Apoio Computacional para Tarefas Linguísticas em Mineração de Opinião

Angelo A. R. Tessaro, Thiago A. S. Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)

Universidade de São Paulo

angeloartessaro@usp.br , taspardo@icmc.usp.br

Objetivos

O objetivo deste trabalho é apoiar computacionalmente o processamento linguístico no contexto de Mineração de Opinião. Isso permite a realização de análises linguísticas em larga escala, que seriam inviáveis apenas com trabalho manual, além de possibilitar a obtenção de dados mais precisos e com mais profundidade, por exemplo na análise de sentimentos em textos opinativos.

Métodos e Procedimentos

Foram estudados artigos científicos a respeito do Processamento de Linguagem Natural e identificadas ferramentas a serem usadas no trabalho descrito. Com o uso das bibliotecas NLTK (Bird, 2006), NLPNet (Fonseca and Rosa, 2013) e PALAVRAS (Bick, 2000) em Python para processamento de texto foi criado um script para a extração de atributos linguísticos (número de caracteres, número de verbos, número de StopWords, etc.) de uma coletânea de “tweets” com teor satírico.

Posteriormente também foi feita a extração de colocações gramaticais em inglês de um arquivo e utilizando o framework Selenium realizou-se a automação da pesquisa dessas colocações e a extração da frequência em que ocorriam no corpus BNC (British National Corpus).

Além disso, por meio de um script em Python e utilizando como referência um léxico UNITEX (Muniz, 2004) do NILC, foi feita a lematização (deflexionar cada uma das palavras para que todas as frequências fossem contabilizadas em um só item na forma canônica) de um conjunto de listas geradas pelo AntConc (Anthony, 2006).

Com o ROUGE (Lin, 2004) comparou-se sumários criados pelo computador com sumários feitos por humanos e os resultados foram organizados em tabelas com medidas de Precisão, Recall e f-Measure.

Por fim, com base no dicionário LIWC em português foram atribuídas polaridades às palavras de textos satíricos. Para o estudo em questão utilizamos apenas as categorias que sugerem emoções Positivas ou Negativas.

Resultados

Foram alcançados os objetivos esperados para cada uma das tarefas realizadas, dadas as limitações impostas pelas ferramentas utilizadas, como por exemplo o dicionário LIWC que não consegue atribuir com tanta precisão a polaridade das palavras.

Conclusões

As atividades desenvolvidas tornam evidentes as vantagens do apoio computacional em tarefas linguísticas, pelo volume de informação que pode ser processado frente a realização manual. Apesar disso, o Processamento de Linguagem Natural ainda tem muitas limitações a serem superadas para se aproximar cada vez mais da compreensão humana.

Referências Bibliográficas

Chin-Yew Lin (2004). ROUGE: A Package for Automatic Evaluation of Summaries.

Laurence Anthony (2006). Concordancing with AntConc: An introduction to tools and techniques in corpus linguistics.

Marcelo C.M. Muniz (2004). A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB.