Novos Métodos e Abordagens para Sumarização de Textos com Base em Informação Discursiva

Vinícius R. Uzêda, Thiago A. S. Pardo, Maria das Graças V. Nunes
Instituto de Ciências Matemáticas e da Computação (ICMC), USP/São Carlos

1. Objetivos

Em trabalho anterior [1], inferiu-se que o uso da *Rhetorical Structure Theory* (RST) [2] no âmbito de sumarização automática tinha eficácia limitada. A avaliação comparativa entre vários dos métodos de sumarização clássicos na área resultou em resumos automáticos de equivalente informatividade. Visando aumentar o escopo da avaliação, mais métodos foram propostos e resolveu-se utilizar algoritmos de Aprendizado de Máquina (AM) em busca de um método que superasse os demais.

2. Material e Métodos

Além dos métodos avaliados em [1], isolou-se o aperfeiçoamento proposto por Marcu em [3] a seu método, e uma variação do mesmo foi implementada. Consultando materiais que remontam aos princípios da RST [4], acrescentou-se uma proposta dos próprios autores da teoria, aqui tratada como método de Mann & Thompson. Avaliaram-se também métodos híbridos, fruto da combinação dos métodos anteriores com outros estatísticos, incluindo o sumarizador GistSumm [5], bem como uma variação do mesmo, e medidas de grafos sintáticos e baseadas na localização das sentenças no texto.

3. Resultados

Utilizando um conjunto de dados de referência da área, foram avaliados 128 diferentes métodos de sumarização, resultantes da permutação das métricas propostas. Para a avaliação, foram utilizados 30 textos que possuem anotação retórica e resumos criados por especialistas. Os resultados, embora extensos, levam às conclusões sintetizadas a seguir. Quanto a proposta de utilização de AM, ainda não há resultados conclusivos, mas há a expectativa de obter-se um método com um desempenho significativamente melhor que os demais.

4. Conclusões

Em geral, percebe-se que os métodos RST não apresentam melhorias significativas quando combinados com Mann & Thompson (na realidade, para alguns métodos, o desempenho foi ligeiramente pior). Ainda, têm um desempenho pior quando combinados com o sumarizador estatístico GistSumm. Pode-se dizer que todos os métodos baseados na RST têm desempenhos equivalentes.

5. Agradecimentos

À FAPESP, CAPES e CNPq.

6. Referências

[1] Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2007). *Estudo e Avaliação de Métodos de Sumarização Automática de Textos Baseados na RST*. ICMC-USP Technical Report. São Carlos-SP, August, 28p.

[2] Mann, W.C.; Thompson, S.A: 1987, 'Rhetorical Structure Theory: A Theory of Text Organization'.

[3] Marcu, D.: 1998, 'To build text summaries of high quality, nuclearity is not sufficient'. In: *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*. Stanford, CA.

[4] Mann, W.C.; Matthiessen, C.M.I.M.; Thompson, S.A. (1992). 'Rhetorical Structure Theory and Text Analysis'.

[5] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V.: 2003, 'GistSumm: A Summarization Tool Based on a New Extractive Method'. In: *6th Workshop on Computational Processing of Portuguese Language – Written and Spoken – PROPOR* (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal.