Sumarização Automática de Textos Estruturados

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 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 dos textos: resumos, ou sumários. A sumarização automática caracteriza-se pela geração de sumários 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. Por exemplo, visa-se à sumarização de artigos científicos, produzindo-se um artigo mais curto, mas que preserve a estruturação original da informação.

2. Material e Métodos

Inicialmente, procedeu-se a uma avaliação subjetiva de textos e sumários produzidos pelo GistSumm. Os seguintes itens apareceram como passíveis de aperfeiçoamento: (a) segmentação sentencial (passo essencial para o processo da sumarização), (b) reconhecimento e tratamento da estrutura textual e (c) detecção de mais de um tópico textual. Em particular, em relação ao item (b), em alguns exemplos analisados, evidenciamos que o sumário gerado era prejudicado em função da divisão de seu conteúdo em seções. Verificou-se que, tratando-se cada seção separadamente, os sumários 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. Com isso, o item (c) anterior é tratado também. Adicionalmente, para lidar com a questão da segmentação, acoplou-se ao GistSumm o segmentador SENTER [2].

3. Resultados e Discussão

Após a realização destas modificações, obteve-se como resultado uma maior aproximação da idéia principal dos textos que continham mais de uma seção. O uso do SENTER levou o sistema a identificar melhor as sentenças do texto, acarretando, portanto, melhor qualidade do sumário. Evidenciou-se também, de forma subjetiva, que o sistema é especialmente mais eficaz na sumarização de artigos científicos. Como próximo passo, uma avaliação mais robusta dessa nova versão do sistema deve ser realizada.

4. Conclusões

A análise de alguns textos que contêm mais de uma seção mostrou a melhoria de seus sumários. A nova característica foi incorporada ao sistema e gerou uma nova versão que está disponível na Internet. Após essas modificações realizadas no sistema, foram levantadas outras características que resultariam em melhores sumários. Dentre elas, a sugestão de que o sistema utilize o título da seção para a produção de um sumário mais centrado no assunto e a sua utilização específica para sumarização de dissertações e teses de pós-graduação.

5. 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] Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.