



Introdução à Sumarização Automática e Algumas Ferramentas de PLN

Thiago A. S. Pardo

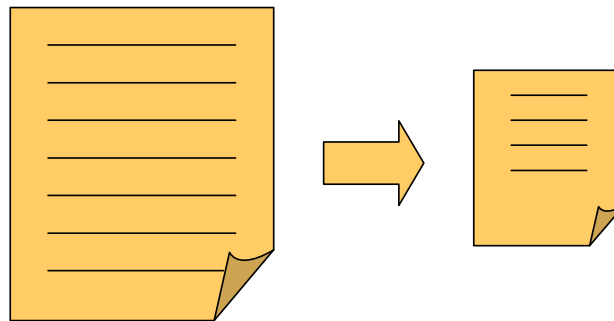
Núcleo Interinstitucional de Lingüística Computacional
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

I Escola Brasileira de Lingüística Computacional
2007



Sumarização

- Produção de uma versão mais curta de um texto-fonte: seu sumário



- Sumário, resumo
 - Extrato e *abstract*



Sumarização

- Permeia o dia a dia das pessoas
 - Sinopse de novelas
 - Resumo de notícias
 - Resenhas de livros e filmes
 - *Abstracts* de artigos científicos
 - Passagens de páginas da internet

[Forum-lp] **Linguateca - Escola de Verão**

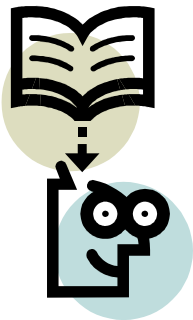
A **Linguateca** tem o prazer de anunciar a Primeira **Escola de Verão**, que terá lugar na Universidade do Porto, Portugal, de 10 a 14 de Julho de 2006. ...

<https://mail.di.fct.unl.pt/pipermail/forum-lp/2006-February/000092.html> - 4k -

[Em cache](#) - [Páginas Semelhantes](#)

Sumarização

- Motivações (humanas)
 - Acesso **rápido** à informação (*aboutness*)
 - **Auxílio à tomada de decisões**
 - Comprar um livro, alugar um filme, ler uma tese?
 - Acessar uma página da internet



Sumarização

- Motivações (humanas)

- Incapacidade de se absorver toda a informação disponível

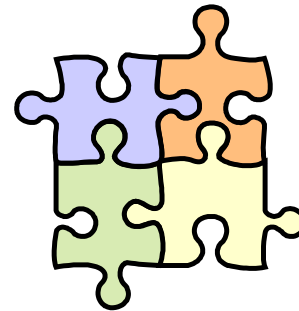
- Estudo de Berkeley (2003)

- 5 milhões de terabytes de nova informação (filme, meio magnético, impressa, *on-line*, etc.)
 - *Web*: 170 terabytes
- Dobro do produzido em 1999
- Aumento de 30% por ano



Sumarização

- Normalmente, resumos são textos
 - Devem apresentar as mesmas características que atribuem 'textualidade' (tessitura) a um texto
 - Coerência e coesão
 - Boa progressão temática
 - Gramaticalidade
 - Legibilidade
 - Etc.
 - Além de
 - Informação relevante



Sumarização: conceitos

- Taxa de compressão
 - O quanto “enxugar” o texto
 - Dependente da aplicação do sumário
- Informação
 - Essencial, complementar, supérflua
 - Dependente da audiência
- Idéia principal
 - Comunicada pelo escritor
 - Entendida pelo leitor





Sumarização: conceitos

- Tipos de sumário
 - Informativo (autocontido)
 - Indicativo (indexador)
 - Crítico (avaliativo)
- Modo de produção
 - Extratos
 - *Abstracts*

Sumarização

- Fatores que influenciam
 - Audiência: genérica ou especializada
 - Objetivo do sumário: substituir o texto-fonte, indexar, criticar
 - Fluência: textual ou fragmentado
 - Fonte: mono ou multidocumento



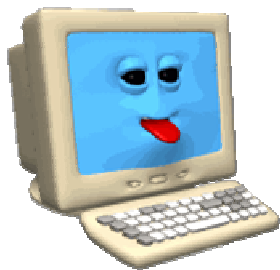


Sumarização

- Humana
 - Grande variedade de sumários para um mesmo texto
 - Processo **quase intuitivo**
- Computacional?
 - **Como** simular a habilidade humana?

Sumarização

- Motivações (lingüístico-computacionais)
 - Acesso somente à **informação relevante**
 - Recuperação de informação
 - Extração de informação
 - Categorização textual
 - Perguntas e respostas
 - Produção de **sumários úteis** aos humanos
 - **Desafio**: o computador deve 'entender' a língua
 - Envolve todas as questões mais complexas de Processamento de Língua Natural (PLN)
 - Interpretação textual
 - Geração textual
 - Avaliação



