Investigação de Métodos de Identificação de Redundância para a Sumarização Automática Multidocumento

Jackson Souza (jackcruzsouza@gmail.com)

01/08/2011 a 31/07/2012

Orientação:
Profa. Dra. Ariani Di Felippo (arianidf@gmail.com)
Prof. Dr. Thiago Pardo (taspardo@icmc.usp.br)

Introdução

- Processamento de Língua Natural (PLN)
  - Sumarização Automática Multidocumento (SAM)
    - Produzir sumários a partir de uma coleção de textos-fonte que abordam um mesmo tópico (MCKEOWN, RADEV, 1995)
  - Sumário
    - Conjunto de sentenças que melhor representam o tópico ou assunto da coleção sem que haja informação repetida entre elas (NEWMAN *et al.*, 2004; HENDRICKX *et al.*, 2009)
  - SAM
    - Tratamento da redundância
      - Identificação e eliminação

Redundância

- Em um conjunto de textos que tratam de um mesmo assunto, é possível encontrar um grande volume de informações em comum ou similar, assim como é possível encontrar informações que são únicas de cada texto.

- Redundância (ou similaridade) entre os textos pode ser Total ou Parcial.

Redundância Total

Redundância total	Sentenças
(i) Identidade de forma e conteúdo	S1: Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
	S2: Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
(ii) Identidade de conteúdo	S1: Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
	S2: Nove pessoas morreram, sendo três crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão registrada em um mercado moscovita, informou a Polícia de Moscou.
(iii) Identidade de conteúdo	S1: De acordo com a assessoria do ministério, a transferência dos vôos de Guarulhos para Viracopos não poderá ser feita neste momento, por que o aeroporto de Campinas necessitará de ampliação, principalmente em terminal de passageiros.
	S2: Para receber os vôos de Cumbica, Viracopos precisará ser ampliado, sobretudo seu terminal de passageiros, segundo nota do Ministério da Defesa.

Redundância Parcial

Redundância parcial	Sentenças
(i) S1 contém X e Y, S2 contém X e Z	S1: A falha no reversor – mecanismo que ajuda o avião a frear – foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado.
	S2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado.
(ii) S1 contém X e Y, S2 contém X.	S1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas.
	S2: Ao menos 17 pessoas morreram após um acidente aéreo na República Democrática do Congo.

Introdução

- Etapa de **transformação** na maioria dos métodos de SAM (RADEV *et al.*, 2004):
  - calcular a importância de cada sentença dos textos-fonte;
  - ranquear as sentenças em função de sua importância;
  - selecionar a sentença de maior pontuação de importância para iniciar o sumário;
  - selecionar a próxima sentença do ranque;
  - **calcular a redundância ou similaridade** da nova sentença candidata em relação à sentença já selecionada para o sumário;
  - selecionar a sentença candidata para compor o sumario se esta contiver pouca sobreposição com a sentença inicialmente selecionada, e
  - repetir os passos para as demais sentenças do ranque até que o tamanho desejado do sumário seja alcançado

Introdução

- Identificação/cálculo da Redundância

  - Métodos superficiais simples (Hatzivassiloglou *et al.*, 1999, 2001)
    - **Estatísticos**
      - word *overlap*
      - noun *overlap*
      - adjective *overlap*
      - verb *overlap*
      - adverb *overlap*
    - **Linguísticos**
      - sobreposição de etiquetas morfossintáticas
      - sobreposição de radicais
      - sobreposição de núcleos de sintagmas nominais
      - sobreposição de palavras sinônimas

  - Métodos superficais compostos (Hatzivassiloglou *et al.*, 1999, 2001)
    - sobreposição de palavras + ordem
    - sobreposição de palavras + distância entre elas

Introdução

- Identificação/cálculo da Redundância
  - Métodos profundos
    - sobreposição de conceitos lexicalizados (Newman *et al*, 2004)
      → pares de sentenças que apresentam maior número de palavras relacionadas na WN.Pr são mais similares
    - sobreposição de núcleos sintagmáticos semanticamente relacionados (Hendrickx *et al*., 2009)
    - sobreposição de conteúdo (Castro Jorde; Pardo, 2010)
      - relações CST de redundância

Objetivos

- Diante da escassez de trabalhos sobre SAM que envolvem o português do Brasil
  - Investigar os principais métodos superficiais (estatísticos e linguísticos) de detecção da redundância entre sentenças
  - Investigar a correspondência desses métodos com outro mais profundo, que identifica a redundância com base no tipo de relação semântica do modelo *Cross-document Structure Theory (CST)* (RADEV, 2000)

Metodologia

- **1. Seleção do *corpus***
  - Características: (i) monolíngue (PB); (ii) jornalístico, (iii) multidocumento e (iv) alinhado no nível retórico, via CST.

