Sumarização: alinhamento do córpus CSTNews

Renata Tironi de Camargo (PPGL/NILC)
Verônica Agostini (ICMC/NILC)

Orientadores: Ariani Di Felippo
Thiago A. S. Pardo

1

Roteiro

- Alinhamento
  - O que é, onde é usado, tipos
- Propostas de mestrado
- Alinhamento do CSTNews
  - Processo, manual, tipos, exemplos, problemas

2

Roteiro

- Alinhamento
  - O que é, onde é usado, tipos
- Propostas de mestrado
- Alinhamento do CSTNews
  - Processo, manual, tipos, exemplos, problemas

3

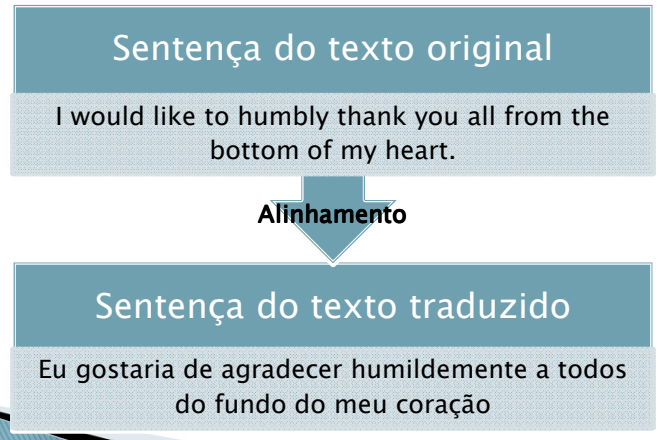
Alinhamento

- Relacionar duas unidades textuais com informação em comum
  - Palavra, sentença, documento...
- Tradução automática
  - Texto e sua tradução
- Sumarização automática
  - Texto(s) e seu sumário
- Outras áreas

4

Exemplos de alinhamento

- Tradução automática, sentencial, 1-1



5

Exemplos de alinhamento

- Sumarização automática, sentencial, 1-n

6

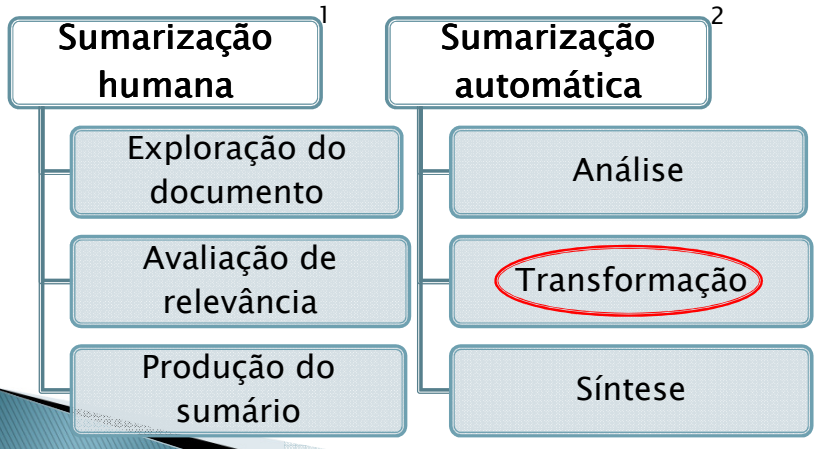
Roteiro

- Alinhamento
  - O que é, onde é usado, tipos
- Propostas de mestrado
- Alinhamento do CSTNews
  - Processo, manual, tipos, exemplos, problemas

7

Proposta de mestrado – Renata

Etapas da Sumarização



8

Proposta de mestrado – Renata

Hipóteses

- Há estratégias recorrentes de seleção de conteúdo na sumarização humana multidocumento
- Há correlação entre tais estratégias e a modelagem CST[3]
- Há estratégias de "recorta e cola"[4] que caracterizam a produção de sumários multidocumento

9

Proposta de mestrado – Renata

Objetivos da pesquisa

a) Investigação e sistematização de estratégias de seleção de conteúdo na SHM em português brasileiro (PB)

b) Investigação da correlação entre as estratégias de seleção e a modelagem CST

c) Investigação das estratégias de "recorta e cola"

d) Formalizar o conhecimento sistematizado em (a), (b) e (c) por meio de regras explícitas que auxiliem a SAM

