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

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

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

fdaniel@grad.icmc.usp.br, sandra@icmc.usp.br, taspardo@icmc.usp.br

1. Introdução

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.

2. Objetivos

Este trabalho tem por objetivo auxiliar a tarefa de atualização do léxico, tornando-a o menos complexa possível para que o tempo gasto com tal tarefa seja investido na eficiência dos aplicativos ou produção de outros. Uma forma de atualizar o léxico é incorporar a ele neologismos, isto é, palavras antigas com novos sentidos ou palavras completamente novas.

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 de forma clara e intuitiva.

3. Resultados

A partir da análise dos métodos e propostas encontradas foi possível desenvolver a arquitetura mostrada na Figura 1 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.

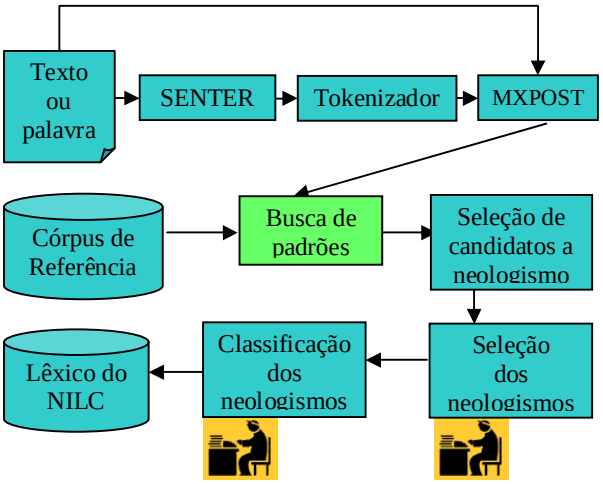

Figura 1. Arquitetura da ferramenta

A primeira fase é marcada pela aquisição da entrada e pela preparação desta para o processamento principal. São aplicados filtros que irão segmentar sentencialmente, tokenizar e colocar etiquetas morfossintáticas no texto.

A segunda fase foi desenvolvida com base em busca e identificação de padrões a partir de medidas estatísticas. Inicialmente tem-se um conjunto de padrões que podem ser possíveis candidatos a neologismo, daí faz-se uma mineração no texto atrás desses padrões, contando também a freqüência com que ocorrem. No resultado da mineração aplica-se um corpus de exclusão que possui os neologismos já encontrados. Em seguida são aplicadas medidas estatísticas e é feito um estudo em cima deste resultado extraindo-se assim os reais candidatos.

A terceira fase é a que possui maior interferência humana, pois é o usuário que irá criar a lista de neologismos a partir da identificação destes na lista de candidatos gerada pela ferramenta. Posteriormente também preencherá uma ficha para cada neologismo com alguns aspectos sobre a classificação deste. Por fim os neologismos encontrados serão armazenados no léxico do NILC, e serão utilizados nas próximas buscas por neologismos.

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, ocorrerá a classificação dos neologismos como já descrito acima.

4. Conclusões

O projeto encontra-se atualmente na terceira etapa. Até o presente momento é possível afirmar que este trabalho conseguirá cumprir seu 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.

Espera-se que esta ferramenta tenha um bom desempenho para que possa ser aplicada de imediato em auxilio a outros estudos feitos pela USP.

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)