Experimentos com Sumarização Automática Extrativa de Textos Científicos

Pedro Paulo Balage Filho, Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes

Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

ppbalage@grad.icmc.usp.br, {taspardo, gracan}@icmc.usp.br

1. Introdução

A quantidade de informação disponível no atual momento é muito grande, impossível de ser apreendida em sua totalidade. Para amenizar o problema, costuma-se procurar versões menores, mais enxutas: resumos. Também chamados de sumários, tratam-se de versões reduzidas dos textos a que se referem e contêm as idéias principais dos mesmos [1].

Apesar de bastante úteis, a produção dos sumários é bastante trabalhosa, visto que é necessária a leitura e a interpretação do texto para, então, perceberem-se suas idéias centrais. Por isso, hoje em dia, buscam-se formas automáticas de produzir esses resumos. A esta área de estudo dá-se o nome de Sumarização Automática de Textos.

Para o português do Brasil, há, atualmente, diversos sumarizadores disponíveis, como apresentado por Rino et al. [2]. Um dos primeiros sumarizadores que surgiu para esta língua e que ainda é bastante utilizado é o GistSumm [3]. O GistSumm é um sumarizador que utiliza métodos estatísticos e/ou empíricos para se obter o sumário.

2. Objetivos

Por se basear em um método superficial relativamente simples e produzir somente extratos (isto é, sumários compostos por sentenças inteiras do texto-fonte, sem modificações), o GistSumm apresenta diversas limitações.

O objetivo deste trabalho é aprimorar o método de sumarização deste sistema para sanar algumas de suas limitações observadas, mais especificamente, o tratamento de textos estruturados, ou seja, textos que apresentam seções/subseções. Mais especificamente, visa-se à sumarização de artigos científicos, produzindo-se um artigo mais curto, mas que preserve a estruturação e distribuição original da informação.

3. Materiais e Métodos

Inicialmente, procedeu-se a uma avaliação subjetiva de textos e sumários produzidos pelo GistSumm. Evidenciamos que o sumário gerado era prejudicado em função da divisão de seu conteúdo em seções: algumas seções eram privilegiadas e outras completamente ignoradas no sumário, perdendo-se grande parte do conteúdo principal do texto-fonte. Verificou-se que, tratando-se cada seção separadamente, ou seja, produzindo sumários individuais para cada seção e depois os justapondo, os sumários resultantes eram melhores. Portanto, o GistSumm foi modificado para ser capaz de reconhecer seções dentro de um texto e tratá-las isoladamente para a composição do sumário final.

Principalmente, foram elaboradas modificações no sistema exclusivas para a estrutura textual de artigos científicos. Para estes foram incluídas duas opções: (a) o sistema manterá no extrato o título da seção

e (b) no mínimo uma sentença de cada seção deverá ser selecionada para o sumário. O item (a) garante que os cabeçalhos das seções permaneçam no extrato, preservando com isso a estrutura textual. O item (b) impõe que todas as seções colaborem com pelo menos uma sentença no sumário final.

4. Resultados

Para avaliação do GistSumm para textos científicos, em comparação com o sistema GistSumm original, duas avaliações distintas foram conduzidas. A primeira avaliação foi baseada em um julgamento por um juiz humano sobre um córpus de 20 textos da língua portuguesa. Com a finalidade de corroborar esta, uma segunda avaliação foi feita utilizando-se a ferramenta ROUGE [4] para avaliação automática de sumários. Nesse caso, foi usado um córpus de 150 textos de língua inglesa. Os resultados obtidos em ambas as avaliações demonstraram que as modificações realizadas no sistema contribuíram para a melhoria de seus sumários em textos que possuem estruturação interna. Para mais detalhes do processo de desenvolvimento e avaliação do sistema, recomenda-se a leitura de [5].

5. Conclusões Finais

Neste trabalho, pudemos mostrar que sumarizadores extrativos simples podem conseguir melhores resultados quando melhorias diretas são incorporadas no processo de sumarização.

6. Agradecimentos

Ao PET-SESU-MEC e às agências de fomento à pesquisa FAPESP, CAPES e CNPq.

7. Referências Bibliográficas

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

[2] Rino, L.H.M.; Pardo, T.A.S.; Silla Jr., C.N.; Kaestner, C.A.; Pombo, M. (2004). A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In the *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence* – SBIA (Lecture Notes in Artificial Intelligence 3171), pp. 235-244. São Luis-MA, Bazil. September, 29 - October, 1

[3] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In *Lecture Notes in Artificial Intelligence* 2721, pp. 210-218. Faro, Portugal. June 26-27.

[4] Lin, C-Y. and Hovy, E.H. (2003). Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In the *Proceedings of Language Technology Conference* – HLT. Edmonton, Canada. May 27 - June 1.

[5] Balage Filho, P.P.; Pardo, T.A.S.; Nunes, M.G.V. (2007). Summarizing Scientific Texts: Experiments with Extractive Summarizers. In the *Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications* – ISDA. Rio de Janeiro-RJ, Brazil. October, 22-24.