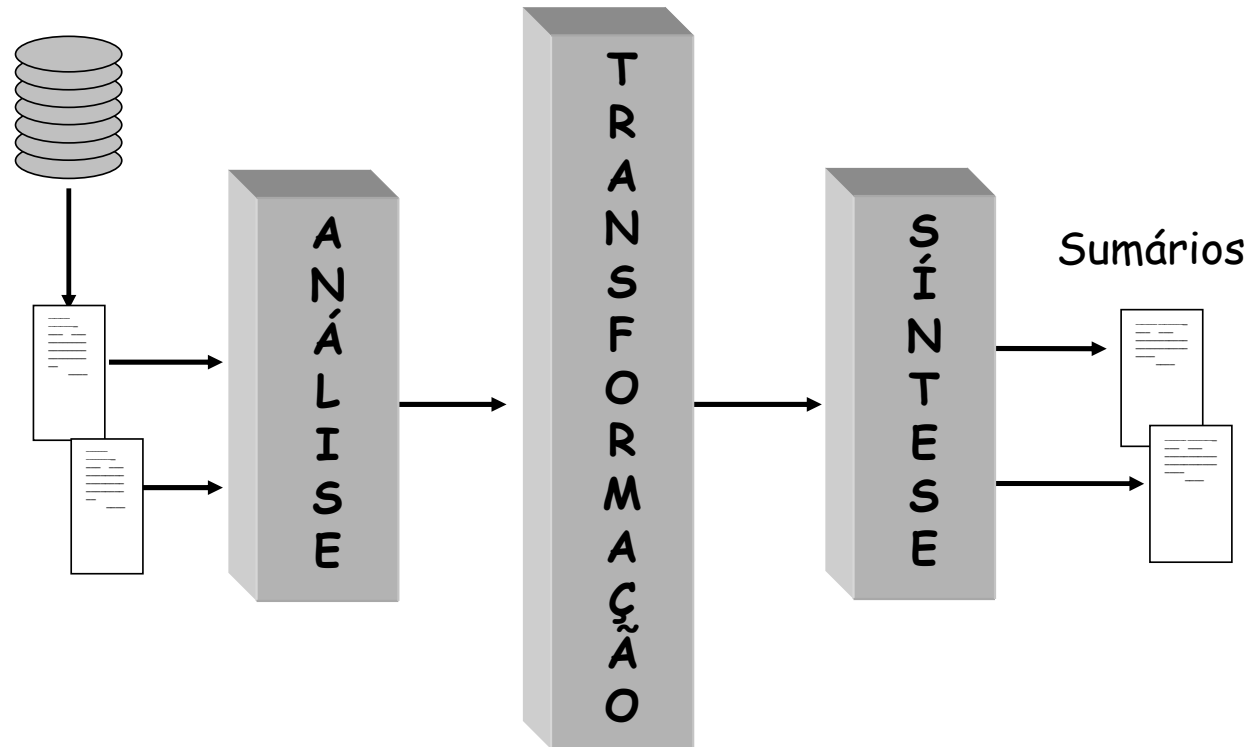


Sumarização Automática

- Financiamento massivo
 - EUA/DARPA, Comunidade Européia, *Pacific Rim*
 - Interesses governamentais e comerciais
- História
 - Primeiro sistema na década de 50
 - Acompanhou a história da IA: 'morte' e 'renascimento' da pesquisa
 - De extratos para *abstracts*
 - Hoje
 - Conferências dedicadas ao tema
 - Da sumarização para as aplicações

Sumarização Automática

Textos-fonte



Operações de sumarização

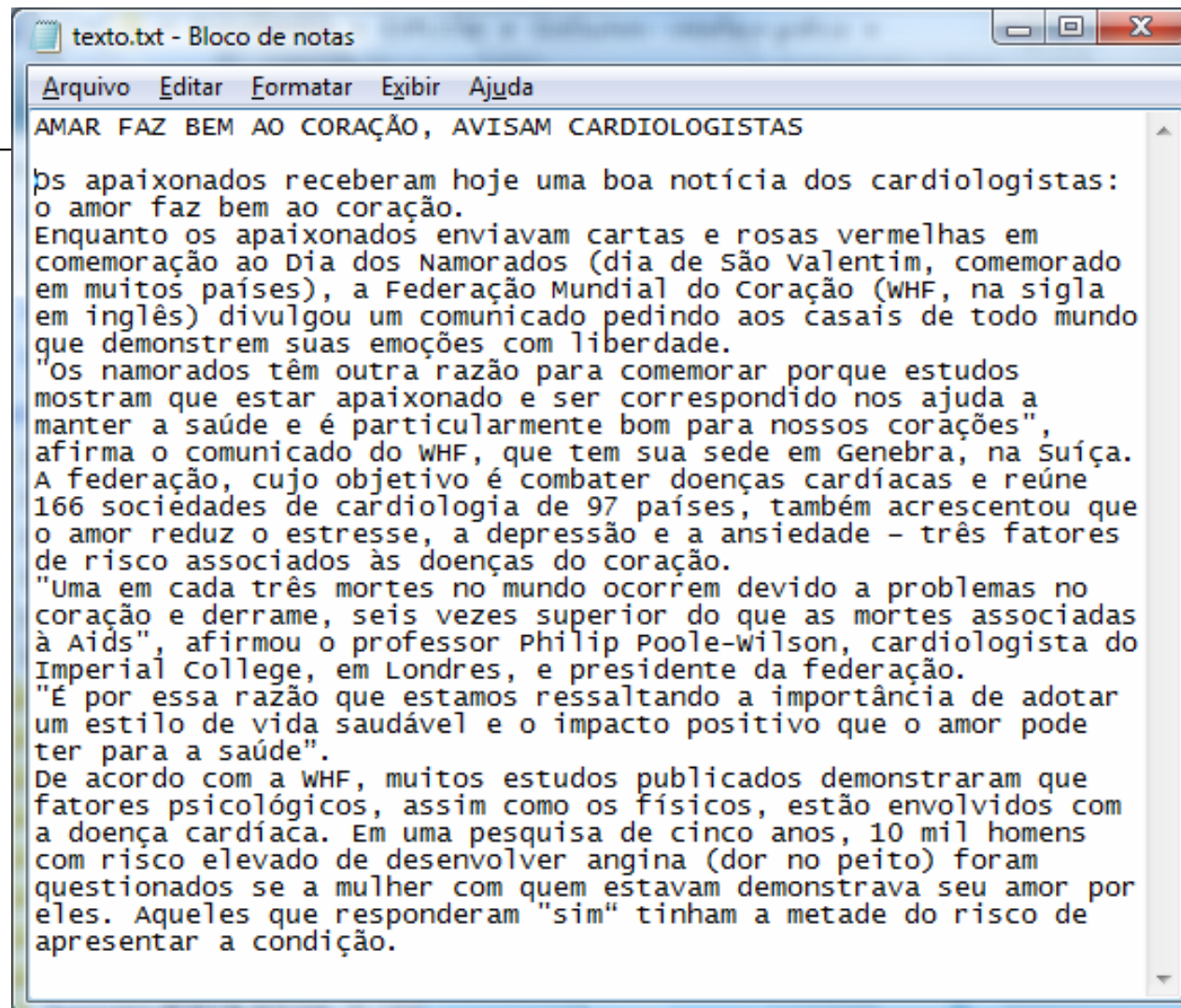


- Seleção/eliminação
 - Seleção do que é relevante ou exclusão do que é irrelevante

- Agregação
 - Associação (*merge*) de informações diversas

- Generalização/substituição
 - Substituição de informações específicas por informação mais geral

Exemplo: texto-fonte



texto.txt - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

AMAR FAZ BEM AO CORAÇÃO, AVISAM CARDIOLOGISTAS

ps apaixonados receberam hoje uma boa notícia dos cardiologistas: o amor faz bem ao coração.

Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês) divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.

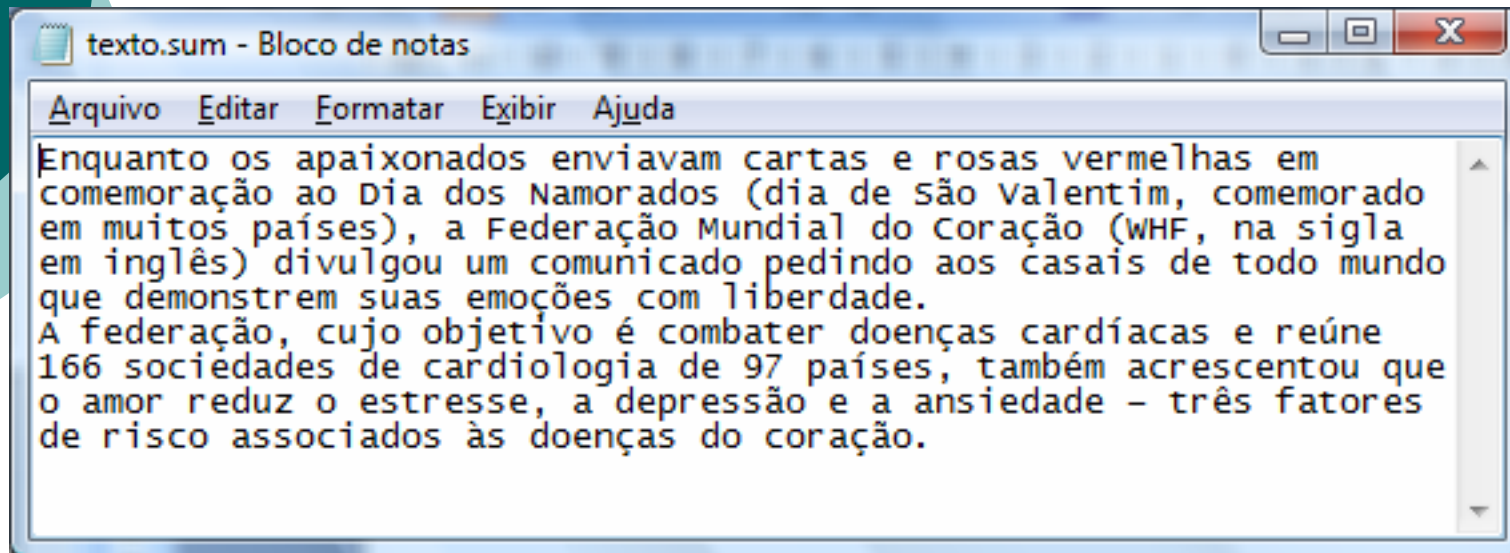
"Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça. A federação, cujo objetivo é combater doenças cardíacas e reúne 166 sociedades de cardiologia de 97 países, também acrescentou que o amor reduz o estresse, a depressão e a ansiedade - três fatores de risco associados às doenças do coração.

"Uma em cada três mortes no mundo ocorrem devido a problemas no coração e derrame, seis vezes superior do que as mortes associadas à Aids", afirmou o professor Philip Poole-wilson, cardiologista do Imperial College, em Londres, e presidente da federação.

"É por essa razão que estamos ressaltando a importância de adotar um estilo de vida saudável e o impacto positivo que o amor pode ter para a saúde".

De acordo com a WHF, muitos estudos publicados demonstraram que fatores psicológicos, assim como os físicos, estão envolvidos com a doença cardíaca. Em uma pesquisa de cinco anos, 10 mil homens com risco elevado de desenvolver angina (dor no peito) foram questionados se a mulher com quem estavam demonstrava seu amor por eles. Aqueles que responderam "sim" tinham a metade do risco de apresentar a condição.

Exemplo: sumário



Sumarização Automática

- Duas abordagens principais
 - **Superficial**: estatística, empírica
 - Processo menos complexo
 - Robustez
 - Resultados piores
 - **Profunda**: lingüística, fundamental
 - Processo mais complexo
 - Especificidade para alguns domínios
 - Resultados melhores
- Abordagens híbridas





Sumarização Automática

- Abordagens superficiais
 - Extratos
- Abordagens profundas
 - Podem produzir *abstracts*
 - Operações de sumarização

Abordagem superficial

- Método das palavras-chave

- Luhn, 1958; Edmundson, 1969; Black e Johnson, 1988
- O escritor do texto utiliza palavras-chave para expressar a idéia principal
 - As palavras-chave se repetem no decorrer do texto
- Seleção de sentenças que contêm palavras-chave para compor o extrato



Abordagem superficial

- Método da localização
 - Baxendale (1958)
 - Sentenças importantes ocorrem em lugares mais proeminentes do texto
 - Início e fim de parágrafo/texto



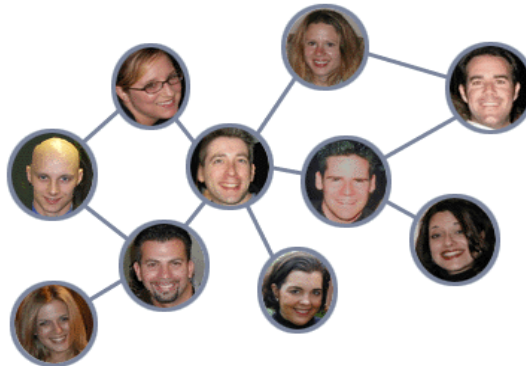
Abordagem superficial

- Método das palavras e frases indicativas
 - Paice (1981)
 - Seleção de sentenças cujo conteúdo é sinalizado como relevante por palavras e frases indicativas
 - Artigo científico: “o objetivo deste trabalho...”
 - Esporte: “resultado”, “placar”



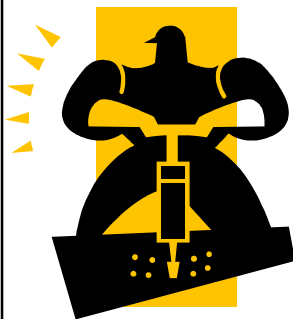
Abordagem superficial

- Método relacional
 - Skorochocko (1971)
 - As sentenças mais importantes são aquelas altamente relacionadas às outras (co-ocorrência de palavras/conceitos)



Abordagem superficial

- Mineração de textos
 - Larocca Neto et al. (2000)
 - TF-ISF (*Term Frequency – Inverse Sentence Frequency*)
 - Quanto mais representativas as palavras de uma sentença, mais importante ela é no texto



Abordagem superficial

- Método da idéia principal
 - Pardo et al. (2003)
 - Há uma sentença identificável no texto que expressa sua idéia principal
 - O sumário é construído a partir desta sentença





Abordagem profunda

- Conhecimento lingüístico e extralingüístico
 - Regras de interpretação e geração textual
 - Modelos de língua
 - *Wordnets*
 - Gramáticas
 - Discurso
 - Identificação do que é relevante no contexto
 - Diversas teorias discursivas

Discurso

- Um texto é mais do que uma simples seqüência de sentenças justapostas
 - Estrutura altamente elaborada
 - Coerência/sentido

"Choveu. O chão está molhado."

Causa-efeito

"Embora tenha chovido, as obras continuaram."

Oposição

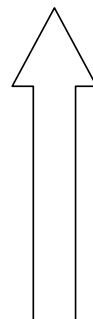
"O menino voltou da escola, fez seus deveres e foi dormir."

Seqüência

Discurso

- Níveis de conhecimento em PLN

Abstração &
complexidade



Pragmática/**Discurso**
Semântica
Sintaxe
Morfologia
Fonética/Fonologia

- Envolve a situação de comunicação (Koch e Travaglia, 2002): escritor e leitor



Teorias discursivas

- Grosz e Sidner (1986): intenções
- Mann e Thompson (1987): retórica
- Jordan (1992) e Kehler (2002): semântica
- Moore e Pollack (1992), Moore e Paris (1993), Korelsky e Kittredge (1993), Moser e Moore (1996), Rino (1996) e Marcu (1999, 2000), entre outros: mapeamentos entre os níveis do discurso



Rhetorical Structure Theory – RST

Mann e Thompson, 1987

- Retórica: parte “palpável” da pragmática (Hovy, 1988)
- Meio pelo qual um texto é organizado para satisfazer um objetivo comunicativo
 - Intenção
- Organização funcional do texto
 - Função de suas partes para o sucesso da comunicação



Rhetorical Structure Theory – RST

Mann e Thompson, 1987

- Estrutura hierárquica do texto
 - Relações retóricas entre proposições (unidades de conteúdo) expressas no texto
 - Em geral, proposições simples são expressas por orações
 - Núcleos e satélites
 - Relações mononucleares e multinucleares
 - Relações intencionais e informativas
 - Intencionais: alteram a inclinação do leitor para algo
 - Informativas: informam o leitor sobre algo



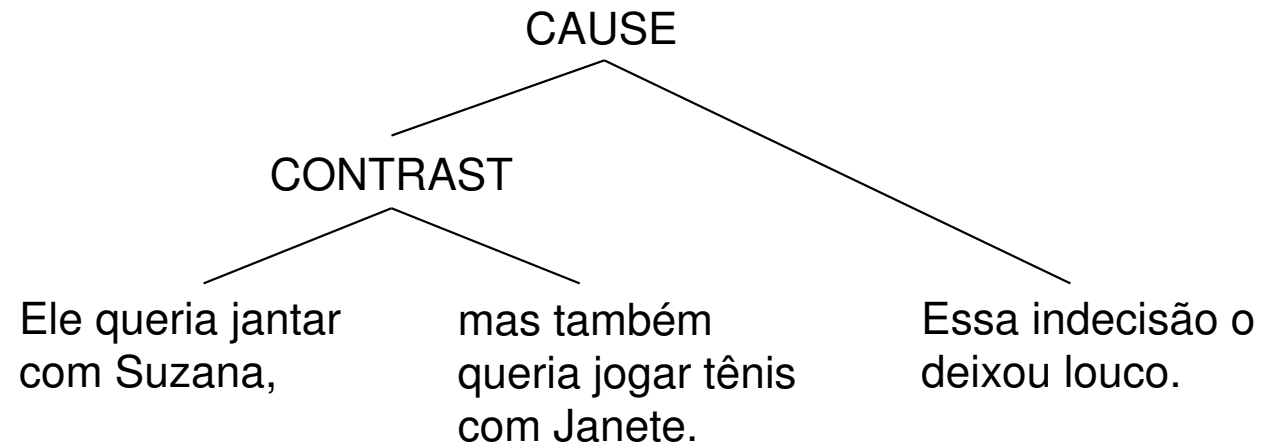
RST: exemplo

Ele queria jantar
com Suzana,

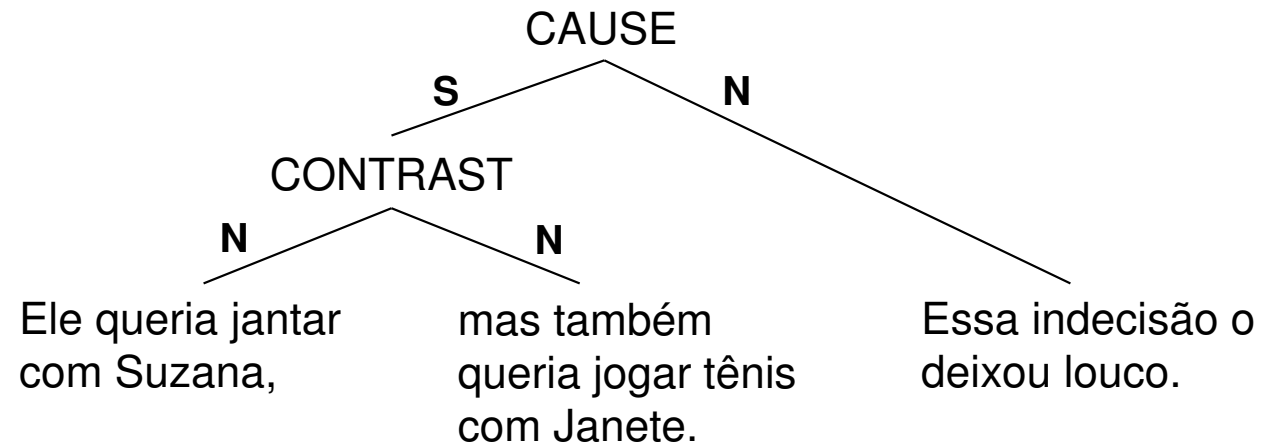
mas também
queria jogar tênis
com Janete.

Essa indecisão o
deixou louco.

RST: exemplo



RST: exemplo





Analísadores discursivos automáticos

- Inglês
 - Marcu (1997, 2000)
 - Corston-Oliver (1998)
 - Schilder (2002)
 - Marcu e Echihabi (2002)
 - Soricut e Marcu (2003)
 - Reitter (2003)
 - Hanneforth et al. (2003)
 - Mahmud e Ramsay (2005)

- Japonês
 - Sumita et al. (1992)

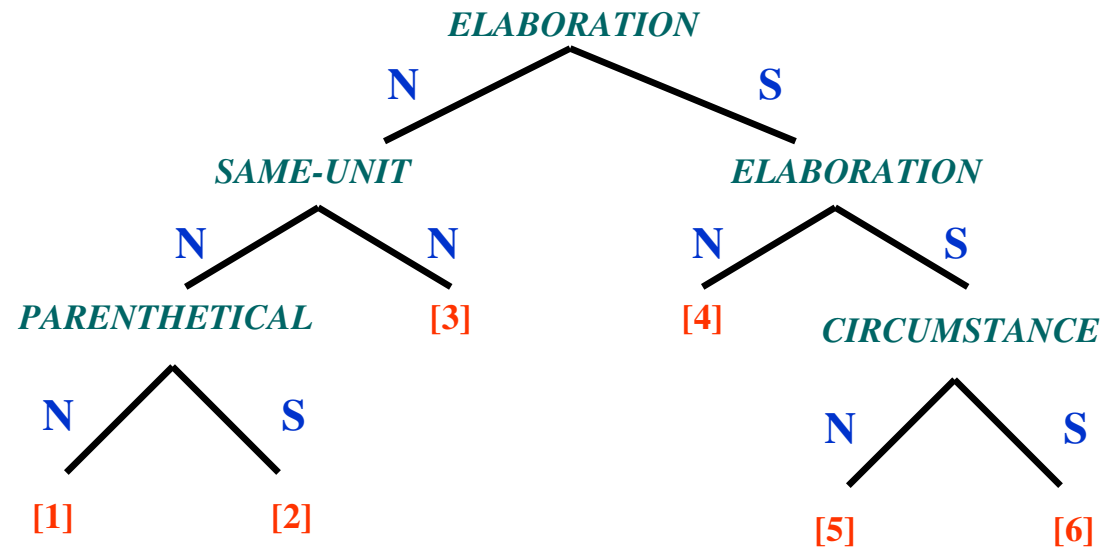
- Português
 - DiZer (Pardo, 2005)



