Clustering and Matching Headlines for Automatic Paraphrase Acquisition

Sander Wubben, Antal van den Bosch, Emiel Krahmer and Erwin Marsi

Tilburg centre for Creative Computing
Tilburg University
The Netherlands
{s.wubben,antal.vdnbosch,e.j.krahmer,e.c.marsi}@uvt.nl

INTRODUCTION

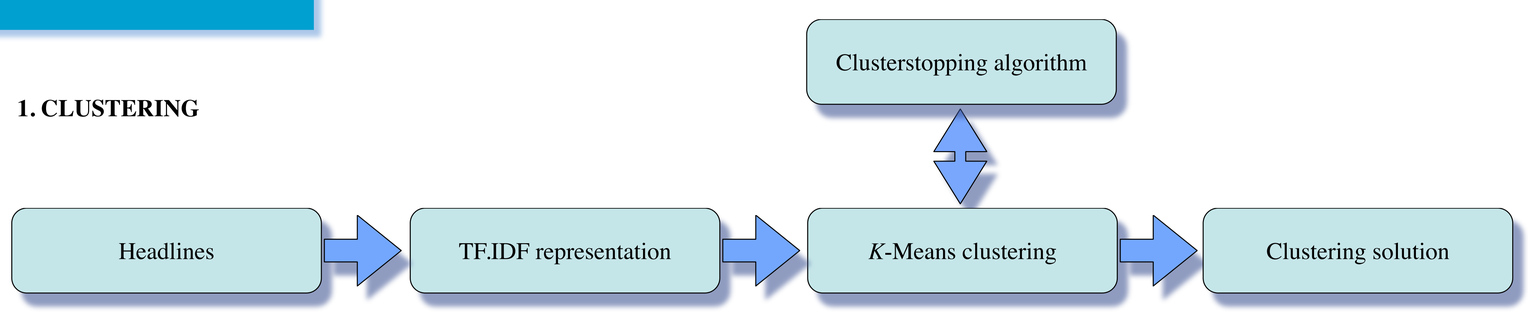
For developing a data-driven text rewriting algorithm for paraphrasing, it is essential to have a monolingual corpus of aligned paraphrased sentences. We make an effort to collect Dutch paraphrases from news article headlines in an unsupervised way to be used in future paraphrase generation. News article 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.

Placenta sandwich? No, urban legend!
Tom wants to make movie with Katie
Kate's dad not happy with Tom Cruise
Cruise and Holmes sign for eighteen million Eighteen million for Tom and Katie
Newest mission Tom Cruise not very convincing Latest mission Tom Cruise succeeds less well Tom Cruise barely succeeds with MI:3
Tom Cruise: How weird is he? How weird is Tom Cruise really?
Tom Cruise leaves family Tom Cruise escapes changing diapers

Table 1. Part of a headline cluster with subclusters

METHOD

1. CLUSTERING



2. PAIRWISE SIMILARITY

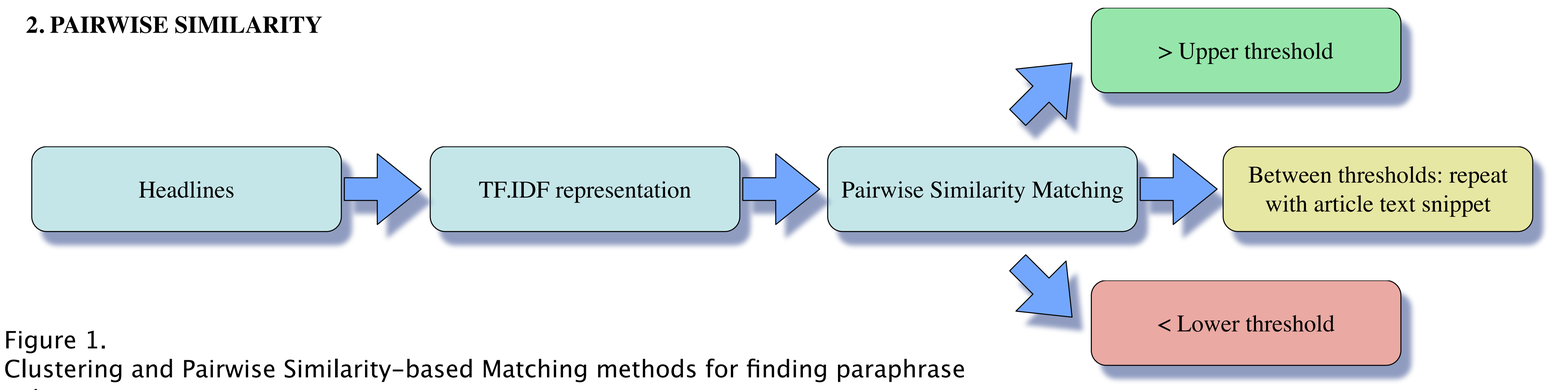


Figure 1.
Clustering and Pairwise Similarity-based Matching methods for finding paraphrase pairs