- **2. Delimitação e teste dos métodos superficiais**
  - Aplicação dos principais métodos superficiais (estatísticos e linguísticos) identificados na literatura ao *corpus*

- **3. Estudo da correlação entre os métodos e as relações CST**
  - Verificar a correlação entre os métodos e o nível de redundância (indicado pelas relações CST), e as relações CST que unem as sentenças de cada par

- **4. Avaliação**
  - Verificar quais métodos são adequados para a identificação de cada nível de redundância, e relação CST

Seleção do *corpus*

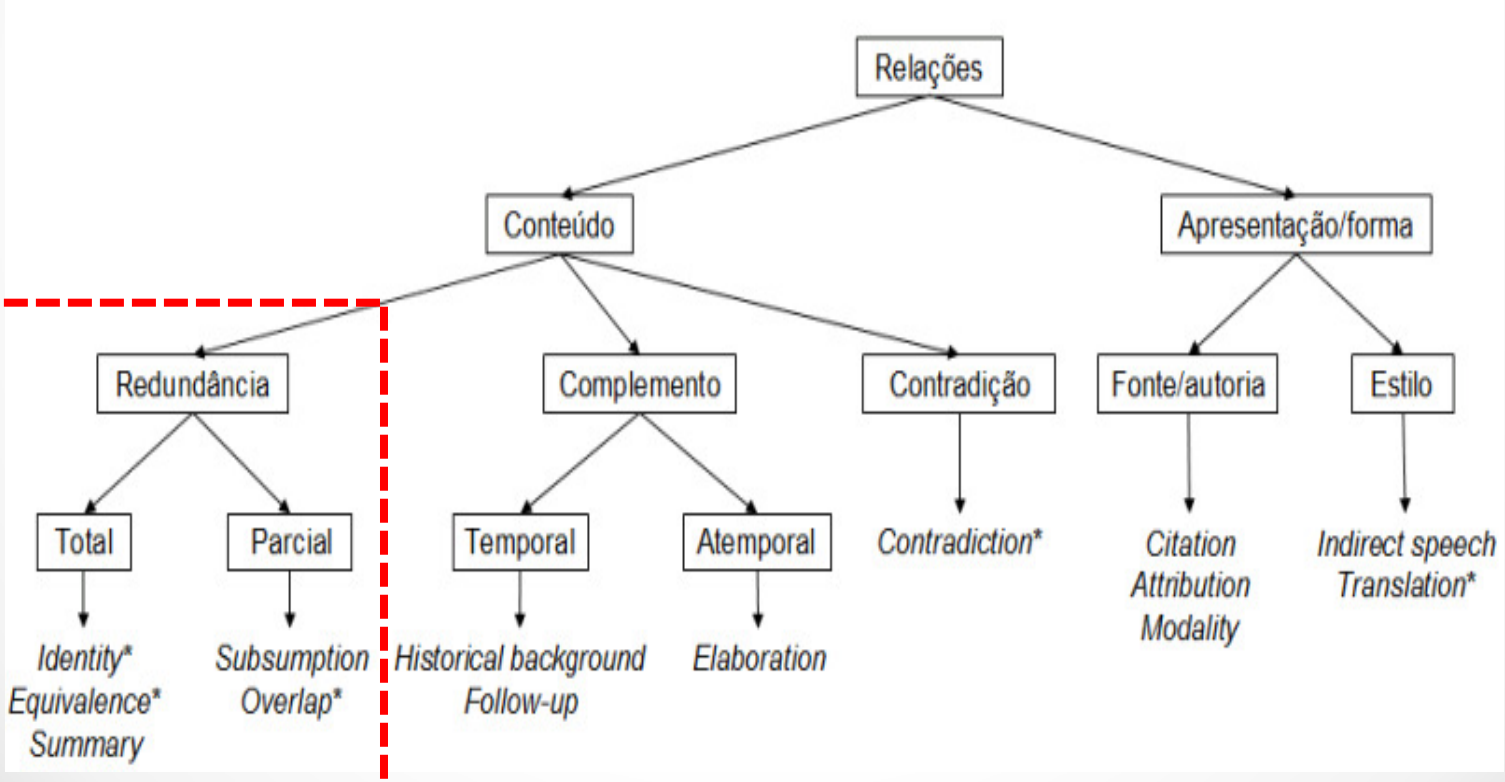
- Tipologia
  - Monolíngue
    - Característica que reflete o objetivo de se estudar a redundância em PB
  - Multidocumento
    - Característica que reflete o objetivo de se estudar a redundância entre textos que tratam de um mesmo assunto
  - Alinhado/ anotação em nível retórico (via CST)
    - Característica que reflete o objetivo de se investigar a possível correlação entre os métodos superficiais de detecção da redundância e as relações CST (RADEV, 2000)

Seleção do *corpus*

- **CSTNews** (Cardoso *et al.* 2011)
  - *Corpus* multidocumento em PB
    - 50 coleções de textos jornalísticos de fontes e domínios variados
    - 2 a 3 textos por coleção (=140 texts)
    - Sumários automáticos e manuais mono e multidocumento
    - 42 sentenças em média por texto-fonte
    - 7 sentenças em média por sumário manual multidocumento
    - E também:
      - Anotações
        - discursiva monodocumento (via RST)
        - discursiva multidocumento (via CST)
        - temporal (HAREM guidelines)
        - sentido dos nomes
        - subtópico/tópico segmentação
        - etc.
- **Recorte**

Seleção do *corpus*

- Montagem de um *subcorpus*
  - 45 pares de sentenças do CSTNews com diferentes níveis de redundância
    - Tipologia de Maziero *et al.* (2010)



Seleção do *corpus*

- ***Subcorpus*** do CSTNews

Tipo de relação	Relação	Quantidade de pares
Redundância total	Identity	5
	Equivalence	6
	Summary	4
Redundância parcial	Subsumption	8
	Overlap	8
Não-redundância*	----	14