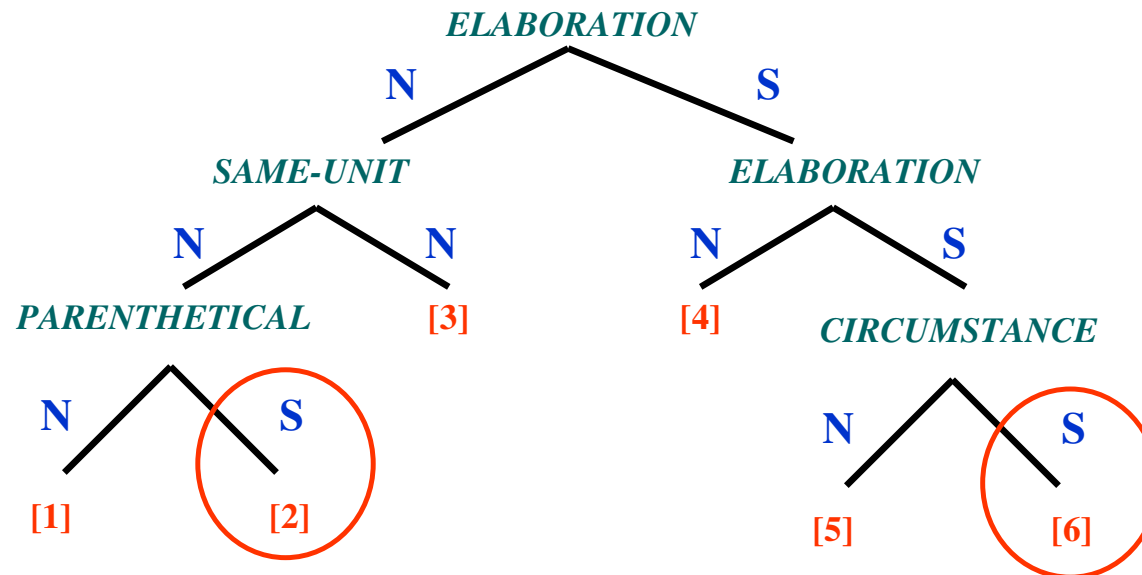
Abordagem profunda

- Idéia básica para sumarização
 - Satélites são informação complementar e, portanto, podem ser eliminados
 - Vários métodos para se escolher que segmentos eliminar
 - Mann e Thompson (1992), Rino (1996), O'Donnel (1997), Marcu (2000)

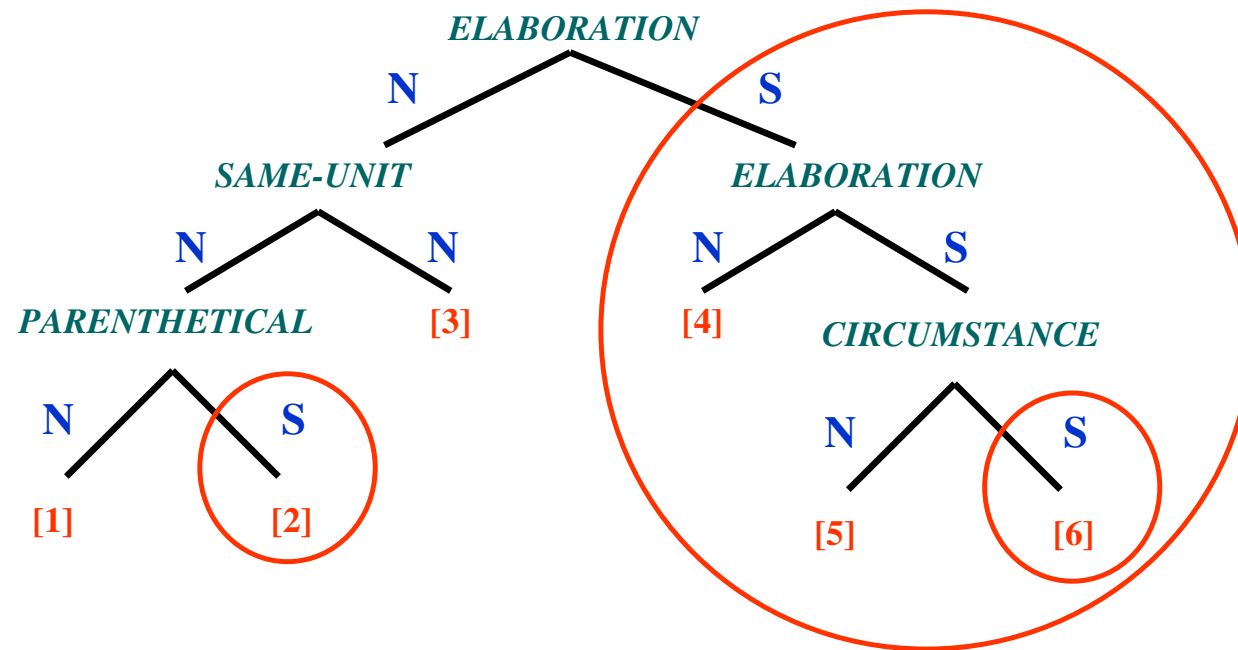
[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem [2] (na região metropolitana de Belo Horizonte), [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente. [4] Os ganhos são atribuídos pela diretoria da fábrica à nova filosofia [5] que vem sendo implantada na empresa desde outubro do ano passado, [6] quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.



[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem [2] (na 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. [4] Os ganhos são atribuídos pela diretoria da fábrica à nova filosofia [5] que vem sendo implantada na empresa desde outubro do ano passado, [6] quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.



[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem [2] (na 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. [4] Os ganhos são atribuídos pela diretoria da fábrica à nova filosofia [5] que vem sendo implantada na empresa desde outubro do ano passado, [6] quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.





Abordagem profunda

- Mann e Thompson (1992)
 - Eliminação de satélites que não são necessários para que as relações retóricas em foco atinjam seus efeitos pretendidos



Abordagem profunda

- O'Donnel (1997)
 - Cada segmento (núcleo e satélite) tem sua importância determinada em função da profundidade na árvore retórica e da relação a qual pertence



Abordagem profunda

- Marcu (2000)
 - A saliência (profundidade na árvore) de um segmento determina sua importância
 - Quanto mais nuclear, mais importante



Abordagem profunda

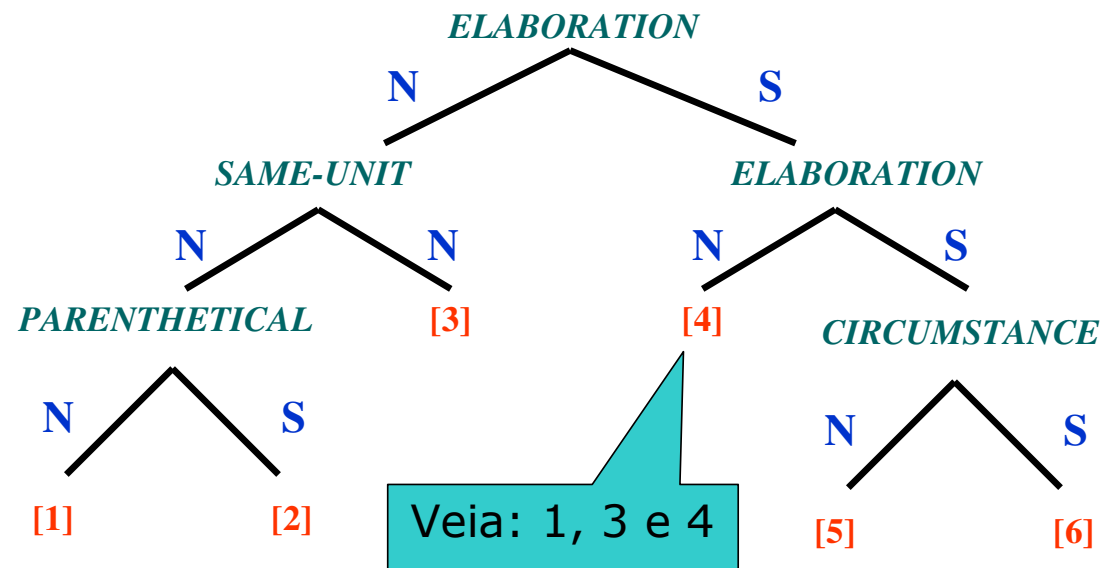
- Rino (1996)
 - A nuclearidade não é suficiente; é necessário considerar o objetivo comunicativo original do texto