- **The data we use was acquired by crawling Google News Netherland 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_β-score with a β of 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.**

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 Pairwise similarity method, optimizing using an $F_{0.25}$-score, the optimum values for the lower and upper threshold are Th_{lower} = 0.2 and Th_{upper} = 0.5**
- **Using pairwise similarity on the 30,000 headline clusters results in roughly 200,000 Dutch paraphrase pairs**

Type	Precision	Recall
k-means clustering clusters only	0.91	0.43
k-means clustering all headlines	**0.66**	0.44
pairwise similarity clusters only	0.93	0.39
pairwise similarity all headlines	**0.76**	0.41

Table 2. Results for both *k*-means clustering and pairwise similarity-based matching on headlines that could be clustered only and on all headlines

Playstation 3 more expensive than competitor Playstation 3 will become more expensive than Xbox 360
Sony postpones Blu-Ray movies Sony postpones coming of blu-ray dvds
Prices Playstation 3 known: from 499 euros E3 2006: Playstation 3 from 499 euros
Sony PS3 with Blu-Ray for sale from November 11th PS3 available in Europe from November 17th

Table 3. Examples of correct (above) and incorrect (below) alignments, according to the human annotations

DISCUSSION & FUTURE WORK

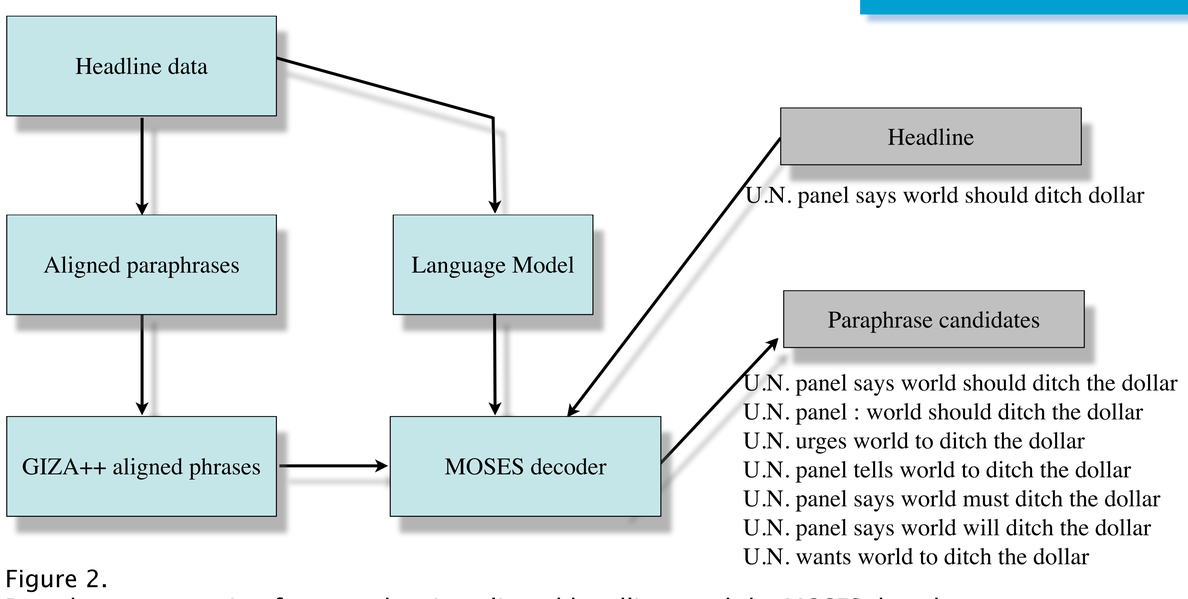


Figure 2.
Paraphrase generation framework using aligned headlines and the MOSES decoder

- **Pairwise similarity-based matching is a good way of collecting syntactically similar paraphrases**
- **Combining this method with crawled headline clusters results in quick and simple acquisition of large amounts of paraphrase data for text generation**
- **Next steps involve using this method for building a phrase based sentential paraphrase generation system for English**

References

Erwin Marsi and Emiel Krahmer, 2007, Annotating a parallel monolingual treebank with semantic similarity relations. In he Sixth International Workshop on Treebanks and Linguistic Theories (TLT'07).

Ted Pedersen and Anagha Kulkarni. 2006. Automatic cluster stopping with criterion functions and the gap statistic. In Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 276–279.

Aknowledgements

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