Sumarização Automática para Simplificação de Textos

Paulo Rodrigues Alves Margarido, Thiago Alexandre Salgueiro Pardo

Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

1. Objetivos

O acesso à informação tem se tornado cada vez mais abrangente, de modo que qualquer pessoa conectada à Internet pode obter todas as informações que necessita a partir de qualquer computador. Porém, apenas o acesso à informação não é suficiente para realizar a inclusão digital de todos, pois pessoas com nível baixo de alfabetismo podem não compreender um texto encontrado *on-line* [1]. Portanto, além de disponibilizar o conteúdo, é importante também que se adapte este conteúdo ao nível de letramento do usuário. Este trabalho tem por objetivo avaliar a contribuição da sumarização automática de textos para fins de inclusão digital.

2. Métodos

A sumarização automática é uma das subáreas clássicas do Processamento de Línguas Naturais (PLN) e consiste em gerar um resumo a partir de um dado texto sem o auxílio de um ser humano. A literatura para métodos de sumarização é abundante, uma vez que diversos pesquisadores já trabalharam com este problema de vários modos, com graus variados de complexidade e de sucesso [3].

Este trabalho consistiu em duas etapas: a primeira visou determinar um método de sumarização aplicável a tarefas de simplificação; a segunda em um experimento piloto com o público-alvo deste trabalho (analfabetos de diversos níveis de letramento) para determinar o impacto de diversas estratégias de simplificação via sumarização, por exemplo, destacando a sentença principal no texto original, destacando algumas sentenças, ou apresentando apenas o sumário.

3. Resultados

A partir da análise de vários métodos de sumarização (alguns clássicos, alguns do estado da arte), foi possível determinar qual deles seria mais adequado para a tarefa em questão, baseando-se em 4 critérios: (i) a precisão do método em selecionar a sentença principal de um texto; (ii) a informatividade dos sumários gerados (baseado na medida ROUGE [2]); (iii) disponibilidade do sistema; (iv) facilidade de uso do sistema.

O experimento piloto realizado com 19 pessoas do público-alvo do trabalho deu indícios de que diferentes estratégias de apresentação da sumarização auxilia pessoas de diferentes escolaridades de modos diferentes, mas em geral as pessoas consideraram que informações do sumário ajudam no entendimento do texto.

4. Conclusões

Este trabalho dá indícios de que é possível utilizar a sumarização automática para auxiliar no trabalho de inclusão digital da população menos favorecida, utilizando métodos portáveis e eficientes de sumarização. Além disso, os resultados do experimento com o público alvo também são encorajadores. No futuro, experimentos em maior escala podem confirmar essa possibilidade.

5. Agradecimentos

Este trabalho contou com o apoio da FAPESP e da Microsoft Research.

6. Referências Bibliográficas

[1] Ribeiro, V. M.: Analfabetismo e alfabetismo funcional no Brasil. Boletim INAF. São Paulo: Instituto Paulo Montenegro (2006)

[2] Lin, C., Hovy, E.H. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of Language Technology Conference (2003)

[3] Mani, I. Automatic Summarization. John Benjamins Publishing Co. Amsterdam (2001)