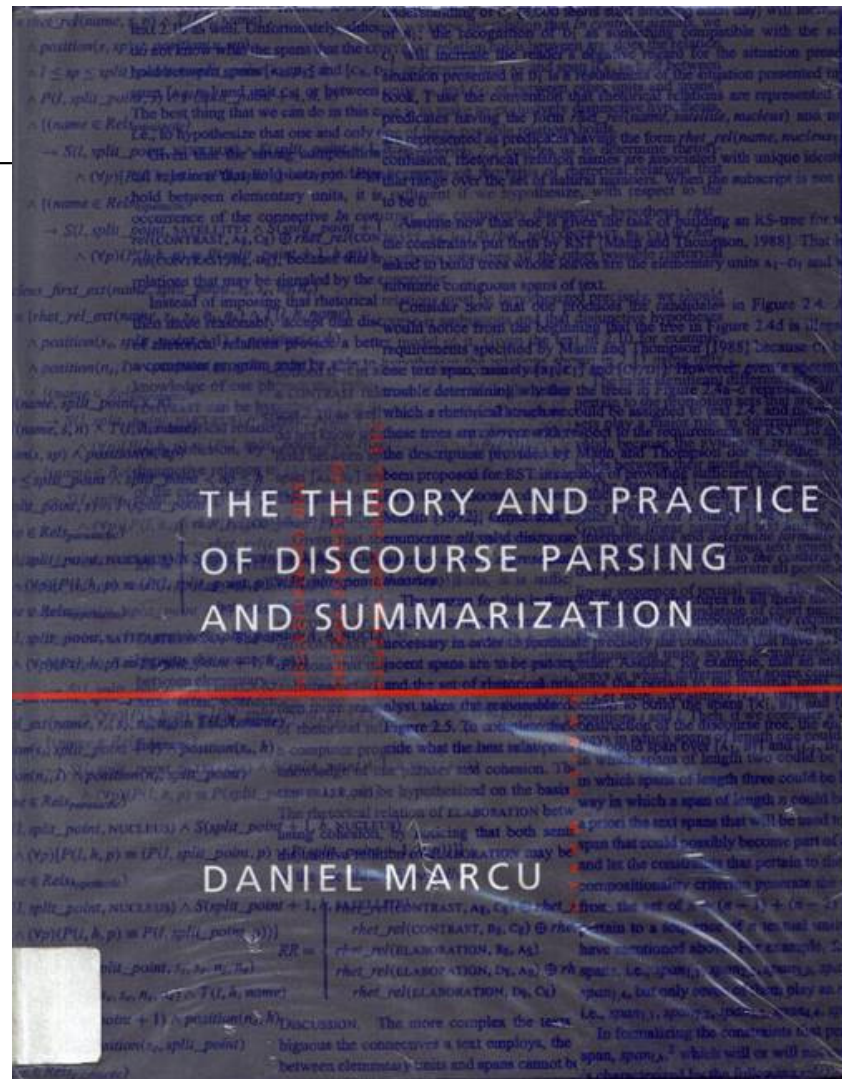
Abordagem profunda

- Teoria das veias (Cristea et al., 1998)
 - Cada segmento da estrutura discursiva contém uma lista de segmentos que possivelmente contêm os antecedentes das anáforas que ocorrem
 - Ao se selecionar um segmento para inclusão no sumário, garante-se a inclusão dos segmentos anteriores que possam conter os antecedentes anafóricos

[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem [2] (na 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. [4] Os ganhos são atribuídos pela diretoria da fábrica à nova filosofia [5] que vem sendo implantada na empresa desde outubro do ano passado, [6] quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.



Discurso e sumarização



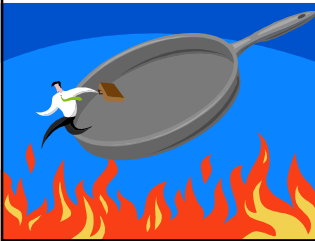
Abordagem superficial/profunda

- Aprendizado de máquina
 - Kupiec et al. (1995), Teufel e Moens (1997)
 - Combinação de características sentenciais para julgamento de relevância para compor o sumário
 - Tamanho, posição, número de substantivos, nuclearidade, etc.



Sumarização multidocumento

- Geração de um único sumário para um conjunto de textos sobre um mesmo assunto
 - Diversos novos problemas
 - Ordenação temporal dos eventos
 - Tratamento de redundância
 - Manutenção da coerência e coesão





Sumarização multidocumento

- CST (*Cross-document Structure Theory*) (Radev, 2000)
 - Baseada na RST
 - Relaciona os segmentos de diversos textos

Avaliação: como decidir o que é melhor?

