Identificação da Redundância na Sumarização Automática Multidocumento: Explorando Métodos Superficiais

Jackson W. C. Souza[1,2], Ariani Di Felippo[1,2], Thiago A. S. Pardo[2]

[1] Departamento de Letras (DL) – Centro de Educação e Ciências Humanas (CECH)
Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905– São Carlos – SP – Brazil

[2] Núcleo Interinstitucional de Linguística Computacional (NILC)
Inst. de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 - São Carlos, - SP - Brazil

{jackcruzsouza, arianidf}@gmail.com, taspardo@icmc.usp.br

1. Introdução

A Sumarização Automática Multidocumento (SAM) visa à produção de sumários a partir de uma coleção de textos que versam sobre um mesmo assunto [Mani, 2001]. Por partir de textos com conteúdo similar, os métodos de SAM precisam identificar e tratar a redundância, pois um sumário deve conter o conteúdo principal da coleção sem repetição de informação. Para a identificação da redundância, há vários métodos que se baseiam em conhecimentos linguísticos superficiais e profundos [p.ex.: Hatzivassiloglou *et al.*, 1999; Newman *et al*, 2004; Hendrickx *et al*., 2009]. Neste trabalho, apresenta-se a investigação de um conjunto de métodos superficiais.

2. Delimitação dos Métodos e Descrição e Caracterização do *Corpus*

Com base principalmente nos trabalhos de Hatzivassiloglou *et al.*(1999), Newman *et al* (2004) e Hendrickx *et al*. (2009), foram especificados 5 métodos linguísticos e 3 estatísticos. Os linguísticos são: (i) sobreposição de padrões morfossintáticos (PdMorf), (ii) sobreposição de verbo principal (Vp), (iii) sobreposição de núcleo de sujeito (Suj), (iv) sobreposição de núcleo de objeto/predicativo principal (ObjPredp) e (v) sobreposição de etiquetas morfossintáticas (EtMorf). O conjunto dos métodos estatísticos engloba os seguintes: (i) *word overlap* (Wol), (ii) *noun overlap* (Nol) e (iii) *verb overlap* (Vol). Além dos métodos advindos da literatura, especificou-se outro, de natureza estrutural, segundo o qual quanto menor a distância entre as posições que as sentenças ocupam em seus respectivos textos-fonte, maior a redundância entre elas. A esse método, deu-se a denominação de Loc.

Para a investigação dos métodos selecionados, partiu-se do CSTNews[1] [Cardoso *et al.*, 2011], *corpus* jornalístico multidocumento composto por 50 coleções de notícias. No CSTNews, os textos de uma coleção estão inter-relacionados em nível sentencial pelas relações da teoria/modelo *Cross-Document Structure Theory* (CST) [Radev, 2000], sendo que algumas delas capturam diferentes níveis de redundância. Diante disso, construiu-se um *subcorpus* do CSTNews cuja composição é apresentada no Quadro 1. As sentenças que compõem os 14 pares com redundância nula foram

[1] http://caravelas.icmc.usp.br/CSTNews/

Anais do III Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana, páginas 22–24, Fortaleza, CE, Brasil, Outubro 21–23, 2013.

selecionadas de coleções que tratam assuntos distintos e, por isso, não possuem relações.

Quadro 1. Composição do *corpus* de análise

Nível de redundância	Relação CST	Quantidade de pares de sentenças
Redundância total	*Identity*	5
	Equivalence	6
	Summary	4
Redundância parcial	*Subsumption*	8
	Overlap	8
Redundância nula	--	14

Para a aplicação dos métodos, caracterizou-se manualmente cada uma das 45 sentenças em função dos atributos relativos aos métodos. No Quadro 2, ilustra-se a caracterização do par 15, composto pelas sentenças "*O prazo foi definido pela Mesa Diretora da Câmara*" e "*O prazo foi definido pela direção da Câmara*", advindas de textos distintos.

