Sumarização Automática de Artigos 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

1. Objetivos

A sumarização automática caracteriza-se pela geração de sumários (resumos) através de métodos computacionais. Um dos primeiros sumarizadores que surgiu para o português e que ainda é bastante utilizado é o GistSumm [1]. Por se basear em um método relativamente simples e produzir somente extratos (isto é, sumários compostos por sentenças inteiras do texto-fonte, sem modificações), o sistema 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.

2. Material 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. 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 tratar seções isoladamente para a composição do sumário. Também foram incluídas duas opções no novo sistema: (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 sua estrutura. O item (b) impõe que todas as seções colaborem com pelo menos uma sentença no sumário final.

3. Resultados e Discussão

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 científicos da língua portuguesa. Com a finalidade de corroborar esta, uma segunda avaliação foi feita utilizando-se a ferramenta ROUGE [2] 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.

4. Conclusões

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

5. Agradecimentos

Este trabalho contou com o apoio da FAPESP, CAPES e CNPq.

6. Referências Bibliográficas

[1] 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.

[2] Lin, C-Y. and Hovy, E.H. (2003). Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In the *Proceedings of HLT Conference*. Edmonton, Canada.