Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts

Priscila Aleixo

Thiago A. S. Pardo

Cenário

- Pesquisa do IDC (*International Digital Center*): 281 exabytes (281 bilhões de gigabytes) em 2007, superando em mais de 10% as previsões
  - Informação redundante, contraditória e complementar

Cenário

- Humanos não conseguem processar tudo

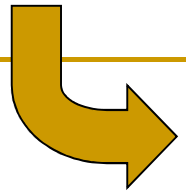
- Aplicações e ferramentas de PLN podem ajudar a processar textos, mas precisam saber lidar com o problema

Análise discursiva multidocumento

- Relaciona segmentos textuais de diversas fontes
  - Causa-efeito, contradição, equivalência semântica, identidade, tradução, citação, etc.

(S1) A colisão 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.



contradição, elaboração

Análise discursiva multidocumento

- Algumas aplicações
  - Sumarização automática multidocumento
    (Barzilay et al., 1999; Zhang et al., 2002; Afantenos et al., 2004)
  - Perguntas e respostas
    (Otterbacher, 2006)
  - Tarefas de interpretação e geração textual em geral
    (Radev, 2000)

Análise discursiva multidocumento

- *Cross-document Structure Theory* (CST)
  (Radev, 2000)

  - Teoria discursiva multidocumento mais difundida

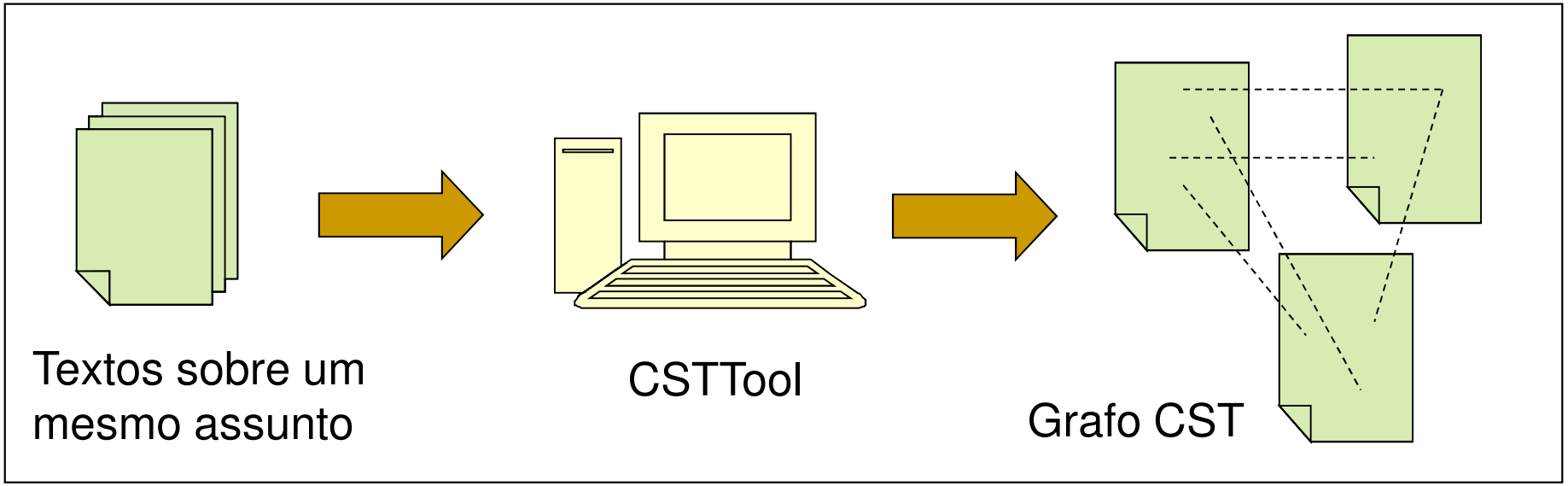
  - Baseada na tradicional RST (Mann e Thompson, 1987)
    - Quebrou várias "regras sagradas" da RST
    - ...mas herdou problemas
    - ...e criou alguns novos

  - Apesar de ser uma das poucas propostas "formalizadas", é alvo de críticas
    - Não é discurso, heterogeneidade dos segmentos, baseada na RST

