a facilidade que há de qualquer pessoa poder publicar na rede, o volume de informações cresce rapidamente, sendo que somente em 2009, foram produzidos 800 exabytes de informação, segundo estimativas da IDC (*Internacional Digital Center*).

Existem diversos tipos de sumários, como trailers de filmes, resumos de artigos científicos, tabelas de temporadas de jogos, revisão de livros, programação de eventos, catálogo de produtos, *abstracts* de teses, etc. No contexto de SA, um sumário pode ser um extrato ou *abstract*. Um extrato é um sumário formado simplesmente pela junção de pedaços inalterados do texto-fonte. No *abstract,* podem ocorrer algumas adaptações e reescritas no texto final.

As abordagens para SA podem ser **superficial** ou **profunda**. Na primeira, utilizam-se métodos estatísticos e/ou empíricos para se obter o sumário, levando em consideração pouco ou nenhum conhecimento lingüístico profundo. Na abordagem profunda, são utilizadas técnicas formais e modelos lingüísticos, que aumentam a sua complexidade de desenvolvimento, mas apresentam melhores resultados. A abordagem profunda leva em consideração regras gramaticais, semântica, conhecimento discursivo e de mundo. Quanto ao número de documentos analisados, a SA pode ser **monodocumento** ou **multidocumento**. A primeira produz um sumário a partir de um único texto-fonte que é dado como entrada para o sistema, e a multidocumento trabalha com diversos textos-fonte como jornais que tratam de um mesmo tópico. O foco deste artigo é a SA multidocumento.

Os trabalhos de SA seguem uma arquitetura genérica proposta por Mani e Maybury (1999), independente da abordagem, dividida nas etapas de análise, transformação e síntese. A Figura 1 ilustra essa arquitetura. O processo de **análise** consiste em extrair uma representação formal do conteúdo de um ou mais textos-fonte para ser processada automaticamente. A **transformação** deve gerar uma representação interna do sumário a partir da representação fornecida na etapa anterior e, desta forma, deve realizar a condensação do conteúdo de uma forma computável, mas ainda não textual. A **síntese** se responsabiliza por gerar em língua natural a representação interna condensada, produzindo o sumário propriamente dito.

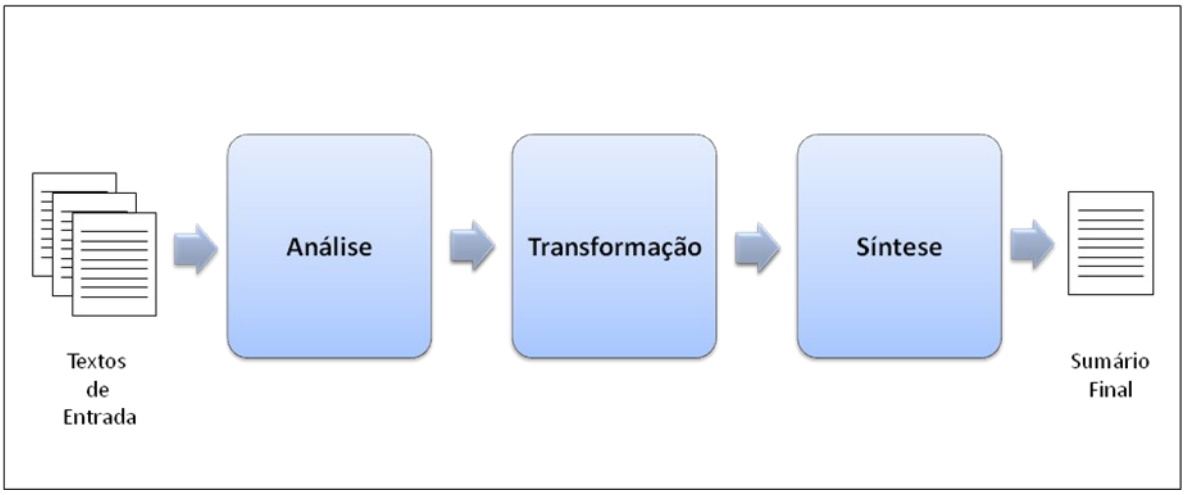


**Figura 1. Arquitetura genérica de um sumarizador automático de textos (Mani e Maybury, 2001)**

As iniciativas de sistemas de sumarização multidocumento que se destacam são SUMMONS (*SUMMarizing Online NewS articles*) de McKeown e Radev (1995) e MEAD de Radev et al. (2001). No Brasil, destaca-se o sistema GistSumm**,** uma ferramenta desenvolvida pelo Núcleo Interinstitucional de Lingüística Computacional (NILC[1]), para sumarização automática multidocumento em português do Brasil (Pardo, 2005). Este sumarizador faz uso de técnicas estatísticas simples para determinar a *Gist Sentence* (a sentença principal do texto) como guia para selecionar as demais sentenças do sumário.

