Abstract

Sentiment analysis refers to the task of natural language processing to determine whether a piece of text contains some subjective information and what subjective information it expresses, i.e., whether the attitude behind this text is positive, negative or neutral. Understanding the opinions behind user-generated content automatically is of great help for commercial and political use, among others. The task can be conducted on different levels, classifying the polarity of words, sentences or entire documents.

In this thesis, we propose and investigate a method of elementary discourse unit (EDU) level sentiment analysis using discourse features. Following prior work, we hypothesize that when we want to predict the sentiment of a certain EDU, we can use the sentiments of other EDUs which stand in some discourse relation with the current one. For example, a Contrast relation is likely to signal that the sentiments of two arguments of this relation are different. Once we know one of the sentiments, we can try to use this information to predict the other one. Some discourse theories have been applied in this context, but no prior work compares relative influence of different discourse theories on the performance of sentiment analysis. To the best of our knowledge, this is the first work comparing different discourse theories in a principled way in the task of sentiment analysis, making use of state-of-the-art discourse parsers.

To discover the discourse relations we employ two discourse parsers based on two different discourse theories, respectively Rhetorical Structure Theory (RST) and the The Penn Discourse Treebank (PDTB). We propose several models to represent the discourse structures and test them on the task of EDU-level sentiment prediction. In the discourse representations the relations are determined by the discourse parsers and sentiment values of connected EDUs are taken from our gold-standard data. We also implemented two baselines using lexical features and features concerning adjacent EDUs. The classification accuracy increases if discourse features are added, and RST-based discourse features outperforms PDTB-based discourse features.