* Sentenças de *clusters* distintos

Delimitação e Teste dos métodos

- Delimitação dos métodos superficiais
  - **Estatísticos**
    - word overlap (MWol)
    - noun overlap (MNol)
    - verb overlap (MVol)
    - adjective overlap (MAdj)
    - adverb overlap (MAdv)

$$Wol\ (S1, S2) = \frac{\#CommonWords}{\#Words(S1)\ +\ \#Words\ (S2)}$$

Delimitação e Teste dos Métodos

- Delimitação dos métodos superficiais
  - **Linguísticos**
    - sobreposição de padrões morfossintáticos (MPdMorf)
    - sobreposição de verbo principal (MVp)
    - sobreposição de núcleo de sujeito (MSuj)
    - sobreposição de núcleo de objeto/predicativo principal (MObj)
    - sobreposição de palavras sinônimas (MSin)
    - sobreposição de etiquetas morfossintáticas (MEtMorf)
  - **Estrutural**
    - localização das sentenças (MLoc)

Delimitação e Teste dos Métodos

- Caracterização das sentenças do subcorpus em função dos atributos linguísticos referentes aos diferentes métodos superficiais
  - Pré-processamento → anotação morfossintática automática
    - LX-Tagger (Branco, Silva, 2004)
    - 96,87% de precisão + *Interface* amigável

Par	Relação	Cluster	Sentenças anotadas
1	Identity (Redundância total)	D2_C1	As/DA vítimas/vítima/CN de_/PREP o/DA acidente/acidente/CN foram/ser/V 14/DGT passageiros/passageiro/CN e/CJ três/CARD membros/membro/CN de_/PREP a/DA tripulação/tripulação/CN
		D3_C1	As/DA vítimas/vítima/CN de_/PREP o/DA acidente/acidente/CN foram/ser/V 14/DGT passageiros/passageiro/CN e/CJ três/CARD membros/membro/CN de_/PREP a/DA tripulação/tripulação/CN

Delimitação e Teste dos Métodos

- Caracterização das sentenças do *subcorpus* em função dos atributos linguísticos referentes aos diferentes métodos superficiais I dos atributos
    - Tabelas em formato Excell

Par	**Atributo linguístico**					
	Método Estrutural	***Método Linguístico***				
	Loc	*PdMorf*	*NSuj*	*Vp*	*NObjPredp*	*Sin*
1	S2	CN PREP CN (2) membro da tripulação	vítima	ser	passageiro, membro	0
	S2	CN PREP CN (2) membro da tripulação	vítima	ser	passageiro, membro	0

Delimitação e Teste dos Métodos

- Caracterização das sentenças do *subcorpus* em função dos atributos linguísticos referentes aos diferentes métodos superficiais I dos atributos
    - Tabelas em formato Excell

Par	Atributo linguístico					
	Método Linguístico					
	Palavra	*Nome*	*Verbo*	*Adjetivo*	*Advérbio*	*Etiqueta morfossintática*
1	vítima, acidente, ir, passageiro, membro, tripulação	vítima, acidente, passageiro, membro, tripulação	ir	--	--	DA CN PREP V DGT CJ CARD
	vítima, acidente, ir, passageiro, membro, tripulação	vítima, acidente, passageiro, membro, tripulação	ir	--	--	DA CN PREP V DGT CJ CARD

Teste dos Métodos

- Aplicação manual dos métodos superficiais
  - Tabelas em formato Excell

Par	Rel. CST	Método											
		Estrut.	Estatístico					Linguístico					
		MLoc	MWol	MNol	MVol	MADJol	MADVol	MPdMorf	MSuj	MVp	MObjPredp	MSin	MEtMorf
01	Ident.	0	1	1	1	**NA**	**NA**	1	Sim	Sim	Sim	Não	1