Na sumarização multidocumento, vários são os desafios a serem tratados, como: evolução de eventos no tempo; narração de eventos com diferentes estilos, focos e perspectivas; informações redundantes, complementares ou contraditórias; ordenação das sentenças e tratamento da coerência e coesão. A seção seguinte aborda a teoria CST que pode ser aplicada para sumarização multidocumento e ainda ajudar na resolução desses problemas.

## 3. CST

Radev (2000) realizou uma análise criteriosa de artigos de notícias de diferentes fontes que tratavam sobre um mesmo tópico e identificou que, em alguns casos, eles apresentavam informações em comum, novas e contraditórias. Como resultado desse trabalho, surgiu a teoria CST (*Cross-document Structure Theory*), um dos modelos semântico-discursivos tradicionalmente usados para lidar com o conteúdo multidocumento. A CST é capaz de relacionar esse conteúdo e lidar com os fenômenos multidocumento, tais como redundância, contradição, informação temporal e outros.

Na sua versão original, a CST continha 24 relações que representavam as relações multidocumento, mas neste conjunto foram apontadas algumas ambigüidades por Zhang et al. (2002), que produziram um refinamento da teoria com 18 relações (Figura 2). Os nomes foram mantidos em inglês como na obra original.

[1] NILC: www.nilc.icmc.usp.br

| RELAÇÃO | DESCRIÇÃO |
|---|---|
| *Identity* | O mesmo texto aparece em mais de um local. |
| *Equivalence (Paraphrase)* | Duas sentenças possuem a mesma informação. |
| *Translation* | Mesma informação em línguas diferentes. |
| *Subsumption* | S1 contém toda informação em S2, mais informação adicional que não está em S2. |
| *Contradiction* | S1 e S2 apresentam informação conflitante. |
| *Historical Background* | S1 fornece contexto histórico da informação em S2. |
| *Citation* | S1 explicitamente cita o documento S2. |
| *Modality* | S1 apresenta uma versão modalizada da informação em S2, por exemplo, “é dito que; se sabe que”. |
| *Attribution* | S1 atribui a versão da informação em S2 usando, por exemplo, “de acordo com o Globo”. |
| *Summary* | S1 resume S2. |
| *Follow-up* | S1 apresenta informação adicional que tem acontecido desde S2. |
| *Indirect Speech* | S1 indiretamente menciona algo que foi diretamente mencionado em S2. |
| *Fulfillment* | S1 afirma a ocorrência de um evento previsto em S2. |
| *Elaboration (Refinement)* | S1 fornece detalhes de alguma informação em S2. |
| *Description* | S1 descreve uma entidade mencionada em S2. |
| *Reader Profile* | S1 e S2 fornecem a mesma informação, porém escrita para leitores diferentes. |
| *Change of Perspective* | A mesma entidade apresenta uma opinião diferente ou apresenta um fato por outro ângulo. |
| *Overlap (partial equivalence)* | S1 informa fatos X e Y, enquanto S2 informa fatos X e Z; Y e Z devem ser não-triviais. |

**Figura 2. Relações CST refinadas por Zhang et al. (2002)**

A Figura 3 exemplifica a aplicação de algumas das relações. Os fragmentos de texto são de diferentes fontes, tratam de um mesmo tópico e podem ser relacionados pelas relações *Contradiction* e *Attribution*. Observa-se que, no primeiro caso, há informações contraditórias, pois S1 diz que a colisão foi no 26º andar e S2 diz que foi no 25º andar, e no segundo caso, a relação *Attribution* se deve ao fato de que a fonte da informação em S1está sendo identificada (Aleixo e Pardo, 2008a).

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

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

**Figura 3. Exemplo de identificação de relações CST (Aleixo e Pardo, 2008b)**

Após o estudo das 18 relações CST refinadas por Zhang et al (2002), Aleixo e Pardo (2008b) encontraram relações redundantes e realizaram um novo refinamento, resultando em 14. Esta nova proposta está sendo utilizada nos trabalhos de sumarização para o português do Brasil. A Figura 4 apresenta o refinamento na forma de tipologia de relações (Jorge, 2010).

