Abstract

One of the challenges for sentiment analysis is the presence of sarcasm. Sarcasm is a form of speech that generally implies a bitter remark toward another person or thing expressed in an indirect or non-straightforward manner. The presence of sarcasm can potentially flip the sentiment of the entire sentence or document, depending on its usage. A sarcasm detector has been developed using sentiment patterns, world knowledge, and context in addition to features that previous works used, such as frequencies of terms and patterns. This sarcasm detector can detect sarcasm on two different levels: sentence-level and document-level. Sentence-level sarcasm detection incorporates basic syntactical features along with world knowledge in the form of a ResearchCyc Sentiment Treebank, which has been created for this project. Document-level sarcasm detection incorporates context by using the sentiments of sequential sentences in addition to punctuation features that occur throughout the entire document.

The results obtained by this sarcasm detector are considerably better than random guessing. The highest F1 score obtained for sentence-level sarcasm detection is 0.687 and the highest F1 score obtained for document-level sarcasm detection is 0.707. These results imply that the features used for this project are useful for sarcasm detection. The pattern features used for sentence-level detection work well. However, the results from the usage of the ResearchCyc Sentiment Treebank on the sentence-level compared to the results without this treebank are approximately the same, partially due to the fact that this treebank has been built off of Stanford's CoreNLP treebank, which includes a limited set of words. Document-level detection indicates that context is an important factor in sarcasm detection. This thesis provides insight to areas that were not previously thoroughly explored in sarcasm detection and opens the door for new research using world knowledge and context for sarcasm detection, sentiment analysis, and potentially other areas of natural language processing.