

Ferramenta para Apoio à Produção Textual baseada na Extração de Conceitos

Rafael Finkelstein¹, Prof. Dr. Eliseo Berni Reategui²

¹ Autor, Engenharia da computação, UFRGS

² Orientador



UFRGS PROPESQ XXV SIC
Salão Iniciação Científica

CH - Ciências Humanas

Introdução

A ferramenta desenvolvida pelo grupo de pesquisa do referido prof., a ferramenta SOBEK, é capaz de identificar conceitos relevantes em um texto a partir da análise da frequência dos termos nesse material textual (conversas em ferramentas de interação). Pode ser usada, portanto, para extrair representações concisas de textos diversos, bem como das interações entre participantes num ambiente virtual.

Palavras de determinadas classes gramaticais não possuem relevância na geração do grafo (e.g. verbos e artigos), e frequentemente gerando nodos irrelevantes que diminuem a precisão das relações entre as palavras, externadas no grafo.

A ferramenta TreeTagger foi implementada para remover tais nodos, com pouco ou nenhum significado relevante para o algoritmo de geração do grafo.

Ainda visando a maior precisão do grafo gerado, o banco de sinônimos *Thesaurus* foi implementado para possibilitar a fusão de sinônimos, muito recorrentes em textos de alto nível.

Em outro projeto da bolsa foi implementado o cliente e servidor de um chat com a atuação de um agente pedagógico interativo que possui várias ações dentre elas, consultas ao google, solicitações de tradução e verificação ortográfica.

Atividades desenvolvidas e resultados obtidos

Pesquisa e implementação da ferramenta *TreeTagger*

O funcionamento da ferramenta é simples. Uma vez instalado, o *TreeTagger* recebe uma entrada de uma *string* (sequência de caracteres) com o texto que se deseja etiquetar e retorna uma *string* para cada palavra no formato *token tag lemma*, por exemplo no texto “o meu tio é português”:

token tag lemma
o DET o
meu ADJ meu
tio NOM tio
é V ser
português ADJ português

Também foi anexada a interface gráfica já existente do SOBEK a opção de escolher quais classes de palavras serão removidas (disponível em português e Inglês) da geração do grafo, como é mostrado na figura 1.

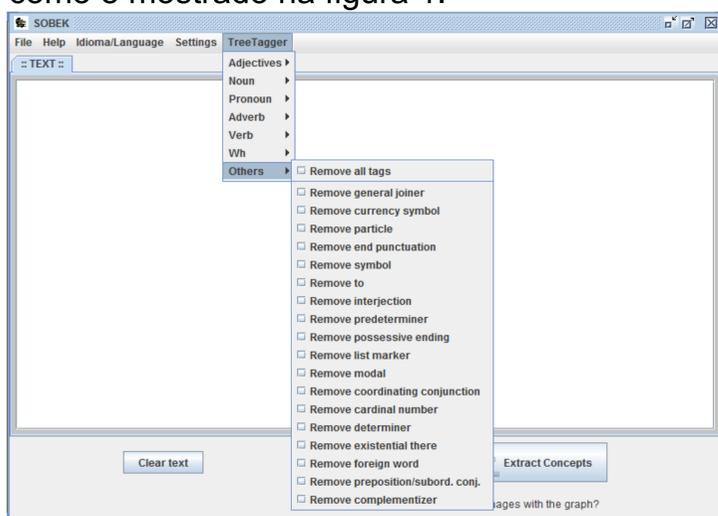


Figura 1: Interface gráfica da implementação da ferramenta TreeTagger na aplicação SOBEK



Figura 2: Interface gráfica do cliente da aplicação de chat

A aplicação *Thesaurus*, tal como o *TreeTagger* foi implementada na linguagem java, utilizando APIs de uso livre para utilização acadêmica. O funcionamento da ferramenta consistem em, quando acionada, verificar se dois conceitos são sinonimos, em caso afirmativo estes dois nodos do grafo são fundidos mantendo o sinônimo de maior frequência no texto.

Por último, como ilustrado pela figura 2 está o chat implementado usando websockets, para comunicação com o servidor, podendo ser acessado online por qualquer *browser*. A interação com o agente pedagógico e dada pro botões que acionam *pop-ups* com as devidas solicitações.