Exploração de Técnicas de Aprendizado de Máquina para Sumarização de Textos Intrasentencial

Daniel Kawamoto, Thiago A. Salgueiro Pardo
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

1. Objetivos

Propõe-se o estudo e a implementação de técnicas de Sumarização Automática (SA), área que atua no desenvolvimento de geração de sumários (resumos) através de métodos computacionais, visando o aperfeiçoamento de técnicas existentes, a exploração de novos caminhos para a área e o desenvolvimento de ferramentas. Em particular, visa-se à sumarização intrasentencial de textos pelo uso de técnicas de aprendizado de máquina.

2. Material e métodos

O foco deste trabalho é estudar técnicas de sumarização intrasentencial, que consiste em compactar as sentenças, mantendo a informação principal e a coerência do texto [1], por exemplo, removendo-se partes da sentença que afetam minimamente seu conteúdo. Prevê-se tal estudo com textos jornalísticos, os quais se caracterizam pela linguagem do dia a dia.

Inicialmente, utiliza-se o parser PALAVRAS [2], o melhor analisador sintático automático para o português do Brasil, para anotar com informações sintáticas as sentenças dos textos do estudo. A partir desses dados, extraem-se para cada palavra de cada sentença uma série de informações, como classe morfossintática, função sintática, posição na sentença, etc. Para completar, extraem-se também informações estatísticas para cada palavra a partir da análise realizada pelo sistema GistSumm [3], um dos primeiros sumarizadores que surgiu para o português e que ainda é bastante utilizado.

Após a tarefa anteriormente descrita, as informações encontradas são utilizadas para construir uma tabela atributo-valor no formato do WEKA, uma ferramenta formada por um conjunto de implementações de algoritmos de aprendizado de máquina. Com essa ferramenta, pretende-se induzir regras que predigam que partes de uma sentença podem ser excluídas em um processo de sumarização.

A Figura 1 ilustra a arquitetura do projeto descrito.

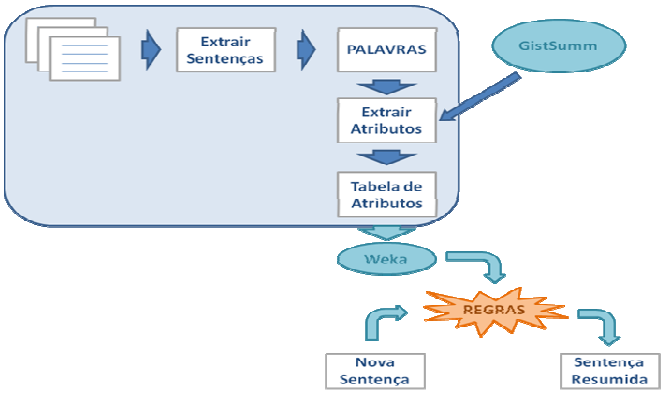

Figura 1: Arquitetura

3. Resultados esperados

O método de sumarização descrito ainda está em fase de implementação e teste. Espera-se obter regras de sumarização intrasentencial (induzidas pelo aprendizado de máquina) e, consequentemente, um novo sumarizador para a língua portuguesa.

4. Conclusões

Até onde se sabe, este trabalho consiste na primeira abordagem de sumarização intrasentencial para o português do Brasil com o uso de aprendizado de máquina.

5. Referências bibliográficas

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

[2] Eckhard B. The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)

[3] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. 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 (2003)