Proposta

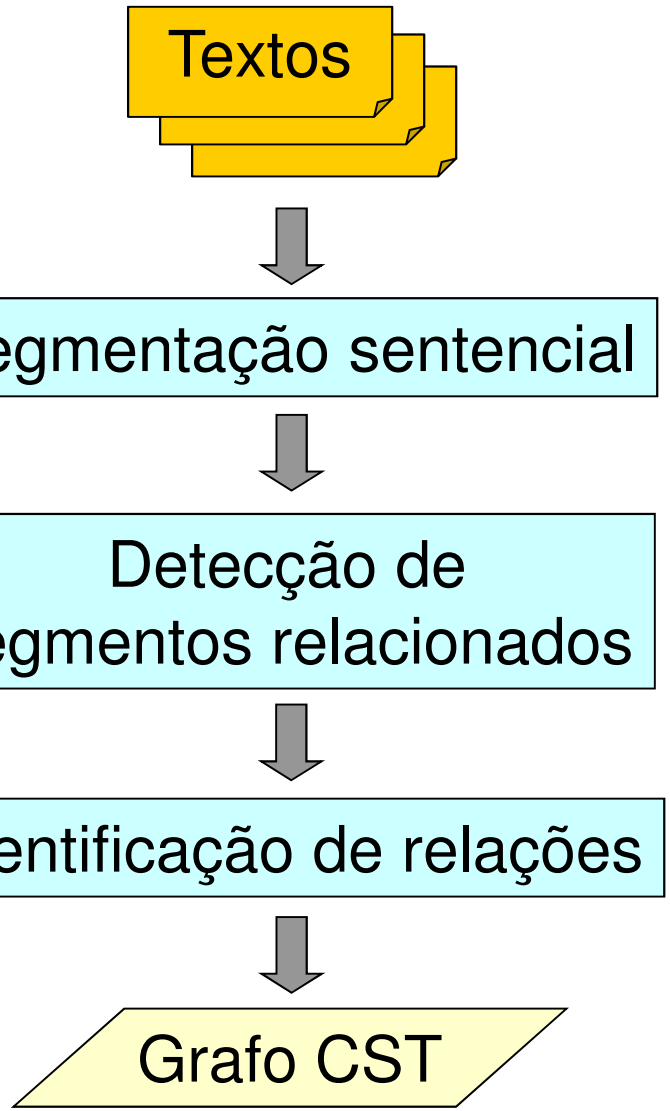
- Validação da teoria para o português
  - Redundância, contradição, complementaridade, ordenação temporal de eventos

- Construção de um analisador discursivo multidocumento automático para o português



CSTTool

- 3 etapas
  (Zhang et al., 2003)



CSTNews

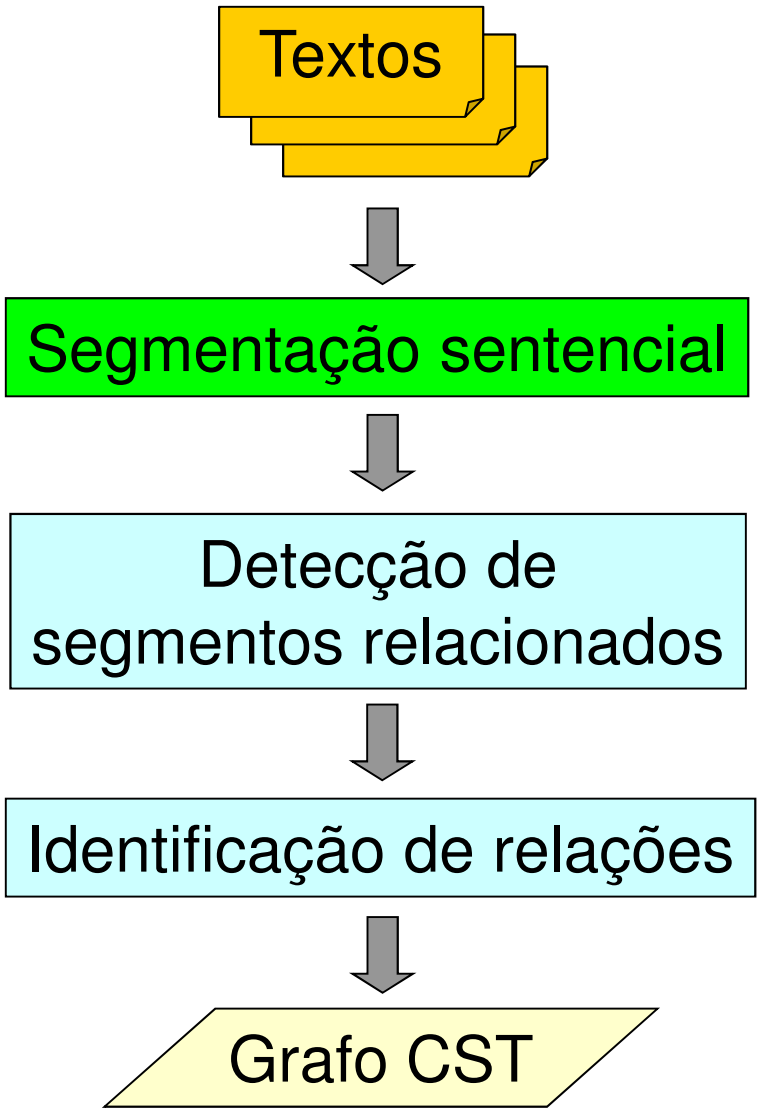
- Construção de um córpus
  - 50 grupos de textos de diferentes domínios
    - Média de 3 textos por grupo
  - Diversas fontes: jornais on-line
    - Linguagem do dia a dia

