Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006) - Poster Section, Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Uma comparação entre sistemas de sumarização automática extrativa

Daniel Saraiva Leite e Lucia Helena Machado Rino

Departamento de Computação, UFSCar
CP 676, 13565-905 São Carlos, SP, Brazil
Núcleo Interinstitucional de Lingüística Computacional
http://www.nilc.icmc.usp.br
{daniel_leite; lucia}@dc.ufscar.br

Resumo. Neste artigo é apresentada uma comparação entre o TextRank e o Supor-2, sendo o primeiro um sumarizador independente de língua baseado somente na representação em grafos das sentenças e o segundo um sumarizador que combina diversas técnicas clássicas utilizando aprendizado de máquina com o auxílio do ambiente WEKA. Adicionalmente, variações do TextRank pela inclusão de algum processamento lingüístico também foram consideradas, resultando em duas versões adicionais de sumarizadores extrativos também consideradas na comparação. Seu objetivo principal foi analisar o potencial do Supor-2 para sumarizar textos em português.

1 Introdução

De modo geral, sumarizar é o processo de produção de uma versão reduzida de um texto, geralmente pela seleção e estruturação de seu conteúdo informativo mais relevante [16]. Desse processo, pode-se originar um *extrato* ou um *sumário,* distinção feita somente pela forma como o conteúdo selecionado do texto é organizado. Um extrato, particularmente, é o texto construído simplesmente pela justaposição dos segmentos em foco, copiados literal e seqüencialmente do texto original.

Considerando essa metodologia extrativa de sumarização automática, este artigo apresenta uma comparação entre quatro sumarizadores de textos em português. O primeiro é o SuPor-2 [7], que combina técnicas clássicas da área numa abordagem envolvendo aprendizado de máquina. Os três outros sistemas são implementações distintas de um mesmo método de sumarização: o TextRank ([11],[12]). Esse sistema foi escolhido devido ao fato de ser considerado independente de língua natural (LN) e já ter sido avaliado para um corpus de textos em português [12]. Neste caso, a comparação apresentada neste artigo (Seção 4) utiliza a mesma metodologia empregada na avaliação do TextRank. As duas variações desse sistema foram delineadas a partir da incorporação de recursos lingüísticos específicos e distintos para o português do Brasil, conforme é descrito na Seção 3, após a apresentação do SuPor-2 (Seção 2). Considerações sobre o desempenho do SuPor-2, objetivo central da comparação relatada, são feitas na Seção 5.

2 O SuPor-2

O SuPor-2 é uma versão modificada do SuPor (Ambiente para **Su**marização Automática de Textos em **Por**tuguês) [13], um sumarizador extrativo que depende de informações fornecidas por um engenheiro de conhecimento para treino e combinação de diversos métodos de extração de informações relevantes. Foram mantidos do SuPor tanto o método de classificação das informações (Naïve-Bayes) quanto os métodos de SA extrativa, a saber: *Tamanho da Sentença* [5], *Posição das Sentenças* [4], *Freqüência das Palavras* [9], *Nomes Próprios* [5], *Cadeias Lexicais* [1], *Importância dos Tópicos* [6] e *Mapa de Relacionamentos* [15]. A modificação introduzida visou facilitar a escolha de características relevantes para o treinamento e, assim, tornar o sistema mais independente do conhecimento do engenheiro, pois não era possível descobrir sistematicamente qual a melhor combinação de métodos para a SA de determinado texto. Tampouco o SuPor permitia recuperar informações que levassem à análise das situações em que seu desempenho flutuava devido a escolhas indevidas de características durante seu treinamento. Em outras palavras, o SuPor-2 resulta da modificação do módulo específico de aprendizado do SuPor, que foi integralmente substituído pelo WEKA [17], um ambiente bastante amigável com o mesmo fim.

O investimento nessa versão foi motivado pelos bons resultados do SuPor, quando comparado a outros sete sumarizadores para o português do Brasil [7].