Par	Rel. CST	Método											
		Estrut.	Estatístico					Linguístico					
		MLoc	MWol	MNol	MVol	MADJol	MADVol	MPdMorf	MSuj	MVp	MObjPredp	MSin	MEtMorf
01	Ident.	0	1	1	1	**NA**	**NA**	1	Sim	Sim	Sim	Não	1
02	Ident.	0	1	1	1	1	1	1	Sim	Sim	Sim	Não	1
03	Ident.	0	1	1	1	1	1	1	Sim	Sim	Sim	Não	1
04	Ident.	0	1	1	1	1	**NA**	1	Sim	Sim	Sim	Não	1
05	Ident.	0	1	1	1	1	**NA**	1	Sim	Sim	Sim	Não	1
06	Sum.	0	0,23	1	0,66	1	0,66	0	Não	Não	Não	Não	1
07	Sum.	0,18	0,8	1	1	1	**NA**	1	Sim	Sim	Sim	Não	1
08	Sum.	0,45	0,30	0,33	0	**NA**	0,5	0	Não	Não	Não	Não	0,66
09	Sum.	0,45	0,36	0,47	0.25	**NA**	0	0	Não	Não	Não	Não	0,92
10	Equi.	0,9	0,66	1	1	0	0	1	Sim	Sim	Não	Sim	0,72
11	Equi.	0,36	0,19	0,36	0	0	0	0	Não	Não	Não	Não	0,82
12	Equi.	0,18	0,45	0,61	0,5	0	0	0,4	Não	Sim	Não	Sim	0,52
13	Equi.	0,27	0,42	0,42	0,5	0,66	0	1	Não	Sim	Não	Sim	0,77
14	Equi.	0	0,68	0,46	0,6	0,5	**NA**	0	Sim	Sim	Não	Não	0,90
15	Equi.	0,09	0,36	0,57	1	**NA**	**NA**	0	Sim	Sim	Não	Não	1
16	Subs.	0	0,47	0,46	0,5	0,4	**NA**	0,5	Sim	Não	Não	Não	0,85
17	Subs.	0	0,22	0,30	0	0	**NA**	0,5	Não	Não	Não	Não	0,88
18	Subs.	0	0,33	0,5	0	**NA**	**NA**	0	Não	Não	Não	Não	0,77
19	Subs.	0,27	0,08	0,20	0,16	**NA**	0	0	Não	Não	Não	Não	0,75
20	Subs.	0	0,2	0,12	0,4	0	0	0	Não	Não	Não	Não	0,75
21	Subs.	0,18	0,5	0,30	0,66	0,8	0,5	0	Não	Não	Não	Não	0,6
22	Subs.	0,36	0,28	0,30	0,4	0	**NA**	0	Não	Não	Não	Não	0,63
23	Subs.	1	0,15	0,13	0,33	**NA**	0	1	Não	Não	Não	Sim	0,71
24	Over.	0,36	0,44	0,66	0	0	0	0,8	Não	Não	Não	Não	0,66
25	Over.	0,36	0,44	0,47	0	0,66	0	0,66	Não	Não	Não	Não	0,5
26	Over.	0,09	0,29	0,12	0	0,5	0	1	Não	Não	Sim	Não	0,28
27	Over.	0,09	0,3	0,28	0,4	**NA**	0	0,66	Não	Não	Não	Não	0,88
28	Over.	0	0,42	0,25	0,28	0	0	0,44	Não	Não	Não	Não	0,9
29	Over.	0,18	0,42	0,6	0	**NA**	**NA**	0	Sim	Não	Não	Não	0.8
30	Over.	0	0,29	0,28	0	0,33	0	0,66	Não	Não	Não	Sim	0,77
31	Over.	0	0,2	0,44	0	0,66	0	0,66	Sim	Não	Não	Não	0.82
32	Perm.	0,18	0	0	0	0	0	0,33	Não	Não	Não	Não	0,625
33	Perm.	0	0	0	0	0	0	0,33	Não	Não	Não	Não	0,84
34	Perm.	1	0,06	0	0	0	0,5	0,4	Não	Não	Não	Não	0,8
35	Perm.	0,09	0	0	0	0	0	0,66	Não	Não	Não	Não	0,84
36	Perm.	0,72	0	0	0	0	0	0	Não	Não	Não	Não	0,77
37	Perm.	0	0,16	0,09	0,4	0	0	0,5	Não	Não	Não	Não	0,7
38	Perm.	0,9	0	0	0	0	0	0	Não	Não	Não	Não	0,66
39	Perm.	0	0	0	0	0	0	0,66	Não	Não	Não	Não	0,66
40	Perm.	0,27	0	0	0	0	0	0,5	Não	Não	Não	Não	0,66
41	Perm.	0,54	0	0	0	0	0	0	Não	Não	Não	Não	0,8
42	Perm.	0,27	0,08	0	0,5	**NA**	**NA**	0,66	Não	Não	Não	Não	0,75
43	Perm.	0	0,08	0	0,28	0	0	0	Não	Não	Não	Não	0,75

Correlação: métodos *vs.* redundância

- Análise manual preliminar
  - Cálculo da **média simples**
  - Identificação do **número de pares** que obtiveram valores iguais ou maiores que a média simples*

    *média simples do MLoc = pares que obtiveram valores iguais ou inferiores à média simples

- Análise automática
  - Aprendizado de Máquina
    - Weka (*Waikato Environment for Knowledge Analysis*) (Frank *et al.*, 2011)

Correlação: métodos *vs.* redundância

- Análise manual preliminar
  - Cálculo da **média simples**
  - Identificação do **número de pares** que obtiveram valores iguais ou maiores que a média simples*****

Método	Nível de redundância		
	Total	Parcial	Nula
MLoc	10/15	9/16	9/14
MWol	8/15	7/16	4/14
MNol	8/15	6/16	1/14
MVol	8/15	7/16	3/14
MPdMorf	8/15	10/16	7/14
MSuj	9/15	3/16	0/14
MVp	11/15	0/16	0/14
MObPredp	6/15	1/16	0/14
MSin	4/15	2/16	0/14
MEtMorf	10/15	11/16	8/14

Correlação: métodos *vs.* redundância

- Análise manual preliminar
  - Cálculo da **média simples**
  - Identificação do **número de pares** que obtiveram valores iguais ou maiores que a média simples*****

Método	Nível de redundância		
	Total	Parcial	Nula
MLoc	10/15	9/16	9/14
MWol	8/15	7/16	4/14
MNol	8/15	6/16	1/14
MVol	8/15	7/16	3/14
MPdMorf	8/15	10/16	7/14
MSuj	9/15	3/16	0/14
MVp	11/15	0/16	0/14
MObPredp	6/15	1/16	0/14
MSin	4/15	2/16	0/14
MEtMorf	10/15	11/16	8/14

1ª observação: MLoc, MPdMorf e MEtMorf parecem não expressar as diferenças de redundância, pois destacaram-se nos pares dos 3 níveis.

Correlação: métodos *vs.* redundância

- Análise manual preliminar
  - Cálculo da **média simples**
  - Identificação do **número de pares** que obtiveram valores iguais ou maiores que a média simples*****

Método	Nível de redundância		
	Total	Parcial	Nula
MLoc	10/15	9/16	**9/14**
MWol	8/15	7/16	4/14
MNol	8/15	6/16	1/14
MVol	8/15	7/16	3/14
MPdMorf	8/15	10/16	7/14
MSuj	9/15	3/16	0/14
MVp	11/15	0/16	0/14
MObPredp	6/15	1/16	0/14
MSin	4/15	2/16	0/14
MEtMorf	10/15	11/16	8/14

1ª observação: MLoc, MPdMorf e MEtMorf parecem não expressar as diferenças de redundância, pois destacaram-se nos pares dos 3 níveis.

