INTRODUCTION

For developing a data-driven text rewriting algorithm for paraphrasing, it is essential to have a monolingual corpus of aligned paraphrased sentences. We make an effort to collect Dutch paraphrases from news article headlines in an unsupervised way to be used in future paraphrase generation. News article headlines are abundant on the web, and are already grouped by news aggregators such as Google News. These services collect multiple articles covering the same event. Crawling such news aggregators is an effective way of collecting related articles which can straightforwardly be used for the acquisition of paraphrases. Two methods of automatically aligning headlines to construct such an aligned corpus of paraphrases are compared: one based on clustering, and the other on pairwise similarity-based matching. We show that the latter performs best on the task of aligning paraphrastic headlines.

METHOD

1. CLUSTERING

- TF-IDF representation
- k-Means clustering
- Clustering solution
- Clustering stopping algorithm

2. PAIRWISE SIMILARITY

- TF-IDF representation
- Pairwise Similarity Matching
- Between thresholds: repeat with article text snippet
- > Upper threshold
- < Lower threshold

RESULTS

- The 825 clusters in the test set contain 1,751 sub-clusters in total. In these sub-clusters, there are 6,685 clustered headlines. Another 3,123 headlines remain unclustered.
- For the k-means similarity method, optimizing using an \( F\_1 \) score, the optimum values for the lower and upper threshold are \( T\_\text{lower} = 0.2 \) and \( T\_\text{upper} = 0.5 \).
- Using pairwise similarity on the 30,000 headline clusters results in roughly 200,000 Dutch paraphrase pairs.

DISCUSSION & FUTURE WORK

- The data we use was acquired by crawling Google News in the period of April-August 2006, resulting in roughly 13,000 Dutch headline clusters.
- 865 of these clusters were subclustered manually as part of the DAESO project (Marsi and Krahmer, 2007). These annotated clusters are used for optimizing our system.
- We aim for high precision rather than high recall: we want a high quality paraphrase corpus. We evaluate the number of correct alignments, and optimize using an \( F\_1 \) score with \( \beta = 0.25 \).
- For the clustering method, the CLUTO clustering package is used and the PK1 cluster-stopping algorithm by Pedersen and Kulkarni (2006) to determine the correct number of sub-clusters within each cluster.
- For the pairwise similarity-based matching we use a cosine similarity function. We adopt two thresholds; if the similarity exceeds the upper threshold, the paraphrase pair is accepted. If it is below the lower threshold, the sentences are not considered paraphrases. When it is between the two thresholds, the procedure is repeated but this time using a text snippet taken from the beginning of the article.

REFERENCES


Acknowledgments

Thanks are due to the Netherlands Organization for Scientific Research (NWO) and to the Dutch HLT Stevin programme. Thanks also to Wouter Booms for originally mining the headlines from Google News. For more information on DAESO, please visit daeso.uvt.nl.