10

Proposta de mestrado – Renata

Etapas Metodológicas

1. **Alinhamento** dos textos-fonte a seu sumário humano
2. **Caracterização** das sentenças dos textos-fonte quanto às estratégias superficiais de seleção de conteúdo
3. **Caracterização** das sentenças dos textos-fonte quanto às relações CST (abordagem profunda)
4. **Identificação** de estratégias de seleção na SHM
5. **Formalização** e **avaliação** das estratégias de seleção de conteúdo
6. **Caracterização** (e formalização?) dos sumários em função das operações "recorta e cola"

11

Exemplo

- Caracterização automática quanto às estratégias superficiais (em andamento)

Sentenca	Localizacao	Tamanho	Tamanho-Norm	Top-Words	Top-Words-Norm	Palavras-do-Titulo	Palavras-do-Titulo-Norm	Frequencia	Frequencia-Norm
S1_D1_C1	COMECO	9	0.473684	5	0.714286	5	0.416667	40	0.634921
S2_D1_C1	MEIO	13	0.684211	4	0.571429	1	0.0833333	43	0.68254
S3_D1_C1	MEIO	11	0.578947	1	0.142857	0	0	25	0.396825
S4_D1_C1	MEIO	16	0.842105	2	0.285714	1	0.0833333	32	0.507937
S5_D1_C1	MEIO	11	0.578947	2	0.285714	1	0.0833333	30	0.47619
S6_D1_C1	MEIO	11	0.578947	2	0.285714	0	0	24	0.380952
S7_D1_C1	MEIO	13	0.684211	4	0.571429	1	0.0833333	33	0.52381
S8_D1_C1	MEIO	10	0.526316	2	0.285714	1	0.0833333	24	0.380952
S9_D1_C1	MEIO	2	0.105263	0	0	0	0	2	0.031746
S10_D1_C1	FIM	16	0.842105	1	0.142857	1	0.0833333	27	0.428571
S1_D2_C1	COMECO	19	1	7	1	2	0.166667	63	1
S2_D2_C1	MEIO	8	0.421053	3	0.428571	0	0	26	0.412698
S3_D2_C1	MEIO	18	0.947368	3	0.428571	12	1	56	0.888889
S4_D2_C1	MEIO	8	0.421053	1	0.142857	0	0	22	0.349206
S5_D2_C1	MEIO	9	0.473684	2	0.285714	1	0.0833333	33	0.52381
S6_D2_C1	MEIO	5	0.263158	1	0.142857	0	0	14	0.222222
S7_D2_C1	FIM	16	0.842105	5	0.714286	2	0.166667	57	0.904762
S1_D3_C1	COMECO	18	0.947368	7	1	3	0.25	62	0.984127
S2_D3_C1	MEIO	8	0.421053	3	0.428571	0	0	26	0.412698
S3_D3_C1	MEIO	18	0.947368	3	0.428571	3	0.25	54	0.857143
S4_D3_C1	MEIO	9	0.473684	2	0.285714	1	0.0833333	33	0.52381
S5_D3_C1	MEIO	5	0.263158	1	0.142857	0	0	14	0.222222
S6_D3_C1	MEIO	16	0.842105	5	0.714286	2	0.166667	57	0.904762

Exemplo

- Caracterização quanto às relações CST (em andamento)

Redundancia	Redundancia-Norm	Complemento	Complemento-Norm	Contradicao	Contradicao-Norm	Forma	Forma-Norm
2	0.666667	4	0.571429	0	0	1	1
2	0.666667	7	1	0	0	1	1
1	0.333333	0	0	1	1	0	0
0	0	1	0.142857	0	0	0	0
2	0.666667	3	0.428571	1	1	0	0
0	0	1	0.142857	0	0	0	0
0	0	1	0.142857	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
2	0.666667	4	0.571429	0	0	0	0
2	0.666667	1	0.142857	0	0	0	0
3	1	3	0.428571	0	0	1	1
1	0.333333	2	0.285714	0	0	0	0
1	0.333333	3	0.428571	0	0	0	0
1	0.333333	1	0.142857	0	0	0	0
1	0.333333	5	0.714286	1	1	0	0
2	0.666667	2	0.285714	0	0	1	1
2	0.666667	1	0.142857	0	0	0	0
1	0.333333	1	0.142857	1	1	0	0
1	0.333333	5	0.714286	0	0	0	0
1	0.333333	1	0.142857	0	0	0	0
2	0.666667	3	0.428571	0	0	0	0