3 O TextRank

O TextRank é baseado no mesmo algoritmo usado pelo Google para julgar a relevância das páginas WEB, o PageRank [2]. Este considera que a importância de uma página é proporcional ao número de recomendações que existem na WEB para ela, as quais são, basicamente, os *links* que apontam para a página. O PageRank constrói um grafo em que os vértices são as páginas e as arestas são suas recomendações. Assim, é capaz de percorrer o grafo para definir a importância das páginas, que são consideradas em seu processo de busca de informações. Seguindo a mesma idéia, o TextRank constrói um grafo similar, em que os vértices são sentenças e as arestas são denotadas pelo grau de similaridade entre pares de sentenças. De forma análoga à do PageRank, o grafo é percorrido para definir a importância das sentenças (vértices) e, então, classificá-las para inclusão no extrato, mediante uma taxa de compressão.

A implementação original desse método [12] realiza o cálculo de similaridade entre um par de sentenças baseando-se apenas na ocorrência de seus termos comuns. As duas variações consideradas na comparação com o SuPor-2 consideram, respectivamente: (1) o pré-processamento do texto-fonte pela remoção de *stopwords* e *stemização*, resultando no cálculo de similaridade pela ocorrência de *stems* comuns; (2) a utilização de um *thesaurus* [3] do português, resultando na consideração de relações de sinonímia e antonímia no cálculo da similaridade entre as sentenças.

4 A Comparação entre os sistemas propostos

Foram comparados os quatro sistemas extrativos antes descritos: o SuPor-2, o TextRank e suas duas variações, aqui indicadas por 'TextRank+Stem+StopRem' e 'TextRank+Thesaurus'. Também foi considerado um *baseline*, o *TopFirst*, que seleciona as

sentenças na ordem em que ocorrem no texto-fonte a partir da primeira, até atingir a taxa de compressão. Este foi o mesmo *baseline* usado por Mihalcea [12], já que foi este o experimento reproduzido. Como tal, somente o grau de informatividade dos extratos foi calculado automaticamente, com o auxílio da ROUGE[1] [8]. Mais particularmente, foi usada a medida ROUGE(1,1), isto é, a que utiliza somente unigramas para verificar co-ocorrências e um sumário de referência para comparação. Também manteve-se o mesmo corpus de teste, o TeMário [14], que contém 100 textos jornalísticos e seus respectivos sumários manuais e extratos ideais e a mesma taxa de compressão (30%). Somente o SuPor-2 exigiu treinamento prévio, efetuado com os extratos ideais do TeMário e o esquema de *10-fold cross validation*.

A Tabela 1 mostra os resultados obtidos para o SuPor-2 e para as nossas duas versões modificadas do TextRank (os resultados dos demais sistemas foram simplesmente copiados de [12]).

Tabela 1. Informatividade dos extratos

Sistema	ROUGE(1,1)
SuPor-2	0,5839
TextRank+Thesaurus	0,5603
TextRank+Stem+StopRem	0,5426
TopFirst	0,4963
TextRank	0,4939

Como se vê, a incorporação de processamento lingüístico ao TextRank trouxe melhoras razoáveis em relação ao TextRank "puro", de 13% com o uso do *thesaurus* e 10% com *stemming* e remoção de *stopwords*. De certa forma, esses resultados já eram esperados, pois o conhecimento lingüístico traz mais subsídios para a conectividade entre informações textuais e, assim, parece natural que os extratos sejam mais informativos. Já o desempenho pior que o *baseline TopFirst*, do TextRank, **pode** indicar que sua aplicação ao português, embora considerada naquele experimento independente de língua natural, só será válida se houver algum pré-processamento específico para o português, como proposto em nossas duas implementações. A última comparação que se faz é entre o SuPor-2 e o 'TextRank+Thesaurus'. Nota-se que há um ganho de informatividade do SuPor-2 de apenas 4%, mas os níveis de processamento empregados são bastante distintos. Enquanto o SuPor-2 considera sete métodos de SA e exige um esforço de treino, essa implementação do TextRank é de menor complexidade computacional e utiliza apenas um thesaurus em seu processamento lingüístico.