Destaque de MLoc quanto à redundância "nula" é coincidência.

Correlação: métodos *vs.* redundância

- Análise manual preliminar
  - Cálculo da **média simples**
  - Identificação do **número de pares** que obtiveram valores iguais ou maiores que a média simples*****

Método	Nível de redundância		
	Total	Parcial	Nula
MLoc	10/15	9/16	**9/14**
MWol	8/15	7/16	4/14
MNol	8/15	6/16	1/14
MVol	8/15	7/16	3/14
MPdMorf	8/15	10/16	7/14
MSuj	9/15	3/16	0/14
MVp	11/15	0/16	0/14
MObPredp	6/15	1/16	0/14
MSin	4/15	2/16	0/14
MEtMorf	10/15	11/16	8/14

2ª observação: MWol, MNol e MVol parecem expressar a existência ou não de redundância, pois se destacaram em muitos pares de redundância total e parcial e em poucos pares de sentenças não redundantes.

Correlação: métodos *vs.* redundância

- Análise manual preliminar
  - Cálculo da **média simples**
  - Identificação do **número de pares** que obtiveram valores iguais ou maiores que a média simples*****

Método	Nível de redundância		
	Total	Parcial	Nula
MLoc	10/15	9/16	**9/14**
MWol	8/15	7/16	4/14
MNol	8/15	6/16	1/14
MVol	8/15	7/16	3/14
MPdMorf	8/15	10/16	7/14
MSuj	9/15	3/16	0/14
MVp	11/15	0/16	0/14
MObPredp	6/15	1/16	0/14
MSin	4/15	2/16	0/14
MEtMorf	10/15	11/16	8/14

3ª observação: MSuj e MVp parecem expressar a redundância total, posto que se destacaram em pares totalmente redundantes

Correlação: métodos *vs.* redundância

- Análise manual preliminar
  - Cálculo da **média simples**
  - Identificação do **número de pares** que obtiveram valores iguais ou maiores que a média simples*****

Método	Nível de redundância		
	Total	Parcial	Nula
MLoc	10/15	9/16	**9/14**
MWol	8/15	7/16	4/14
MNol	8/15	6/16	1/14
MVol	8/15	7/16	3/14
MPdMorf	8/15	10/16	7/14
MSuj	9/15	3/16	0/14
MVp	11/15	0/16	0/14
MObPredp	6/15	1/16	0/14
MSin	4/15	2/16	0/14
MEtMorf	10/15	11/16	8/14

4ª observação:
Os métodos MSuj, MVp, MObPredp e MSin não obtiveram valores iguais ou superiores à média simples nem nenhum dos pares da categoria "redundância nula".

Correlação: métodos *vs.* redundância

- Análise automática
  - Aprendizado de Máquina (Weka)
    - Algoritmos com as melhores precisões: PART e J48

PART (Precisão de 97.7%)

Se MNol ≤ 0.09 então **nulo**	(14)
Senão se MVp = não e MEtMorf<=0.9 e Mloc ≤ 0.27 então **parcial**	(12)
Senão se MVp = sim então **total**	(11)
Senão se MPdMorf ≤ 0.33 então **total**	(5/1)
Senão **parcial**	(3)

Correlação: métodos *vs.* redundância

- Análise automática
  - Aprendizado de Máquina (Weka)
    - Algoritmos com as melhores precisões: PART e J48

PART (Precisão de 97.7%)

Se MNol ≤ 0.09 então **nulo** (14)
Senão se MVp = não e MEtMorf<=0.9 e Mloc ≤ 0.27 então **parcial** (12)
Senão se MVp = sim então **total** (11)
Senão se MPdMorf ≤ 0.33 então **total** (5/**1**)
Senão **parcial** (3)

Par	Relação	Cluster	Sentenças originais
22	Subs.	D2_C15	Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
		D3_C15	Anteriormente, a Polícia havia informado sobre nove mortos, sendo três deles crianças, e 25 feridos.