13

Proposta de mestrado – Renata

Próximas Etapas

1. **Caracterização** das sentenças dos textos-fonte com relação a **outros atributos**
   - Número de substantivos, verbos, advérbios, adjetivos
   - Discurso direto (sim/não)
   - RST[5] – nuclearidade (sim/não)
   - RST – relação (elaboration, cause, contrast...)
2. **Identificação** de estratégias de seleção na SHM
3. **Formalização** e **avaliação** das estratégias de seleção de conteúdo
4. **Caracterização** (e formalização?) dos sumários em função das operações "recorta e cola"

14

Proposta de mestrado – Verônica

- Objetivos
  - Investigação e exploração de técnicas de alinhamento sentencial (sumários- textos fonte) com variados níveis de conhecimento linguístico
    - Produção de um alinhador automático (para textos jornalísticos, a princípio)

15

Proposta de mestrado – Verônica

- Hipóteses
  - Alinhamentos multidocumento refletem, em certa medida, os fenômenos multidocumento
    - Contradição, redundância, etc.
  - Métodos com mais conhecimento linguístico trarão melhores resultados
  - Alinhamento em nível sentencial é suficiente para se obter bons resultados
    - No mínimo próximos aos da literatura
  - Terão mais alinhamentos do tipo 1-N do que 1-1

16

Proposta de mestrado – Verônica

- Lacunas
  - Falta de estudos significativos da sumarização multidocumento
  - Sumarização multidocumento ainda traz resultados insatisfatórios

17

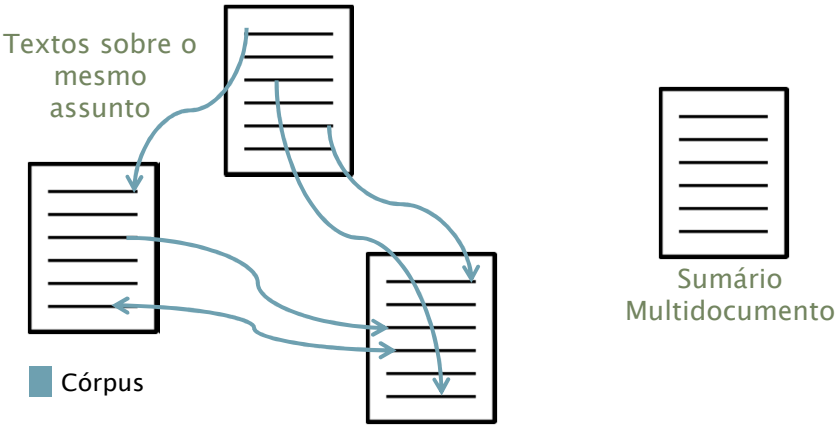
Proposta de mestrado – Verônica

- Três abordagens propostas
  - Métodos baseline
    - Palavras iguais, posição sentenças
  - Relações CST
    - Guiar o alinhamento com as relações
  - Aprendizado de máquina
    - Atributos: quantidade de palavras iguais, quantidade de relações CST, tipo das relações CST, quantidade de classes morfossintáticas iguais, posição relativa nos textos