CSTNews

- Anotação nos moldes do CSTBank (Radev et al., 2004)
  - XML
  - Anotação por 2 especialistas treinados
    - Ambiente de edição amigável
    - 14 relações CST
    - Kappa=0.26, na média
  - Indicação automática de segmentos para relacionar
    - Humano consegue ver sentido em tudo
    - Radev (2000): relações CST acontecem entre segmentos com alta similaridade lexical
      - *Word overlap*

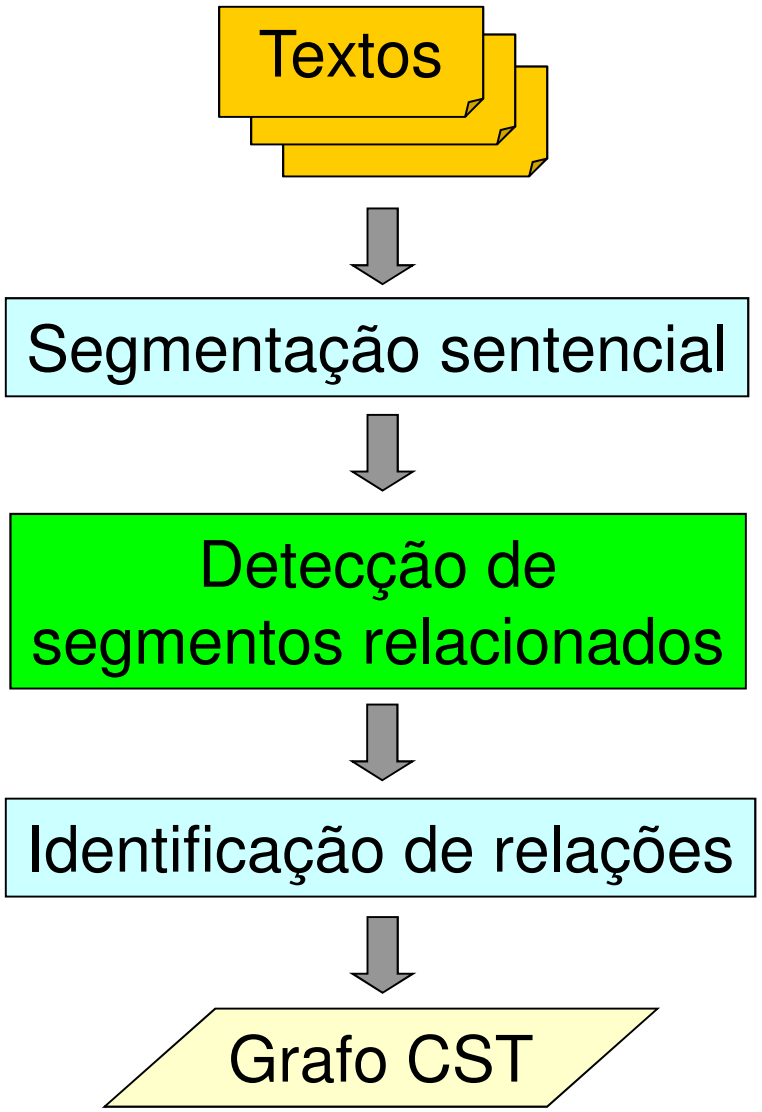
CSTTool

- Segmentação
  - Sentença é o nível de análise por excelência
  - SENTER (Pardo, 2006)



CSTTool

- Detecção de segmentos relacionados
  - Teste das melhores medidas da literatura
    - *Word overlap*
    - Cosseno
  - Adição de recursos lingüísticos
    - Stoplist
    - Wordnet.Br
    - Lematizador