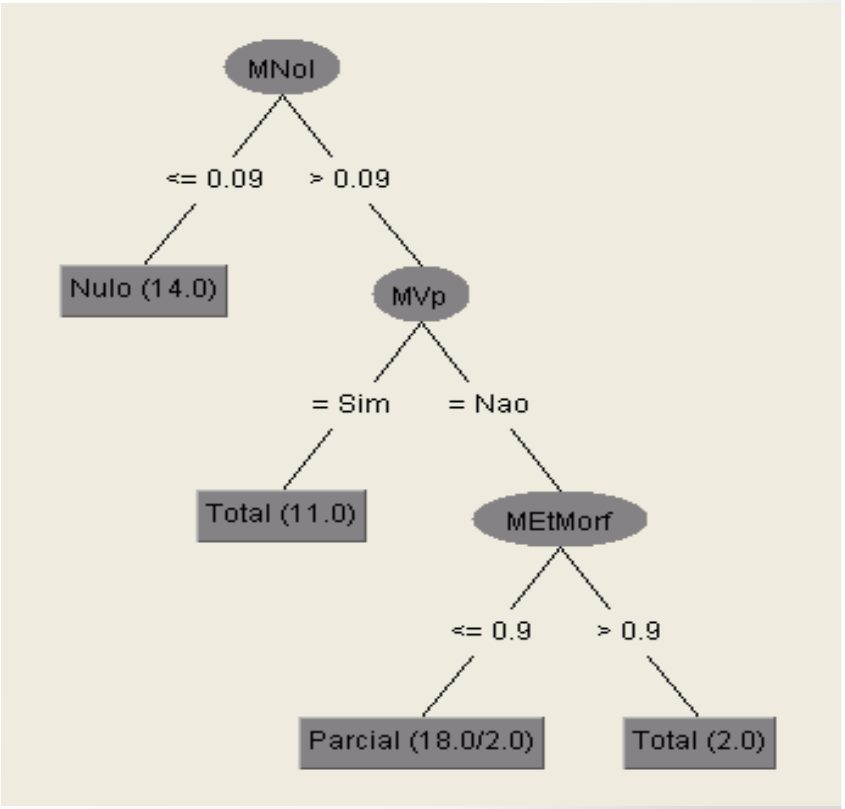
Par	Rel. CST	Método											
		Estrut.	Estatístico					Linguístico					
		MLoc	MWol	MNol	MVol	MADJol	MADVol	MPdMorf	MSuj	MVp	MObjPredp	MSin	MEtMorf
22	Subs.	0,36	0,28	0,30	0,4	0	NA	0	Não	Não	Não	Não	0,63

Correlação: métodos *vs.* redundância

- Análise automática
  - Aprendizado de Máquina
    - Weka
      - Algoritmos com as melhores precisões: PART e J48

J48 (Precisão de 95.5%)

MNol <= 0.09 então **nulo** (14.0)
MNol > 0.09
    MVp = sim então **total** (11.0)
    MVp = não
        MEtMorf <= 0.9 então **parcial** (18.0/2.0)
        MEtMorf > 0.9 então **total** (2.0)



Correlação: métodos *vs.* relações CST

Método	Relação CST					
	Identity	Summary	Equivalence	Subsumption	Overlap	Permuta
MLoc	5/5	3/4	4/6	5/8	5/8	9/14
MWol	5/5	1/4	2/6	4/8	4/8	4/14
MNol	5/5	2/4	3/6	5/8	4/8	1/14
MVol	5/5	2/4	3/6	5/8	2/8	2/14
MPdMorf	5/5	1/4	3/6	3/8	6/8	9/14
MSuj	5/5	1/4	3/6	2/8	2/8	0/14
MVp	5/5	1/4	5/6	1/8	0/8	0/14
MObPredp	5/5	1/4	0/6	2/8	1/8	0/14
MSin	5/5	0/4	4/6	1/8	3/8	1/14
MEtMorf	5/5	3/4	3/6	5/8	6/8	8/14

Correlação: métodos *vs.* relações CST

- Análise automática
  - Aprendizado de Máquina (Weka)
    - Algoritmos com as melhores precisões: PART e J48

PART (Precisão de 88,88%)

Se MNol <= 0.09 então **Permuta** (14.0)
Senão se MWol <= 0.8 e MVp = não e MEtMorf <= 0.9 e MLoc <= 0.36 e
MPdMorf <= 0.5 então **Subsumption** (10.0/3.0)
Senão se MVp = não AND MPdMorf > 0.33 então **Overlap** (7.0/1.0)
Senão se MWol <= 0.8 AND MVp = sim então **Equivalence** (6.0/1.0)
Senão se MPdMorf > 0.5 então **Identity** (5.0)
Senão **Summary** (3.0)

Correlação: métodos *vs.* relações CST

- Análise automática
  - Aprendizado de Máquina (Weka)
    - Algoritmos com as melhores precisões: PART e J48

