Investigação de Mapas de Relacionamento para Sumarização Multidocumento

Rafael Ribaldo, Thiago A. S. Pardo
Instituto de Ciências Matemática e de Computação - ICMC/USP São Carlos
ribaldo@usp.br, taspardo@icmc.usp.br

Objetivos

A tarefa de sumarização automática multidocumento consiste em se produzir automaticamente um único sumário a partir de um grupo de textos sobre um mesmo assunto. Neste projeto, são exploradas estratégias de construção de sumários com base em um dos "mapas de relacionamentos" propostos por Salton et al. (1997), em que se tenta representar no sumário os principais subtópicos presentes nos textos de origem ao mesmo tempo em que se faz a manutenção da informatividade do sumário produzido.

Método

O principal método investigado neste trabalho é o método chamado "denso segmentado" proposto por Salton et al. (1997). Neste método, adaptado para a sumarização multidocumento por Ribadlo et al. (2012), representam-se os textos de origem em um grafo, em que cada vértice é uma sentença e as arestas indicam a similaridade lexical entre as sentenças correspondentes. Tendo o grafo construído, identificam-se grupos de sentenças mais topicalmente relacionadas para, então, selecionarem-se as sentenças mais relevantes de cada subtópico para compor o sumário final. Utiliza-se a técnica TextTiling (Hearst, 1997) adaptada para a língua portuguesa para auxiliar na identificação e agrupamento de subtópicos.

Resultados e Conclusões

Este trabalho está em andamento e, atualmente, foca-se no estudo e na implementação das técnicas relacionadas à identificação e agrupamento de subtópicos. A construção do grafo e o método de seleção de sentenças estão completamente desenvolvidos e poderão ser utilizados tão logo os subtópicos sejam satisfatoriamente detectados. A próxima etapa deste projeto consiste na avaliação dos resultados, que será feita utilizando-se um conjunto de dados de referência, o córpus CSTNews (Cardoso et al., 2011), composto por 50 grupos de textos e sumários construídos manualmente. Os sumários produzidos automaticamente serão comparados com os sumários humanos deste córpus, produzindo-se medidas de desempenho do método de sumarização investigado. Essas medidas também permitirão comparar os resultados obtidos neste projeto com os resultados já existentes na área, como os apresentados por Ribaldo et al. (2012).

Agradecimentos

À FAPESP, pelo suporte a este projeto.

Referências

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. (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*, pp. 88-105.

Hearst, M. (1997).TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, Vol. 23, N. 1, pp. 33-64.

Ribaldo, R.; Akabane, A.T.; Rino, L.H.M.; Pardo, T.A.S. (2012). Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In the *Proceedings of the 10th International Conference on Computational Processing of Portuguese* (LNAI 7243), pp. 260-271. April 17-20, Coimbra, Portugal.

Salton, G.; Singhal A.; Mitra, M; Buckley C. (1997). Automatic Text Structuring And Summarization. *Information Processing & Management*, Vol. 33, No, 2, pp. 193-207.