Avaliação de Métodos Automáticos de Sumarização de Textos Baseados em Conhecimento Discursivo

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

O processo de sumarização automática de textos consiste na produção de uma versão mais curta de um texto, ou seja, seu resumo. Alguns métodos de sumarização são baseados em conhecimento empírico e estatístico, produzindo resultados duvidosos muitas vezes. Em outra linha, alguns métodos fazem uso massivo de conhecimento lingüístico de variados níveis e complexidades, visando a produção de sumários melhores, próximos dos produzidos por humanos. Nessa linha, muitos métodos se baseiam em conhecimento discursivo, seguindo a teoria *Rhetorical Structure Theory* (RST) [1], uma das mais difundidas atualmente na área.

Nesse trabalho, realizou-se uma avaliação comparativa dos principais métodos de sumarização automática de textos baseados na RST. Nós também propusemos um novo método e comparamos nossos resultados a um sumarizador extrativo simples.

2. Material e Métodos

Os métodos avaliados foram os propostos por Marcu [2, 3], por O'Donnell [4] e Ono et al. [5]. Cada um destes métodos aborda aspectos diferenciados do discurso para a produção do sumário. O método proposto por nós é descrito em detalhes em [6]. O sumarizador extrativo utilizado é o GistSumm [7].

3. Resultados

Após avaliar os resultados, pudemos perceber 2 fatos proeminentes: os métodos baseados na RST produzem resultados similares; todos os métodos baseados na RST se desempenharam melhores do que o sumarizador extrativo.

4. Conclusões

Em geral, nossos resultados indicam que não importa qual método baseado em RST seja escolhido, pois todos apresentam resultados semelhantes. Por outro lado, como esperado na área, todos eles foram melhores do que o método extrativo.

5. Agradecimentos

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

6. Referências

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

[2] Marcu, D.: 1997a, 'From discourse structures to text summaries'. In: I. Mani & M. Maybury (Eds.), *Proceedings of ACL Workshop on Intelligent Scalable Text Summarization* (pp. 82 - 88). Madrid, Spain.

[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] O'Donnell, M.: 1997, 'Variable-Length On-Line Document Generation'. In: *Proceedings of the 6th European Workshop on Natural Language Generation*. Gerhard-Mercator University, Duisburg, Germany.

[5] Ono, K.; Sumita, K.; Miike, S.: 1994, 'Abstract generation based on rethorical structure extraction'. In: *Proceedings of Coling*.

[6] 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.

[7] 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.