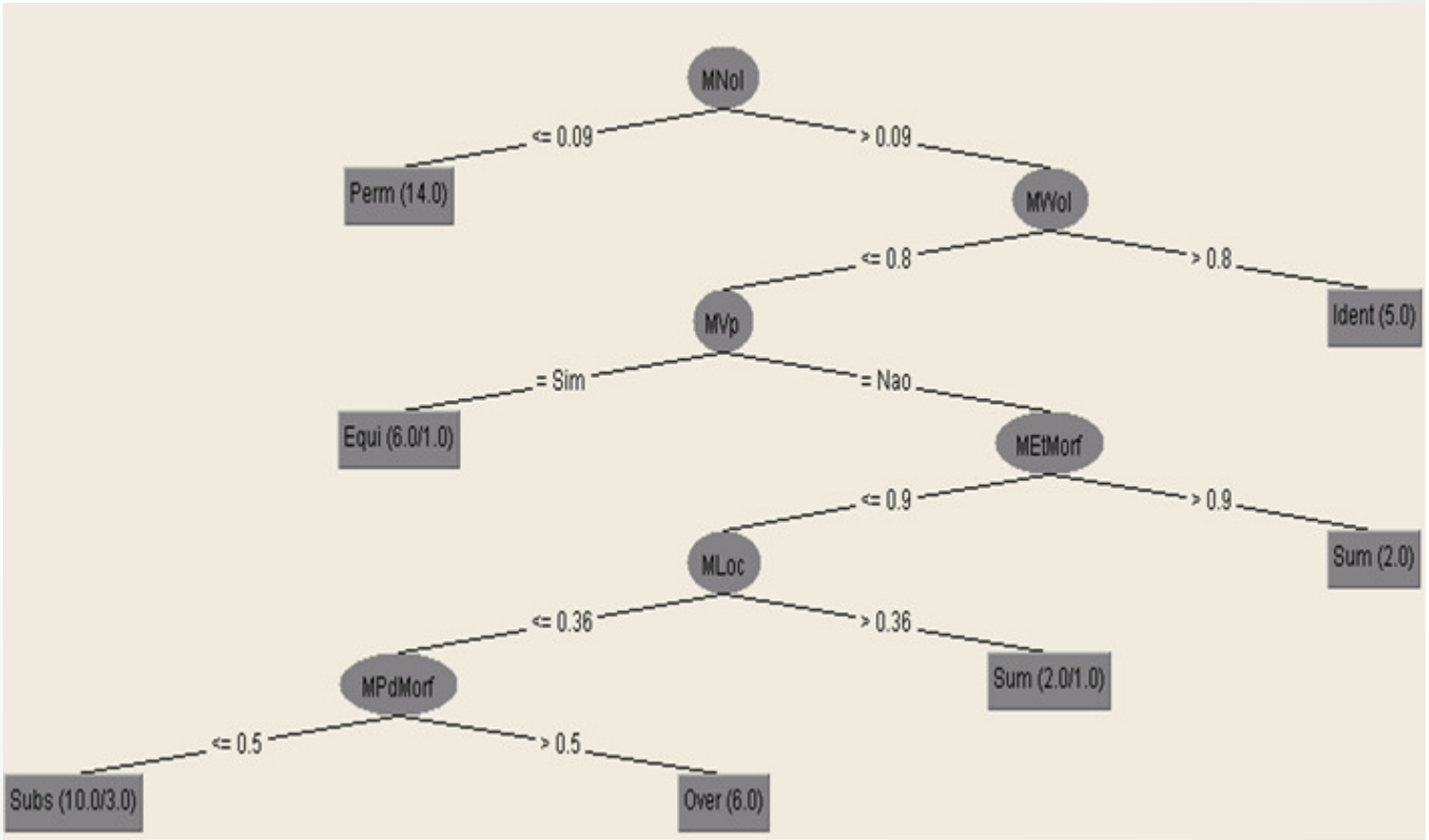
J48 (Precisão de 88,88%)

MNol <= 0.09 então **Permuta** (14.0)
MNol > 0.09
| MWol <= 0.8
| | MVp = sim então **Equivalence** (6.0/1.0)
| | MVp = não
| | | MEtMorf <= 0.9
| | | | MLoc <= 0.36
| | | | | MPdMorf <= 0.5 então **Subsumption** (10.0/3.0)
| | | | | MPdMorf > 0.5 então **Overlap** (6.0)
| | | | MLoc > 0.36 então **Summary** (2.0/1.0)
| | | MEtMorf > 0.9 então **Summary** (2.0)
| MWol > 0.8 então **Identity** (5.0)

Correlação: métodos *vs.* relações CST

- Análise automática
  - Aprendizado de Máquina (Weka)
    - Algoritmos com as melhores precisões: PART e J48

J48



Considerações finais

- Gerais
  - O conjunto de dados de análise é pequeno, já que é composto por 45 pares de sentenças, assim, os resultados são preliminares a respeito dos métodos
  - O conjunto de métodos é composto apenas por atributos "superficiais"
  - As análises manual e automática são complementares

Considerações finais

- Específicas
  - Métodos *vs* nível de redundância
    - Os métodos Mwol, Mnol, MVp e Msin foram os atributos preferenciais dos algoritmos PART e J48
    - MNol é capaz de identificar pares que não são redundantes dos que apresentam alguma redundância
    - A redundância identificada como padrão foi a "parcial"
    - O PART precisa apenas de 4 atributos (métodos) (MNol, MVp, MEtMorf e MLoc) para identificar a redundância entre as sentenças com 97.7% precisão
    - O J48 precisa de 5 atributos (MNol, MVp, MEtMorf, MLoc e MPdMorf) para identificar a redundância entre as sentenças com 95.5% precisão

Considerações finais

- Específicas
  - Métodos vs relações CST
    - Os métodos Mwol, Mnol, MVp e MSin foram os atributos preferenciais dos algoritmos PART e J48
    - MNol é capaz de identificar com precisão as sentenças que não estão relacionadas por relações CST nos algoritmos PART e J48
    - Os algoritmos PART e J48 utilizam o mesmo conjunto de métodos (atributos)
      - MNol, MWol, MVp, MEtMorf, MLoc e MPdMorf
    - Os algoritmos PART e J48 classificam as sentenças quanto às relações CST com a mesma precisão: **88,88%**
    - A relação identificada como padrão pelo algoritmo PART foi "Summary"

Trabalhos futuros

- TCC
  - Investigação de métodos de identificação de redundância baseados em atributos profundos
    - atributos semânticos
      - entidades nomeadas (Palmer, Day, 1997; Aranha, 2007)
      - hiperonímia/hiponímia (Fellbaum *et al,* 1998)
      - *polarização semântica (Ilari, 1984; Linegarger, 1987; Vitral, 1999)

1. A aviação de Israel realizou durante a madrugada desta segunda-feira, dia 7, **ataques** a 150 alvos no Líbano.

2. A Força Aérea israelense lançou uma série de **bombardeios** contra o Líbano nesta segunda-feira.

1. A pista principal do **Aeroporto Internacional de São Paulo** (Cumbica), em Guarulhos, será totalmente reformada em março de 2008, segundo informações do Ministério da Defesa anunciadas nesta segunda-feira, 6.

2. O Ministério da Defesa anunciou nesta segunda-feira (6) que em março do ano que vem uma das pistas do **Aeroporto de Guarulhos** será fechada para reformas de seu trecho central.

