DISCURSO – PARTE 3

SCC5908 Tópicos em Processamento de Língua Natural

Thiago A. S. Pardo

MUDANÇA DE PERSPECTIVA

- Web e explosão de informação
- Necessidade de lidar com grande quantidade de textos/documentos
- Aplicações dedicadas e "mais inteligentes"
  - WolframAlpha, Qwiki
  - Google News, Google Trends
- Modelos surgem para modelar esse mundo

2

Desafios Multidocumento

- O cenário multidocumento é uma "selva" a ser desbravada
  - Evolução de eventos no tempo
  - Narração dos eventos com diversos estilos, perspectivas diferenciadas e em momentos variados
  - Diferentes focos sobre uma mesma informação central
  - Expressões referenciais diferentes
  - Informação redundante
  - Informações complementares
  - Informações contraditórias
    - Evolução de um evento, com relatos parciais ou em momentos diferentes
    - Erros
    - Discordâncias e perspectivas diferentes
  - Ordenação das informações
  - Coerência e coesão
  - Multilinguismo
  - Fofocas, boatos, etc.

3

Desafios Multidocumento

- Mídias novas, formatos diferenciados, fatores até então inexistentes
  - Blogs, microblogs e redes sociais
  - Linguagens próprias
  - Links
  - Ranque de páginas e busca orgânica
    - Estratégias para ganhar importância
  - Número de amigos, botão "curtir", etc.

4

Cross-document Structure Theory

(Radev, 2000)

Um pouco de história

- Trigg e o sistema TextNet (1983, 1986)
  - Argumentação em textos científicos
- RST (Mann e Thompson, 1987)
  - Retórica - monodocumento
- Radev e Mckeown (1995): SUMMONS e seus operadores
  - Relações entre *templates*
- Radev (2000): CST

6

Um pouco de história

- Afantenos et al. (2004, 2008) e críticas a CST
  - Relações muito genéricas, granularidades diferentes
  - Proposta de relações sincrônicas (mesmo tempo) e diacrônicas (mesma fonte) para domínio específico
- Etoh e Okumura (2005) e o refinamento da CST para o japonês
- *RTE Challenge* (Giampiccolo et al. 2007)
  - Algumas relações: *entailment*, *contradiction*, etc.
- Maziero et al. (2010) e a tipologia de relações
- Murakami et al. (2009, 2010) e relações entre opiniões e fatos para o japonês
  - Relações: *agreement*, *confinement*, *conflict* e *evidence*

7

CST

- Teoria discursiva multidocumento
- 24 relações para documentos que versam sobre um mesmo assunto
- Motivada por questões de aplicações
  - Sumarização, principalmente

8

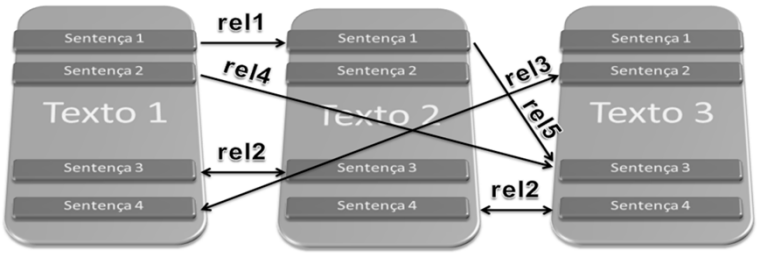
CST

- Radev (2003), durante construção do CSTBank
  - *Since it describes relationships that hold across multiple documents rather than across spans of text within the same document, it makes no assumptions about authors' intentions in creating cohesion in texts*

9

CST

- Modelo semântico-discursivo de estruturação multidocumento
  - São definidas relações entre partes (de quaisquer granularidades) dos documentos/textos



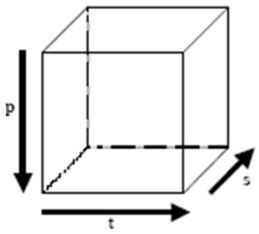
10

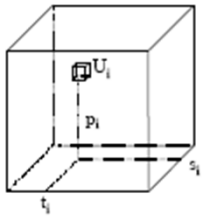
CST

- Estruturas de dados complementares
  - **Cubo multidocumento**: fonte (*source*), tempo (*time*) e posição (*position*) dos segmentos textuais
  - **Grafo multidocumento**: relações multidocumento