5 Considerações finais

Os resultados apresentados na seção anterior levam a questionar as limitações dos métodos independentes de LN, cujo patamar de informatividade não ultrapassa, atualmente, os 50%. Se considerada a complexidade do SuPor-2 em relação ao TextRank, poderíamos argüir que não valeria a pena o esforço da implementação. No entanto, considerando seu ganho de informatividade em relação à versão independente de LN (de 18%) e a demanda atual por melhores meios de transmitir mais informação, o questionamento que se apresenta é que, para se construir sumarizadores automáticos que, de fato, tenham utilidade prática, não

[1] Disponível em http://www.isi.edu/~cyl/ROUGE/ [Agosto/2006]

parece possível desconsiderar algum conhecimento lingüístico. Neste sentido, a própria melhora do TextRank ao incorporar um thesaurus sugere que recursos lingüísticos mais sofisticados devem ser explorados, se o foco for maior informatividade.

Reconhecimentos

Este trabalho conta com o apoio do CNPq.

Referências

1. Barzilay, R., Elhadad, M.: Using Lexical Chains for Text Summarization. In: Mani, I., Maybury, M. T. (eds.): Advances in Automatic Text Summarization. MIT Press (1997) 111-121
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, (1998), 1-7
3. Dias-da-Silva, B. C., Oliveira, M. F., Moraes, H. R., Paschoalino, C., Hasegawa, R., Amorin, D., Nascimento, A. C.: Construção de um Thesaurus Eletrônico para o Português do Brasil. In: Proc. of the V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000), São Carlos, Brasil (2003) 1-11
4. Edmundson, H.P.: New methods in automatic extracting. Journal of the Association for Computing Machinery 16 (1969) 264-285
5. Kupiec, J., Pedersen, J., Chen, F.: A trainable document summarizer. In: Fox, E. A., Ingwersen , P., Fidel, R. (eds.): Proc. of the 18th ACM-SIGIR Conference on Research & Development in Information Retrieval. Seatlle, WA (1995) 68-73
6. Larocca Neto, J., Santos, A.D., Kaestner, A.A., Freitas, A.A.: Generating Text Summaries through the Relative Importance of Topics. Lecture Notes in Artificial Intelligence, No. 1952. Springer-Verlag (2000) 200-309
7. Leite, D. S., Rino, L. H. M.: Selecting a Feature Set to Summarize Texts in Brazilian Portuguese. In: Sichman, J. S. et al. Proc. of 18th Brazilian Symposium on Artificial Intelligence (SBIA'06) and 10th Ibero-American Artificial Intelligence Conference (IBERAMIA'06). Lecture Notes in Artificial Intelligence, No. 4140, Springer-Verlag (2006) 462-471
8. Lin, C. Y.: ROUGE: A package for automatic evaluation of summaries. In: Proc. of the Workshop on Text Summarization Branches Out (WAS-2004), Barcelona, Spain (2004) 74-81
9. Luhn, H.: The automatic creation of literature abstracts. IBM Journal of Research and Development 2 (1958) 159-165
10. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. MIT Press (1999)
11. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, July (2004)
12. Mihalcea, R.: Language Independent Extractive Summarization. In: Proc. of the 43th. Annual Meeting of the Association for Computational Linguistics, Companion Volume (ACL 2005), Ann Arbor, MI, June (2005)
13. Módolo, M.: SuPor: an Environment for Exploration of Extractive Methods for Automatic Text Summarization for Portuguese (in Portuguese). MSc. Dissertation. Departamento de Computação, UFSCar (2003)
14. Pardo, T.A.S., Rino, L.H.M.: TeMário: A corpus for automatic text summarization (in Portuguese). NILC Tech. Report NILC-TR-03-09 (2003)
15. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic Text Structuring and Summarization. In: Information Processing & Management 33 (1997) 193-207
16. Sparck-Jones, K.: What might be in a summary? Information Retrieval 93. Universitatsverlag Konstanz (1993) 9-26
17. Witten, I. H, Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann, San Francisco (2005)