CSTTool

- Seleção aleatória de alguns grupos de textos do córpus
- Teste das medidas na detecção dos segmentos relacionados
  - Cobertura sobre precisão
- Combinação exaustiva das medidas com os recursos lingüísticos (individuais e juntos)

CSTTool

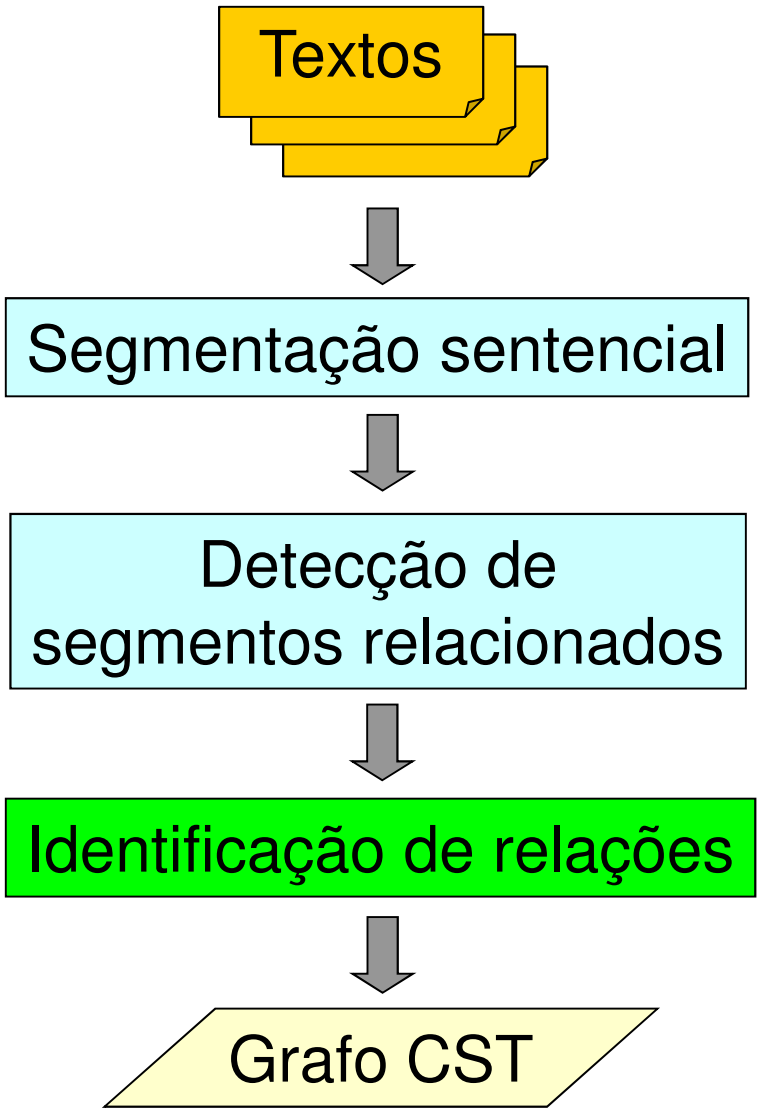
- Resultados recorrentes
  - Stoplist aumenta precisão, mas penaliza cobertura
  - Lematização e sinonímia não alteram a medida *word overlap*
  - Com medida do cosseno, lematização aumenta precisão, mas não penaliza cobertura
  - Melhor medida: cosseno com lematização, corte entre 0.1 e 0.2
    - Cobertura entre 93 e 100%

