Investigação e Desenvolvimento de uma Ferramenta de Detecção Automática de Neologismos da Língua Portuguesa

Daniel Feitosa, Sandra Maria Aluísio, Thiago A. S. Pardo

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

1. Objetivos

Na área de Processamentos de Línguas Naturais (PLN), um dos recursos básicos para o desenvolvimento e bom desempenho da maioria dos softwares é o léxico [1]. O desenvolvimento de léxicos computacionais, como os utilizados nas aplicações, é um processo incrivelmente caro e demanda muito tempo, por exigir trabalho manual de especialistas, que, em geral, são lingüistas de formação. É consenso também que, com o passar do tempo, novas palavras surgem – os neologismos – e outras entram em desuso. Portanto, além do alto custo para criação, existe também um alto custo para manter estes léxicos atualizados para que não sejam descartados. Este trabalho tem por objetivo auxiliar esta tarefa tornando-a o menos complexo possível para que o tempo gasto com tal tarefa seja investido na eficiência dos aplicativos ou produção de outros.

2. Métodos

A identificação computacional de neologismos é uma área de pesquisa que foi pouco explorada até hoje. Portanto, não existe um método definido para tal processo. A literatura traz alguns métodos para encontrar neologismos formais [2] e propostas de ferramentas de um propósito mais geral [3], além de toda base lingüística pertinente ao assunto.

Este trabalho consiste em três etapas: a primeira visa estudar a literatura existente a fim de formar uma base teórica para a construção da ferramenta; a segunda em investigar e construir uma arquitetura de sistema para o propósito do trabalho; a terceira em implementar o sistema e construir uma interface amigável tendo em vista que parte considerável dos usuários finais serão lingüistas e que, portanto, os resultados devem ser apresentados deforma clara e intuitiva.

3. Resultados

A partir da análise dos métodos e propostas encontradas foi possível desenvolver uma arquitetura para a ferramenta, onde se identificou a divisão desta em três fases distintas: pré-processamento, processamento principal (com manutenção de uma base de dados de neologismos) e pós-processamento. A segunda fase foi desenvolvida com base em busca e identificação de padrões a partir de medidas estatísticas. Na primeira e terceira fase, é onde ocorre interação com o usuário, para entrega do texto e apresentação dos resultados. A terceira fase será mais complexa, pois, na apresentação dos resultados, os neologismos que forem confirmados devem ser classificados pelo usuário para alimentar a base.

4. Conclusões

O projeto encontra-se atualmente na terceira etapa. Até o presente momento é possível afirmar que este trabalho conseguirá cumprir com o objetivo. Porém, o processo não será completamente automático, pois gerará uma lista de palavras candidatas da qual caberá ao usuário dizer quais são neologismos e em seguida classificá-los.

5. Referências Bibliográficas

[1] Palmer, M. Multilingual resources, multilingual information management: current levels and future abilities. (2001)

[2] REY, Alain. Néologisme: un pseudoconcept? (1976)

[3] Janicijevic, T. and Derek W. NeoloSearch: Automatic detection of neologisms in French Internet documents. (1997)