Comparação de Técnicas de Normalização Morfológica na Análise de Similaridade Textual

Guilherme Passero[1,2], Rudimar Luís Scaranto Dazzi[2], Aluizio Haendchen Filho[1]

[1]Núcleo de Inteligência Artificial e Sistemas Inteligentes (NIASI)
Centro Universitário de Brusque (UNIFEBE) – Brusque, SC – Brasil

[2]Laboratório de Inteligência Aplicada (LIA)
Universidade do Vale do Itajaí (UNIVALI) – Itajaí, SC – Brasil

guilherme.passer@gmail.com, rudimar@univali.br, aluizioh@terra.com.br

Abstract. *Morphological normalization is an important part of Natural Language Processing Systems and aids the detection of similar terms in texts. This paper presents an approach for comparison of morphological normalization techniques in the context of text similarity analysis. This proposal regards similarity analysis in two levels: lexical and semantic. We suggest the use of a recent workshop dataset with 10.000 pairs of sentences and its human assigned similarity index for performance analysis of these techniques. This proposal may be adapted to assist natural language processing projects choice of morphological normalization approach.*

1. Introdução

Na etapa de pré-processamento textual de Sistemas de Processamento de Linguagem Natural – SPLNs – é comum o uso de operações de normalização morfológica como *stemming*, lematização e extração de n-gramas. Essas operações auxiliam o reconhecimento de termos conceitualmente similares, onde palavras como “menino”, “meninos”, “meninas” e “menininhos”, por exemplo, são reduzidas ao radical “menin” por um *stemmer*, ou ao lema “menino” por um lematizador.

A análise de similaridade textual tem várias aplicações em SPLNs, como em sistemas de detecção de plágio, recuperação de documentos, sumarização de textos etc. Haja vista a importância da normalização morfológica em SPLNs e o grande número de algoritmos disponíveis, são necessários estudos para avaliar sua eficácia em cada cenário, como em [Balakrishnan e Ethel 2014] e [Guimarães, Meirose e Moraes 2015].

Nesse contexto, este trabalho apresenta uma proposta para comparação de técnicas de normalização morfológica considerando o seu impacto na análise de similaridade textual.

2. Solução Proposta

Uma revisão do estado da arte sobre análise de similaridade textual na língua portuguesa é necessária para indicar todas as técnicas de normalização morfológica atualmente aplicadas e, portanto, sujeitas à comparação. Haja vista a ausência na literatura de uma revisão com esse enfoque, sugere-se inicialmente as técnicas comparadas por [Guimarães, Meirose e Moraes 2015] na tarefa de classificação de textos: (i) um

stemmer para o português baseado em regras de substituição/remoção de caracteres; (ii) um lematizador probabilístico; e (iii) uma abordagem de extração de n-gramas em nível de caractere.

O workshop de Avaliação de Similaridade Semântica e Inferência Textual – ASSIN – da Conferência Internacional de Processamento Computacional da Língua Portuguesa de 2016 – PROPOR – disponibilizou um banco com 10.000 pares de sentenças anotadas por humanos em relação ao seu índice de similaridade e vínculo textual (*textual entailment*) [Fonseca *et al.* 2016]. Esse conjunto de dados será utilizado no estudo comparativo proposto, onde serão aplicadas técnicas de análise de similaridade textual e observado o índice de similaridade computado em relação ao atribuído por um humano.

O desempenho das operações de normalização morfológica testadas, combinadas às técnicas de análise de similaridade textual, será obtido através das medidas adotadas no workshop ASSIN: coeficiente de correlação de Pearson e erro médio ao quadrado (*mean squared error*). Em seguida, os resultados serão comparados para verificar a influência da abordagem de normalização morfológica escolhida em relação às técnicas de análise de similaridade aplicadas.

No momento está em andamento um estudo de caso onde serão avaliadas as seguintes técnicas para cada nível de análise: (i) léxico: distância Levenshtein entre cadeias de caracteres; e (ii) semântico: abordagem Solo Queue apresentada em [Hartmann 2016]. Atualmente existem diversos modelos para análise de similaridade textual. Desse universo, selecionamos a técnica (i) por sua simplicidade e uso recorrente na análise de similaridade léxica e a técnica (ii) pelo resultado vencedor no workshop ASSIN para o português brasileiro [Fonseca *et al.* 2016].

3. Considerações Finais

Aqui se propôs uma abordagem para comparar o impacto de diferentes técnicas de normalização morfológica na tarefa de análise de similaridade textual. Os resultados a serem obtidos no experimento proposto podem variar de acordo com o conjunto de dados de teste, com isso esse experimento não visa estabelecer a melhor técnica de normalização morfológica disponível. Não obstante, a abordagem proposta pode ser adaptada para auxiliar na escolha de técnicas de normalização morfológica em projetos futuros de SPLNs.

As técnicas avaliadas podem ter custo computacional significativamente diferentes, assim sugerimos que futuramente seja avaliada também a relação entre o custo e eficácia.

Referências

Balakrishnan, V. e Ethel, L.-Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. Lecture Notes on Software Engineering, v. 2, n. 3, p. 262–267.

Fonseca, E. R., Santos, L. B., Aluísio, S. M. e Criscuolo, M. (2016). ASSIN: Avaliação de Similaridade Semântica e Inferência Textual. PROPOR – International Conference on the Computational Processing of Portuguese.

Guimarães, G. T., Meirose, M. V. e Moraes, S. M. W. (2015). n-Gramas de Caractere como Técnica de Normalização Morfológica para Língua Portuguesa: Um Estudo em Categorização de Textos. Proceedings of Symposium in Information and Human Language Technology, p. 211–220.

Hartmann, N. S. (2016). Solo Queue at ASSIN: Combinando Abordagens Tradicionais e Emergentes. PROPOR – International Conference on the Computational Processing of Portuguese.