18

Proposta de mestrado – Verônica

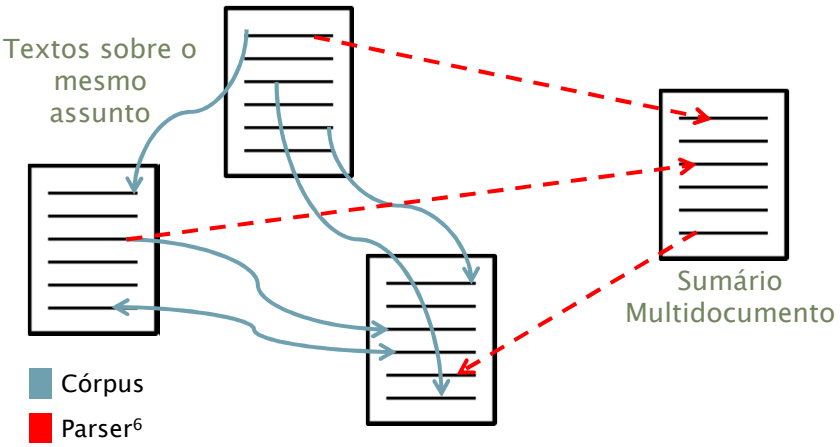
- Abordagem 2 – Relações CST



19

Proposta de mestrado – Verônica

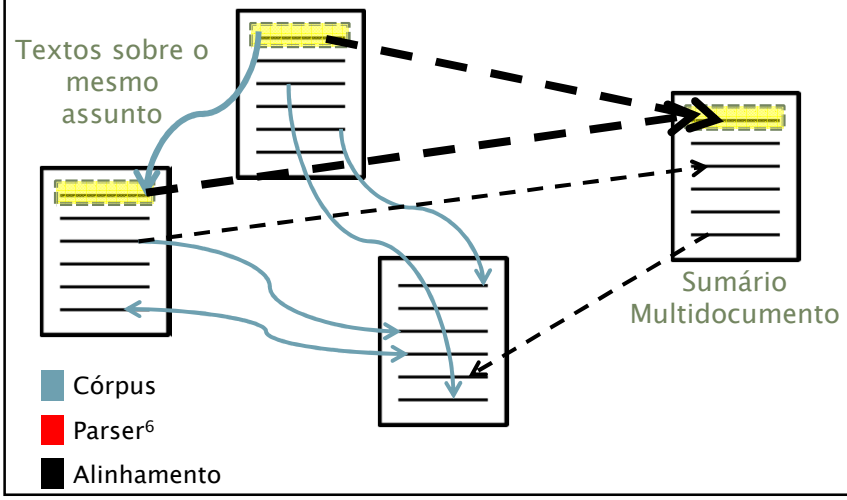
- Abordagem 2 – Relações CST



20

Proposta de mestrado – Verônica

- Abordagem 2 – Relações CST



21

Roteiro

- Alinhamento
  - O que é, onde é usado, tipos
- Propostas de mestrado
- Alinhamento do CSTNews
  - Processo, manual, tipos, exemplos, problemas

22

Alinhamento do CSTNews

- Características do córpus CSTNews[7]
  - Textos jornalísticos em português do Brasil
    - mundo, política, cotidiano, ciência, esportes, dinheiro
  - 50 grupos de textos e 1 sumário humano para cada grupo
  - Média: 3 textos por grupo
  - De 10 a 89 sentenças no grupo (média: 42)

23

Alinhamento do CSTNews

- Processo
  - Treinamento
    - Dois textos
  - Anotação
    - De 1h a 2h por dia
    - Aproximadamente 2 meses
    - Criação de regras para o manual
  - Concordância
    - Cinco textos, temas variados

24

Tipos de alinhamento

Quantidade	Tipos de Alinhamento												
	1-0	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
	2	71	90	67	36	37	13	5	5	1	1	2	1

Total sentenças dos sumários	Total sentenças dos documentos	Total sentenças alinhadas
331	2067	877

25

Exemplos de alinhamento

- Fácil
  - **Sumário**: Antes de chegar à Jamaica, Dean matou ao menos nove pessoas nas ilhas de Santa Lúcia, Dominica, República Dominicana e Haiti, no Caribe.
  - **Documento**: O furacão matou ao menos nove pessoas em sua passagem pelas ilhas do Caribe.

26

Exemplos de alinhamento

- Difícil
  - **Sumário**: Vários moradores e turistas nas regiões, inclusive brasileiros, foram retirados dos locais, enquanto outros estão se preparando para a passagem do furacão.
  - **Documento**: Na Jamaica, muitos estocaram alimentos, água, lanternas e velas.