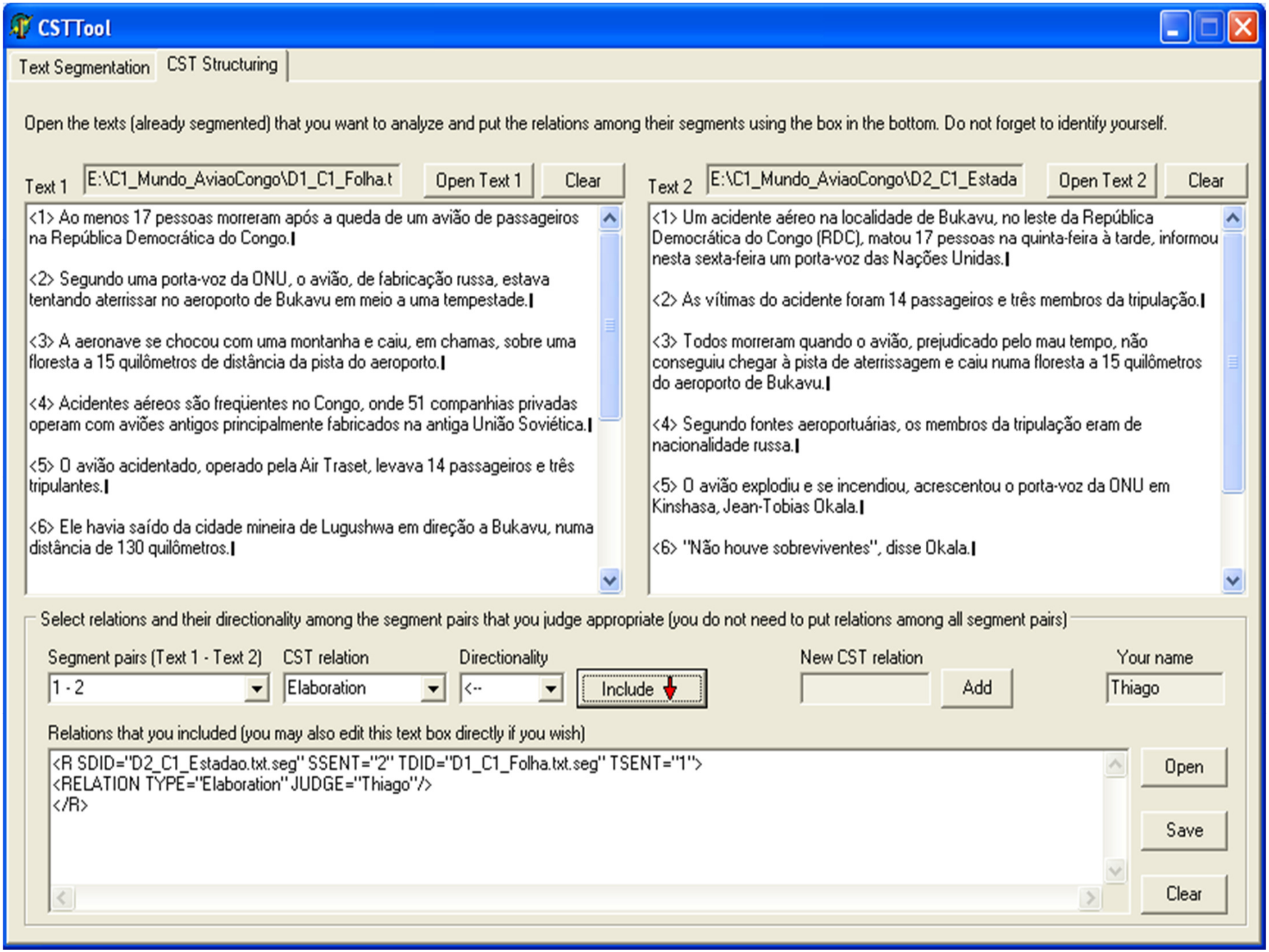
CSTTool

- Precisão geral
  - Segmentos relacionados e não relacionados
    - Novamente, medida do cosseno com lematização
      - Precisão de 89%

CSTTool

- Trabalho futuro
  - Técnicas de aprendizado de máquina




CSTTool
Text Segmentation
CST Structuring
Open the texts (already segmented) that you want to analyze and put the relations among their segments using the box in the bottom. Do not forget to identify yourself.
Text 1
E:\C1_Mundo_AviaoCongo\D1_C1_Folha.t
Open Text 1
Clear
<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.|
<4> Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.|
<5> O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.|
<6> Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros.|
Text 2
E:\C1_Mundo_AviaoCongo\D2_C1_Estada
Open Text 2
Clear
<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 três 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.|
<4> Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.|
<5> O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.|
<6> "Não houve sobreviventes", disse Okala.|
Select relations and their directionality among the segment pairs that you judge appropriate (you do not need to put relations among all segment pairs)
Segment pairs (Text 1 - Text 2)
1 - 2
CST relation
Elaboration
Directionality
<--
Include
New CST relation
Add
Your name
Thiago
Relations that you included (you may also edit this text box directly if you wish)
<R SDID="D2_C1_Estadao.txt.seg" SSENT="2" TDID="D1_C1_Folha.txt.seg" TSENT="1">
<RELATION TYPE="Elaboration" JUDGE="Thiago"/>
</R>
Open
Save
Clear

Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts

- www.nilc.icmc.usp.br