Exploração de Métodos Linguísticos para Sumarização Automática Multidocumento

Guilherme Gonçalves, Thiago A. S. Pardo

Núcleo Interinstitucional de Linguística Computacional – NILC
Instituto de Ciências Matemáticas e de Computação – ICMC
Universidade de São Paulo – USP
guilherme3.goncalves@usp.br, taspardo@icmc.usp.br

Objetivos

A sumarização Automática Multidocumento (SAM) consiste na produção automática de um único sunário (também chamado resumo) a partir de um grupo de textos sobre um mesmo tópico ou sobre tópicos relacionados (Mani, 2001). A SAM seleciona as informações potencialmente mais relevantes identificando e excluindo informações desnecessárias e menos importantes, evitando que se perca tempo na leitura de uma grande quantidade de documentos. Nesse projeto, são explorados métodos linguísticos para a SAM. Tais métodos linguísticos foram identificados por Camargo (2013) em seus estudos, com base em análise de córpus. O intutito dessa exploração é produzir um prototipo que seja capaz de aplicar os métodos de Camargo de forma automática, gerando como saida sumários multidocumento melhores, mais informativos para o leitor.

Métodos e Procedimentos

Primeiramente estudou-se o trabalho de Camargo (2013), que foca na seleção de conteúdo do sumário por meio de conhecimento linguístico. Camargo desenvolveu um conjunto de regras linguísticas, com base em atributos textuais, para indicar que sentenças dos textos são mais relevantes. Em termos de implementação, inicialmente, neste trabalho, foi necessário realizar a representação computacional dos textos-fonte, textos de onde as informações são extraídas. Foi possível, então, aplicar os métodos linguísticos de forma automática. Com isso, faz-se a seleção do conteúdo relevante para formar o sumário. Após selecionar o conteúdo, é necessário retirar informações que se repetem, ou seja, aquelas informações que são redundantes. Gera-se, por fim, a representação em língua natural do conteúdo selecionado, que, neste caso, é a língua portuguesa.

Resultados

O desenvolvimento do protótipo esta em andamento. Atualmente, o protótipo realiza a representação interna dos textos-fonte, aplica os métodos linguísticos de forma automática, e seleciona o conteúdo para o sumário norteado pelos métodos linguísticos aplicados. A próxima etapa, que está em andamento, consiste na remoção de informações redundantes. Posteriormente, pretende-se avaliar o sistema, por meio do nível de informatividade e semelhança com os sumários multidocumento produzidos por humanos.

Conclusões

Uma ferramenta que realize a sumarização multidocumento de textos seria interessante pelo fato de poupar tempo para a assimilação de conhecimento por parte do leitor. No entanto, a ferramenta em questão, que no caso utiliza métodos linguísticos, está em desenvolvimento e precisa ser aperfeiçoada. Esta depende ainda de outros softwares para a realização de seus métodos, com isso, o desenvolvimento desses softwares base para o protótipo influenciam fortemente no desempenho do protótipo.

Referências Bibliográficas

Camargo, R. T. (2013). *Investigação de estratégias de sumarização humana multidocumento*. Dissertação de Mestrado. Universidade Federal de São Carlos.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co. Amsterdam.