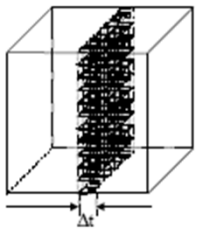
11

CST: cubo multidocumento





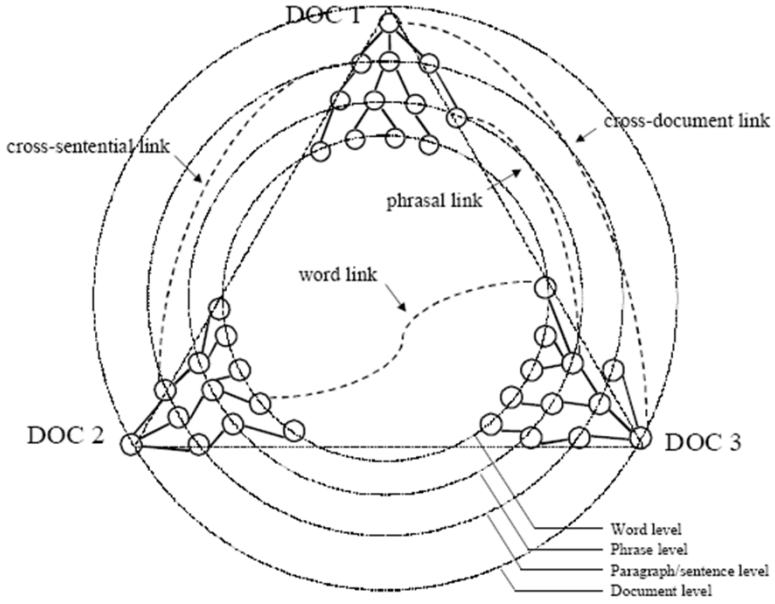
Unidade de documento U_i, sendo que um documento é uma seqüência $U_1 ... U_n$ projetada nos eixos fonte e tempo



Recorte no tempo

12

CST: grafo multidocumento



13

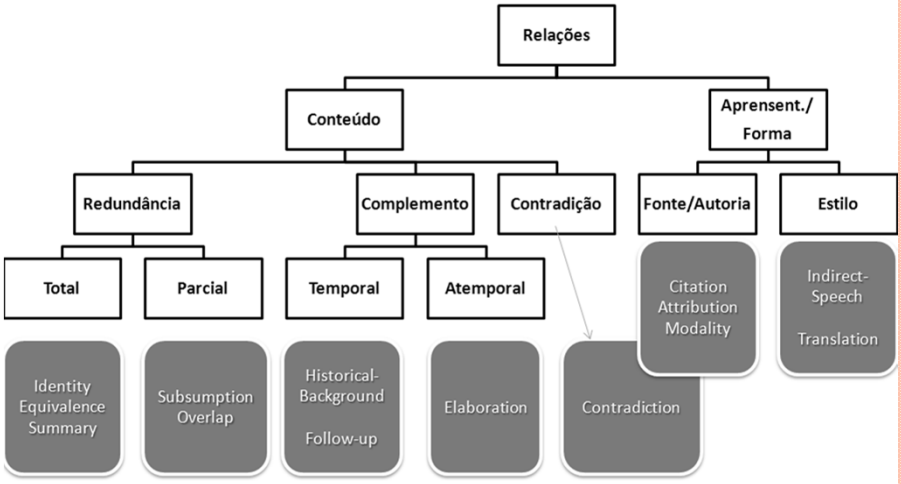
CST

- Relações originais

Identity	Modality	Judgment
Equivalence	Attribution	Fulfillment
Translation	Summary	Description
Subsumption	Follow-up	Reader profile
Contradiction	Elaboration	Contrast
Historical background	Indirect speech	Parallel
Cross-reference	Refinement	Generalization
Citation	Agreement	Change of perspective

TIPOLOGIA PARA O PORTUGUÊS

- Menos relações, refinadas e organizadas



Exemplo

- Quais as relações?

D1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 13 pessoas na quinta-feira à tarde, informou nesta sexta feira um porta-voz das Nações Unidas.

D2: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. Segundo um porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. O Congo tem um histórico de queda de mais de 30 aviões.

Exemplo

- *Contradiction*, *overlap*, *historical background* (←)

D1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 13 pessoas na quinta-feira à tarde, informou nesta sexta feira um porta-voz das Nações Unidas.

D2: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. Segundo um porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. O Congo tem um histórico de queda de mais de 30 aviões.

EXERCÍCIO

- Parte 1 – estudo de um conjunto de relações

18

- Parte 2 – análise do par de textos

O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.
Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador --Cahe descartou pancreatite ou úlcera.
"Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".
Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.
Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.

BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana.
Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.
"Agora está estável. Mesmo com esta melhora, ele continuará internado", disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão).
Cahe reforçou que Maradona ainda tem problemas.
"Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave", afirma, em entrevista ao diário La Nación.
No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão.
Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino.
Sua filha, Dalma, foi ao estádio assistir ao jogo.

- Parte 2 – análise do par de textos

O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.
Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador --Cahe descartou pancreatite ou úlcera.
"Maradona teve uma recaída na h
aguda. Agora está estável. Apesa
melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".
Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.
Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.

Algum problema?

BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana.
Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.
"Agora está estável. Mesmo com esta melhora, ele continuará internado", disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão).
Cahe reforçou que Maradona ainda tem
áticos dele na avaliação não
estão equilibrados e ele não está bem. Mas não é nada grave", afirma, em entrevista ao diário La Nación.
No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão.
Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino.
Sua filha, Dalma, foi ao estádio assistir ao jogo.