Cumbica
Aeroporto de Guarulhos
Aeroporto Internacional de São Paulo
Aeroporto Internacional de Guarulhos
...

Referências Bibliográficas

- ALEIXO, P.; PARDO, T.A.S. CSTNews: um corpus de textos jornalísticos anotados segundo a Teoria Discursiva Multidocumento CST (*Cross-document Structure Theory*). **Série de Relatórios Técnicos do ICMC**, São Carlos-SP, n. 326, 12p., 2008.
- BRANCO, A; SILVA, J. Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4, 2004, Lisbon. **Proceedings...** Lisbon, 2004, p. 507-510.
- CARDOSO, P.C.F.; MAZIERO, E.G.; JORGE, M.L.C.; SENO, E.M.R.; DI FELIPPO, A.; RINO, L.H.M.; NUNES, M.G.V.; PARDO, T.A.S. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3, 2001, Cuiabá, Brasil. **Proceedings...** Cuiabá, 2004, p. 88-105.
- FELLBAUM, C. (Ed.). **WordNet: an electronic lexical database**. Ca, MA: MIT Press, 1998.
- FERREIRA, A. B. H. **Novo dicionário eletrônico Aurélio da língua portuguesa**. Curitiba: Ed. Positivo, 2004. 1 CD-ROM
- FREGE, G. Lógica e filosofia da linguagem. Tradução: Paulo Alcoforado. São Paulo: Cultrix/Edusp, 1978.
- FRANK, E. WITTEN, I. H. HALL, M.A. **Data Mining: Pratical Machine Learning Toos and Techniques**. 3a Ed. MK. Waikato, 2011.
- HATZIVASSILOGLOU, V.; KLAVANS, J. L.; ESKIN, E. Detecting text similarity over short passages: exploring linguistic feature combinations via Machine Learning. In: JOINT SIGDAT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND VERY LARGE CORPORA, 1999, College Park, Maryland. **Proceedings...** Maryland, 1999, p. 203-12.
- ___ et al. SIMFINDER: A Flexible Clustering Tool for Summarization. In: NAACL WORKSHOP ON AUTOMATIC SUMMARIZATION, 2001, Pittsburg (PA), USA. **Proceedings...** Pittsburg, USA, 2001.
- HENDRICKX, I.; DAELEMANS, W.; MARSI, E., KRAHMER, E. Reducing redundancy in multi-document summarization using lexical semantic similarity. In: WORKSHOP ON LANGUAGE GENERATION AND SUMMARISATION, 2009, Singapore. **Proceedings...** Singapore, 2009, p. 63–66.
- HOUAISS, A.; VILLAR, M. de S. **Dicionário eletrônico Houaiss da língua portuguesa**. (versão 1.0). Rio de Janeiro: Editora Objetiva, 2001. 1 CD-ROM.
- JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall: New Jersey, 2009.

Referências Bibliográficas

- LAGE, N. **Estrutura da notícia**. 5ª ed. São Paulo: Ática, 2002.
- NEWMAN, E.; DOMN, W.; STOKES, N.; CARTHY, J.; DUNNION, J. Comparing redundancy removal techniques for multi-document summarization. In: STARTING AI RESEARCHERS' SYMPOSIUM, 2, 2004, Valencia. **Proceedings..**Valencia, 2004, p. 223-28.MCKEOWN, K.; RADEV, D.R. Generating summaries of multiple news articles. In: INTERNATIONAL ACM-SIGIR, 18, 1995, Seattle. **Proceedings…**Seattle, 1995, p. 74-82. MANI, I. **Automatic Summarization**. John Benjamins Publishing Co., Amsterdam, 2001.
- ___.; MAYBURY, M.T. **Advances in automatic text summarization**. The MIT Press, Cambridge, MA. 1999. MARTINS, C. B. et al. Introdução à Sumarização Automática. **Rel. Técnico RT-DC 002/2001**, Departamento de Computação, UFSCar, São Carlos. Abril, 2001. 38p.
- MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying Multidocument Relations. In: NLPCS, 7, 2010, Funchal, PT. **Proceedings...** Funchal, 2010, p. 60-69.
- ___.; PARDO, T.A.S.; DI FELIPPO, A.; DIAS-DA-SILVA, B.C. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TIL) 6, 2008, Vila Velha-ES, Brasil. **Proceedings...**2008, p. 390-392.
- RADEV, D. R. et al. MEAD-a platform for multidocument multilingual text summarization. In: International Conference on Language Resources and Evaluation (LREC), 4, 2004, Lisbon. **Proceedings…** Lisbon, 2004, **Proceeings…** p. 1-4.
- ___. A common theory of information fusion from multiple text sources, step one: cross-document structure". In: ACL Signal Workshop on Discourse and Dialogue, 1, 2000, Hong Kong, **Proceedings…** Hong Kong, 2000, p. 74-83.
- SINCLAIR, J. Corpus and text: basic principles. In: Wynne, M. (Ed.). **Developing linguistic corpora: a guide to good practice**. Oxford: Oxbow Books, 2005. p.1-16. Disponível em: <www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>. Acesso em: 02 ago. 2010.
- SPARCK JONES, K. Discourse modeling for Automatic Summarization. **Tech. Report No. 290**. University of Cambridge. UK, February, 1993.
- ____. Automatic summarising: a review and discussion of the state of the art. **Technical Report UCAM-CL-TR-679**. University of Cambridge. 2007.