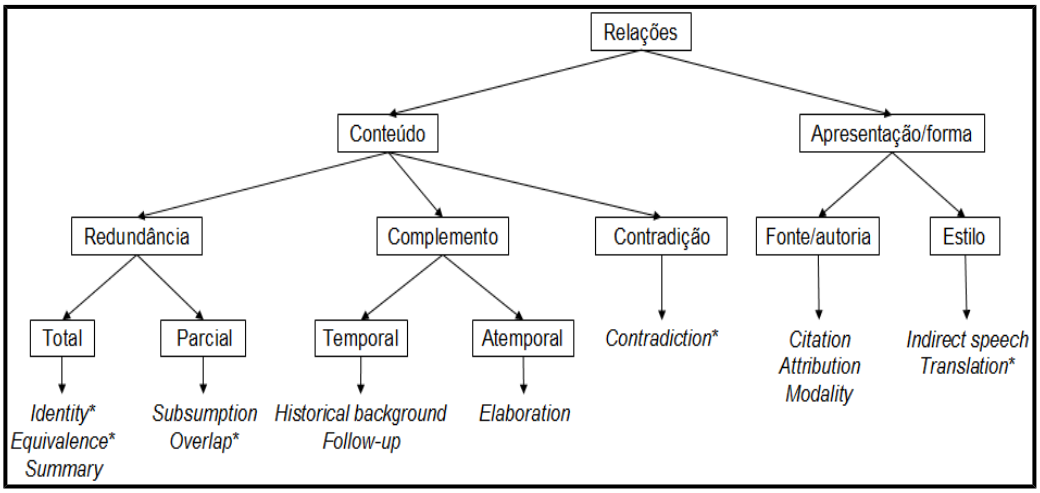


**Figura 4. Classificação de relações CST (Jorge, 2010)**

Já existem alguns recursos e ferramentas que utilizam a teoria CST, como o CSTBank (Radev et al., 2004), o CSTNews (Aleixo e Pardo, 2008b) e a CSTTool (Aleixo e Pardo, 2008a). O CSTBank é considerado um dos primeiros recursos desenvolvidos para pesquisas com CST, e trata-se de um corpus de documentos na língua inglesa composto por 6 coleções de textos jornalísticos, onde cada coleção tem em média 8 textos sobre o mesmo tópico. O CSTNews é composto por uma coleção de 50 textos jornalísticos de domínios variados e é visto como a primeira experiência de anotação para o português. Cada coleção possui em média 3 documentos de diferentes fontes que versam sobre o mesmo assunto, que foram retirados de sites de notícias. A partir deste último, foi desenvolvida a ferramenta CSTTool, projeto que faz parte do grupo NILC. CSTTool é um *parser* discursivo multidocumento semi-automático para textos escritos em português brasileiro.

A seção seguinte apresenta a proposta de pesquisa em sumarização de textos.

## 4. Proposta de pesquisa

Este artigo apresenta a proposta de pesquisa de investigar, desenvolver e avaliar métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo, como as teorias CST e RST (*Rhetorical Structure Theory) (*Mann e Thompson, 1987) e ontologias na produção de sumários mais informativos e coerentes.

Na linha de abordagem profunda de SA monodocumento, a teoria RST descreve que todo texto tem seus segmentos relacionados por relações do discurso, como por exemplo, causa-efeito, contraste e colaboração.

De acordo com Gruber (1995), uma ontologia é uma especificação formal e explicita de uma abstração, uma visão simplificada de um domínio de conhecimento. Uma ontologia modela um domínio de conhecimento, definindo um vocabulário comum. Já existem trabalhos que utilizam ontologias no processo de sumarização como por exemplo, Afantenos et al. (2004).

Tendo conhecimento de que a RST fornece o conhecimento necessário para lidar com conteúdo monodocumento e a CST com multidocumento, será investigado se o uso

das duas teorias em uma mesma aplicação poderá ou não contribuir na geração de sumários melhores. A proposta de pesquisa também trabalha com a hipótese de que ontologias podem fornecer o mecanismo necessário para que sejam realizadas inferências sobre as informações textuais e para que as informações sejam estruturadas e organizadas semanticamente e, portanto, contribuindo para a produção de sumários com mais qualidade.

As estratégias de sumarização serão aplicadas a textos em português de caráter jornalístico, por possuírem linguagem clara e do dia a dia, seguindo a linha atualmente em desenvolvimento no Brasil. Serão privilegiados domínios para os quais haja ontologias (ou recursos similares, como taxonomias) disponíveis. É importante ressaltar que há várias iniciativas de construção de ontologias para a língua portuguesa, como pode ser verificado no portal OntoLP (www.inf.pucrs.br/~ontolp/), que cataloga os recursos ontológicos disponíveis. Acredita-se, portanto, que não será necessária a construção de ontologias, devendo o trabalho ser focado na exploração de métodos de SA propriamente ditos.

## 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. e Pardo, T.A.S. (2008a). *CSTTool: Uma Ferramenta Semi-automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 321. São Carlos-SP, Maio, 14p.

Aleixo, P. e Pardo, T.A.S. (2008b) *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP, Junho, 15p.

Gruber, T. R. (1995) Toward principles for the design of ontologies used for knowledge sharing. In: *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers.

Jorge, M.L.C. (2010) Sumarização automática multidocumento: seleção de conteúdo com base no Modelo CST (Cross-document Structure Theory). Dissertação (Mestrado em Ciências de Computação e Matemática Computacional). Universidade de São Paulo, São Carlos.

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

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

Mann, W.C.; Thompson, S.A. (1987) *Rhetorical Structure Theory: Toward a functional theory of text organization*. Vol. 8, N. 3,243-281p.

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

Pardo, T.A.S. (2005). *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP, Fevereiro, 8p.

Radev, D. R., McKeown, K.R. (1998) Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, 4:469–500.

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*. Hong Kong.

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

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*.

Zhang, Z. S.; Blair-Goldensohn, S.; Radev, D. R. (2002) Towards CST-Enhanced Summarization. In the *Proceedings of AAAI 2002 Conference*. Edmonton, Alberta.