27

Exemplos do manual

- Alinhar de acordo com a ação
- **Sumário**: Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que giram um ao redor do outro, denominado Oph 162225-240515, o primeiro planemo duplo.
- **Documento**: O pesquisadores Ray Jayawardhana e Valentin D. Ivanov informam a descoberta na edição de quinta-feira do serviço online Science Express, mantido pela revista Science.

28

Exemplos do manual

- Alinhar fragmentos de informação
- **Sumário**: A expectativa de lideranças da Câmara e do Conselho de Ética é que pouco mais de 10% dos 69 deputados denunciados no relatório parcial da CPI abrirão mão de seus mandatos.
- **Documento**: As renúncias têm que ser publicadas até terça-feira, quando o presidente do Conselho de Ética, deputado Ricardo Izar (PTB-SP), vai instaurar os processos de perda de mandato contra os 69 deputados acusados pela CPI dos Sanguessugas de envolvimento com a máfia das ambulâncias.

29

Exemplos do manual

- Alinhar com contradição
- **Sumário**: Às 9h, a cidade tinha oito pontos de alagamento, sendo dois intransitáveis.
- **Documento**: O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às 9h30 desta segunda-feira.

30

Exemplos do manual

- Alinhar com contradição
- **Sumário**: As ações são atribuídas à facção criminosa Primeiro Comando da Capital (PCC), que já comandou outros ataques em duas ocasiões.
- **Documento**: As ações criminosas podem ter sido ordenadas pelos líderes do Primeiro Comando da Capital (PCC), que haviam prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo.

31

Exemplos do manual

- Alinhar uma informação geral a outra específica
- **Sumário**: A Receita Federal intensificou a fiscalização sobre as declarações das pessoas físicas neste ano.
- **Documento** 2: Balanço da fiscalização, divulgado nesta segunda-feira pela Receita mostra que as autuações cresceram 316,5% nos sete primeiros meses deste ano e chegaram a R$ 1,339 bilhão.
- **Documento** 3: O volume de recursos recolhido com multas passou de R$ 326,1 milhões para R$ 1,339 bilhão.

32

Disponibilização dos resultados

- XML
  - Exemplo: Cluster 31

```
<align SENT="1">
   <DOC="D1_C31_Folha.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
   <DOC="D2_C31_Estadao.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
   <DOC="D2_C31_Estadao.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
</align>

<align SENT="2">
   <DOC="D1_C31_Folha.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
   <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
   <DOC="D2_C31_Estadao.txt.seg" SENT="6" TYPE="none" JUDGE="veronica"/>
</align>

<align SENT="3">
   <DOC="D1_C31_Folha.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
   <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
</align>
```

33

Referências

- [7] Cardoso, P.C.F.; Maziero, E.G.; Castro Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. October 26, Cuiabá/MT, Brazil.
- [1] Cremmins, E.T. The art of abstracting. Arlington, Virginia: Information Resources Press, 1996.
- [1] Endres-Niggemeyer, B. (1998) Summarization Information. Springer, Berlin.
- [4] Hasler, L. (2007) From Extracts to Abstracts: Human Summary Production Operations for Computer-Aided Summarisation.
- [4] Jing, H. and McKeown, K. R. (1999) The decomposition of human-written summary sentence. In the Proceedings of the 22th International ACM-SIGIR, New York, p. 129-136.
- [4] Jing, H. and McKeown, K. R. (2000) Cut and paste based text summarization. In the Proceedings of the NAACL Conference, San Francisco, p. 178-185.

34

Referências

- [2] Mani, I., Maybury, M.T. Advances in automatic text summarization. The MIT Press, Cambridge, MA. 1999.
- [5] Mann, W.C. and Thompson, S.A. (1987). Rhetorical structure theory: A theory of text organization. Tech. rep. ISI/RS-87-190, University of Southern California, 83 pp.
- [6] Maziero, E.G. and Pardo, T.A.S. (2011). Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning. In the Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pp. 1-10. October 24-26, Cuiabá/MT, Brazil.
- [3] 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*, pp. 74-83.

35

Obrigada!

renatatironi10@gmail.com
agostini87@gmail.com

36