Humour Recognition

-Ayush Jhunjhunwala
(201364162)
Under the guidance of Dr. Radhika Mamidi

Aim

- Use automatic classification techniques to distinguish between humorous and non-humorous text computationally.

- An analysis of the content-based features identified during the process to point out some of the most useful features for future studies of humour.

Dataset

- We restrict our investigation to the type of humour found in one-liners.
- Humorous Data: a set of about 16,000 one-liners were used in the humour-recognition experiment.

  Eg: Progress goes so fast that you have to run like everything to stay where you are.

- Non-humorous Data: a set of about 36,165 one-liners author quotes were used for the experiment.

  Eg: A. A. Milne - Did you ever stop to think, and forget to start again?

Method

1. Identification of humour-specific stylistic features:

   a. **Alliteration**: We identify and count the number of alliteration/rhyme chains in each example in our dataset. The chains are automatically extracted using an index created on the top of the CMU pronunciation dictionary.

   b. **Antonymy**: The lexical resource we use to identify antonyms is WORDNET, and in particular the antonymy relation among noun, verbs adjectives, and adverbs.

   c. **Adult slang**: To form a lexicon required for the identification of this feature, we have extracted all the slang words from the website - Daily Loaf. Next, we check for the presence of the words in this lexicon in each sentence in the corpus.

Method (cont.)

2.   A feature vector was made using the above three stylistic features for each sentence in the corpus.

3.   **Support Vector Machines (SVM)** was used to classify the humorous and non-humorous data using the above feature vector and the dataset as input.

4.   All evaluations were performed using 10-fold cross-validations, for accurate estimates.

Results

After 5-fold cross-validation,

Accuracy: 69.90%

After 10-fold cross-validation,

Accuracy: 69.91%

Here accuracy is the percentage of correctly classified one-liners in the complete dataset.

Challenges

- To identify a set of features that were both significant and feasible to implement using existing machine-readable resources.

- Variation of classification performance with respect to data-size.

- To improve the accuracy of the results by identifying more sophisticated features.

Conclusion

- The result obtained in the automatic classification experiments prove that computational approach is viable solution for the task of humour recognition.