- Cenário

- Diversos métodos
 - Superficiais, profundos e híbridos
 - Variedade de fontes de conhecimento

- Diversos tipos de sumários
 - Extratos e *abstracts*
 - Genérico ou especializado
 - Textual ou fragmentado
 - Informativo, indicativo ou crítico
 - Mono e multidocumento
 - Taxa de compressão



- Muitos sumários bons para um mesmo texto

Quesitos avaliáveis



- Desempenho computacional
 - Complexidade do algoritmo, uso de memória, etc.

- Usabilidade
 - Interface, consistência, flexibilidade, etc.

- **Resultados**
 - Qualidade



Forma de avaliação

- **Intrínseca**

- Qualidade do resultado
 - Quão bom é o sumário?

- **Extrínseca**

- Aplicação em um contexto
 - O quanto o uso de sumários melhorou a recuperação de informação?



Julgamento humano

- *On-line*
 - Humanos treinados
 - Tempo, dinheiro
 - Questões derivadas da subjetividade
 - Boa descrição da tarefa, concordância
- **Off-line**
 - Reproduzível, rápida e barata



O que se avalia

- *Glass-box*
 - Módulos do sistema
 - Crítica mais elaborada

- **Black-box**
 - Resultado final do sistema
 - O que realmente importa!



Comparação de resultados

- **Avaliação comparativa**
 - Grandes eventos internacionais
 - SUMMAC, DUC
 - *Roadmaps*
- Avaliação autônoma

Como moldar a avaliação?

- Sparck Jones e Galliers (1996)
 - Tão importante quanto a forma de avaliação é saber o que se quer avaliar





Medidas de avaliação intrínseca

- Dois principais aspectos (Mani, 2001)
 - Qualidade textual
 - Informatividade do sumário
 - Em relação a um corpus de textos com sumários humanos, em geral



Medidas de avaliação intrínseca

- Qualidade

- Julgamento humano, normalmente
 - Fluência, facilidade de leitura, clareza, legibilidade, concisão
 - Referências anafóricas, explicação para siglas e abreviaturas
 - Integridade das estruturas presentes no texto (listas e tabelas)
 - Coerência e coesão
 - Ortografia e gramática

Medidas de avaliação intrínseca

- Informatividade
 - Cobertura e precisão
 - Informação em comum entre o sumário automático e um sumário ideal
 - Sumário ideal (*gold standard*): humano

$$C = \frac{\textit{senten\c{c}as Ideal} \cap \textit{senten\c{c}as Autom\c{a}tico}}{\textit{senten\c{c}as Ideal}}$$

$$P = \frac{\textit{senten\c{c}as Ideal} \cap \textit{senten\c{c}as Autom\c{a}tico}}{\textit{senten\c{c}as Autom\c{a}tico}}$$





Medidas de avaliação intrínseca

- Medida de utilidade (Radev et al., 2000): extratos
 - Cada sentença do texto-fonte é pontuada (por humanos) de acordo com sua importância
 - O sumário recebe uma nota que corresponde à soma das notas de suas sentenças
- Sobreposição de conteúdo
 - Similar à precisão e cobertura, mas considera proposições simples (unidades factuais de informação)
- *Retention rate* (Mani, 2001)
- Preservação da ideia principal



Medidas de avaliação intrínseca

- ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003)
 - <http://www.isi.edu/~cyl/ROUGE/>
 - Automática, com julgamento próximo do humano
 - Co-ocorrência de n-gramas entre sumário automático e sumário(s) de referência
 - Seqüências de palavras: 1 a 4
 - Ceticismo: “medidas automáticas podem ser enganadas”
 - Explosão das pesquisas em avaliação em sumarização



Avaliação extrínseca

- Sumários em contexto
 - Categorização de textos
 - Perguntas e respostas
 - Recuperação de informação
- Os sumários podem não ser bons para o ser humano, mas podem ser bons para a máquina
 - Às vezes, textualidade não é necessária



Avaliação extrínseca

- Categorização de textos (Mani et al., 1998)
 - Atribuir uma classe aos textos: economia, informática, política, etc.
 - Em vez do humano/computador processar o texto todo, processa somente o sumário
 - Taxa de acerto deve aumentar
 - Menos informação irrelevante
 - Tempo demandado deve diminuir
 - Menos informação para processar



Avaliação extrínseca

- Perguntas e respostas (Morris et al., 1992; Hovy e Lin (2000))
 - Preparam-se perguntas para um grupo de textos
 - Humanos respondem as perguntas
 - Sem ler nada
 - Lendo os sumários
 - Lendo os textos



Avaliação extrínseca

- Recuperação de informação (Mani et al., 1998; Tombros e Sanderson, 1998; Jing et al., 1998)
 - Duas abordagens
 - **Indexação de sumários** em vez dos textos completos
 - Aumento da taxa de acerto na recuperação
 - Somente informação relevante é indexada
 - **Apresentação de sumários** junto com os resultados da busca
 - Aumento da satisfação do usuário

Futuro da avaliação

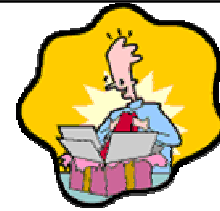
- Avaliação
 - Extrínseca
 - *Off-line*
 - *Black-box*
 - Comparativa
 - Multidocumento
- DUC 2005
 - Difícil superar o método *baseline*
 - “Métodos cada vez mais complicados para se selecionar a primeira sentença dos textos”
 - Estagnação da área?
 - Mudança de paradigma



Futuro da avaliação

- Avaliações conjuntas
 - Avaliam o estado da arte
 - Ditam direções de pesquisa





“Ferramentas” de PLN

- Além de sumarizadores...
 - Etiquetadores morfossintáticos (*taggers*)
 - Analísadores sintáticos (*parsers*)
 - Analísadores semânticos e discursivos
 - Corretores gramaticais
 - Alinhadores textuais
 - Bases de dados lexicais
 - Etc.

Introdução à Sumarização Automática e Algumas Ferramentas de PLN



www.nilc.icmc.usp.br
taspardo@icmc.usp.br