Quadro 2. Caracterização linguística das sentenças do *corpus*

Par	Atributos Linguísticos								
	Loc	**Palavra**	**Nome**	**Verbo**	**Padrão morf.**	**Núcleo/ sujeito**	**Verbo principal**	**Núcleo/ Objeto**	**Etiqueta morf.**
15	S3	prazo, foi, definido, mesa, diretora, câmara	prazo, mesa, diretora, câmara	ser, definir	[NProp+Prep + NProp]	prazo	definir	mesa	Art, N, V, Prep, NProp
	S2	prazo, foi, definido, direção, câmara	prazo, direção, câmara	ser, definir	[N+Prep+ NProp]	prazo	definir	direção	Art, N, V, Prep, NProp

Após a descrição dos atributos linguísticos, aplicou-se manualmente cada método. A aplicação consistiu na verificação da sobreposição dos atributos entre as sentenças de cada um dos 45 pares. Na Tabela 1, apresentam-se os resultados da aplicação dos métodos ao par 15 (redundância parcial) do *corpus*.

Tabela 1. Exemplo de aplicação dos métodos superficiais

Par	Nível	Método								
		Estrut.	**Estatístico**			**Linguístico**				
		Loc	Wol	Nol	Vol	PdMorf	Suj	Vp	ObjPredp	EtMorf
15	Total	0,09	0,36	0,57	1	0	Sim	Sim	Não	1

3. Teste dos Métodos e Resultados

Os valores obtidos pelos métodos foram submetidos a vários algoritmos do ambiente de aprendizado de máquina *Weka* (*Waikato Environment for Knowledge Analysis*) [Frank *et al*., 2011], que aprendem padrões estatisticamente relevantes. Os índices mais altos de precisão foram obtidos pelo algoritmo PART, que gera regras no formato *se, então*. No Teste 1, os valores obtidos pelos 9 métodos foram submetidos em conjunto. Nos Testes 2, 3, 4, 5, 6, 7, 8, 9 e 10, os valores de cada um dos 9 métodos foram testados individualmente. No Teste 11, os valores dos 2 métodos de melhor desempenho individual foram submetidos em conjunto. Na Tabela 2, apresentam-se os testes ranqueados em função da precisão das regras aprendidas.

No Teste 1, aplicação conjunta dos 9 métodos gerou as regras em (1), que obtiveram 97,7% de precisão Dos 45 pares, apenas 1 foi classificado erroneamente pela aplicação da regra 4.

(1) 1. Se Nol<=0.09 então nulo (14 acertos)
2. Senão se Vp = não e EtMorf <=0 .9 e Loc <=0.27 então parcial (12 acertos)
3. Senão se Vp=sim então total (11 acertos)
4. Senão se PdMorf<=0.33 então total (5 acertos /**1 erro**)
5. Senão parcial (3 acertos)

As regras geradas somente com base nos valores de Nol obtiveram os mais altos índices de precisão individual, 91,1% (Teste 4). Na sequência, destacam-se as regras geradas em função somente de Wol, que obtiveram 80% (Teste 3). A combinação de Nol e Wol obteve (91,1%) (Teste 11), mesma precisão obtida por Nol em isolado.

Tabela 2. Precisão dos métodos segundo o algoritmo PART

Teste	Método									Precisão (%)
	Loc	Wol	Nol	Vol	PdMorf	Suj	Vp	ObjPredp	EtMorf	
1	■	■	■	■	■	■	■	■	■	97,7
4			■							91,1
11		■	■							91,1
3		■								80
8							■			57,7
5				■						55,5
10									■	55,5
6					■					53,3
7						■				53,3
9								■		46,6
2	■									42,2

4. Considerações finais

Da investigação apresentada, observa-se que: (i) os métodos superficiais em questão discriminam com precisão (97,7%) os diferentes níveis de redundância; (ii) o método Nol é suficientemente capaz de discriminar com os níveis de redundância com alta precisão (91,1%). No futuro, pretende-se confirmar essas observações em um *corpus* maior.

Referências

Cardoso, P.C.F. *et al* (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, p. 88-105.

Frank, e.; Witten, I. H.; Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. 3a Ed. MK. Waikato, 2011.

Hatzivassiloglou, V.; Klavans, J. L.; Eskin, E. (1999). Detecting text similarity over short passages: exploring linguistic feature combinations via Machine Learning. In the Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-1999), p. 203-12.

___., Gravano, L., Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In the Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-00), p. 224–231.

Hendrickx, I.; Daelemans, W.; Marsi, E., Krahmer, E. (2009). Reducing redundancy in multi-document summarization using lexical semantic similarity. In the Proceedings of the Workshop on Language Generation and Summarisation, pp. 63–66.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co, Amsterdam.