ANÁLISE VIA CST

- Ser humano é capaz de relacionar tudo com tudo, praticamente
- Restrição sobre a CST
  - Relações ocorrem entre segmentos com alguma similaridade lexical (Radev et al., 2004)
    - Uso de medidas como *word overlap*, cosseno, etc.
- E a análise de 3 textos ao mesmo tempo? E de 4? É um problema?

21

CST

- Para o inglês
  - Córpus CSTBank
    - Poucos textos
    - Baixa concordância entre anotadores
  - Parser discursivo limitado

22

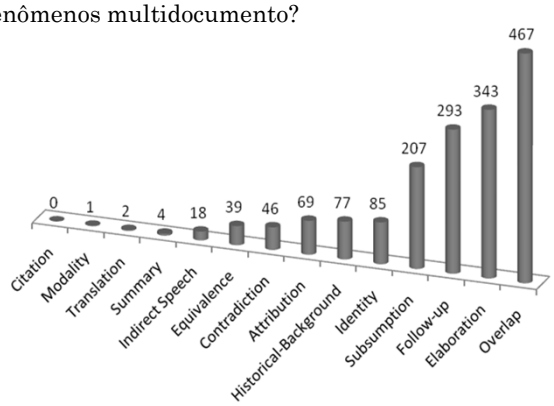
CST

- Para o português
  - Córpus CSTNews
    - 50 grupos de textos jornalísticos, 140 textos no total
    - Boa concordância entre anotadores
  - Ferramenta semi-automática de anotação de textos
    - CSTTool
  - Parser discursivo amplo

23

CST

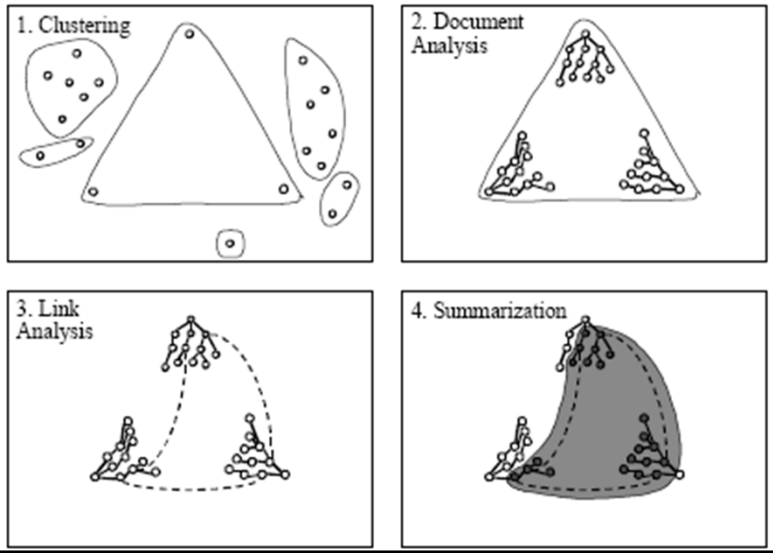
- Para o português
  - CSTNews e suas relações
    - Fenômenos multidocumento?



24

SUMARIZAÇÃO: UMA APLICAÇÃO

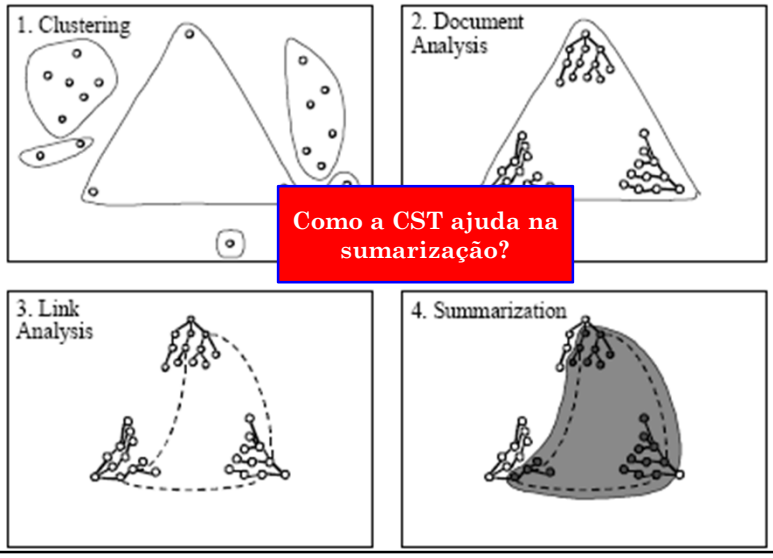
- Vários trabalhos, método padrão



25

SUMARIZAÇÃO: UMA APLICAÇÃO

- Vários trabalhos, método padrão



Como a CST ajuda na sumarização?

26

CST e Modelos Multidocumento

- Questão
  - E o relacionamento entre textos de assuntos diferentes?

27