Análise Discursiva aplicada à Mineração de Opinião: Classificação de Subjetividade para a Língua Portuguesa

Luana B. Belisário, Thiago A. S. Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

luana.belisario@usp.br, taspardo@icmc.usp.br

Objetivos

Neste resumo estendido, relatamos a continuidade de um trabalho anterior de classificação automática de subjetividade para o português [1], que visa distinguir sentenças que exprimem opiniões (ditas "subjetivas") daquelas que são factuais ("objetivas"). Em particular, combinamos o modelo discursivo RST (*Rhetorical Structure Theory*) [2] com outros atributos em métodos de aprendizado de máquina, visando melhorar a classificação de subjetividade obtida e explicitar a contribuição do conhecimento discursivo fornecido pela RST.

Métodos e Procedimentos

Para a modelagem da tarefa como aprendizado de máquina, agrupamos os atributos em três categorias: lexicais, baseados em medidas de centralidade de grafos e discursivos. Enquanto as duas primeiras categorias são baseadas no trabalho descrito em [1], os atributos discursivos são obtidos da representação fornecida pela RST, que estrutura o conteúdo textual em uma árvore, onde os nós folha representam os segmentos textuais e os nós internos indicam as relações que ocorrem entre os segmentos, além de lhes atribuir importância, distinguindo "núcleos" (informação essencial) e "satélites" (complementares).

Os atributos lexicais incluem características gramaticais dos textos, por exemplo, proporção de advérbios de negação e intensidade presentes na sentença e a proporção de palavras subjetivas. Os atributos de centralidade são obtidos de três medidas de centralidade: *Eigenvector centrality*, *Katz index*, e *PageRank*. A maioria dos atributos discursivos representa marcadores booleanos da presença ou ausência das diferentes relações do modelo RST.

Os métodos foram treinados e testados (com *3-fold cross-validation*) em um córpus de resenhas de livros, composto por 350 sentenças manualmente rotuladas como subjetivas ou objetivas e sabidamente desafiador para a área.

Resultados

Verificou-se que, conforme os grupos de atributos vão sendo incorporados ao processo, os resultados melhoram. Utilizando-se todos os atributos, foi possível atingir 77,1% de acurácia geral com o uso de SVM (dentre muitas outras técnicas avaliadas). Além disso, em uma análise qualitativa, notou-se que há relações discursivas que sinalizam a subjetividade. Por exemplo, sentenças subjetivas costumam conter relações de *justify*, *concession* e *contrast*, enquanto as objetivas têm *sequence*, *circumstance* e *cause*.

Conclusões

Mostramos as contribuições de diferentes tipos de atributos para a classificação de um conjunto de dados desafiador. Como trabalho futuro, prevê-se a exploração de outros modelos discursivos para a tarefa.

Referências Bibliográficas

[1] Belisário, L.B.; Ferreira, L.G.; Pardo, T.A.S. (2020). Evaluating Richer Features and Varied Machine Learning Models for Subjectivity Classification of Book Review Sentences in Portuguese. Information, Vol. 11, N. 9, pp. 1-14.

[2] Mann, W.C. and Thompson, S.A. (1987). Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190.