

Multi Document Centroid-based Text Summarization

Dragomir Radev
School of Information
and Department of EECS
University of Michigan
radev@umich.edu

Adam Winkel
Department of EECS
University of Michigan
winkela@umich.edu

Michael Topper
Department of EECS
University of Michigan
mtopper@umich.edu

1 Introduction

1.1 Text summarization

Text summarization is the process of taking a text document and creating a compressed version that consists of the most useful information for the user.

One distinguishes between single-document summarizers (SDS) and multi-document summarizers (MDS). Multi-document summarization is much more complicated than single-document summarization. Factors that make multi-document summarization more difficult include:

- Multiple articles can be written by different authors, having different writing styles and document structure.
- Multiple articles might have contradictory views of the same event. A useful summarizer should be able to detect and handle this situation.
- Facts and views can change over time, documents written at different times may have conflicting information.

1.2 The MEAD toolkit

MEAD is a publicly available toolkit for multi-document summarization (Radev et al., 2000; Radev et al., 2002). The toolkit implements multiple summarization algorithms (at arbitrary compression rates) such as position-based, TF*IDF, largest common subsequence, and keywords. The methods for evaluating the quality of the summaries are both intrinsic (such as percent agreement, precision/recall, and relative utility) and extrinsic (document rank).

1.3 Centroid-based summarization

MEAD uses a form of multi-document summarization called Centroid-based summarization. A cluster of documents with a common topic are used to produce a cluster centroid, consisting of words which are central to all of the documents in the cluster. The cluster centroid is then used to rank sentences based on their relevance to the topic of the cluster.

2 The MEAD summarization process

The MEAD summarizer consists of three components: A feature extractor, a sentence scorer and a sentence reranker. MEAD first computes a value for user-defined features of each sentence using the feature extractor. The features used in the Mead Demo are Position, Centroid, and Length. Once the features are computed, the sentence scorer gives a value to each sentence based on a linear combination of their features. Sentences are then ordered according to their scores. The sentence reranker then adds sentences to the summary beginning with the highest scoring sentence. The reranker calculates the similarity of the sentence about to be added with all of the sentences already in the summary. If the similarity is above a given threshold, the sentence is not added to the summary and the reranker moves on to the next sentence. Sentences are added to the summary until the amount of sentences in the summary corresponds to the compression rate.

3 NewsInEssence

NewsInEssence is a system for finding and summarizing multiple news articles on the web. The NewsI-

nEssence system allows the user three different options for selecting a cluster of documents to summarize.

The user may specify a query by typing in key words in the query box. NIE will search the pre-existing clusters and list the ones that contain news articles related to your query.

Another option for selecting a query is by specifying a seed story. The user can enter a URL of a news article from a list of supported news sites (currently BBC, Yahoo News, CNN, MSNBC and USA Today). NIE then uses this seed article to find other articles related to the same topic.

Users are also able to select pre-existing clusters to summarize. One can select a cluster under Google News Clusters, or also select a cluster created by another user by selecting one of the clusters under Recent User Clusters.

Once the cluster has been selected or created, the user can specify which articles to use as well as a compression rate for the summary. The summary is then produced and displayed.

NewsInEssence differs from other news trackers like Columbia's Newsblaster. Newsblaster gives short summaries from clusters of related Documents. NewsInEssence allows users to give stories on arbitrary topics as seeds and then it retrieves related stories to these. NIE also allows users to control which news sources to query, how long the summary should be, how similar to each other should articles in the same cluster be. Users can also see which source contributed to a given sentence in the summary and see the original document with that sentence highlighted. NewsInEssence can be found on the web at <http://www.newsinessence.com>

FindNews is a part of NewsInEssence that allows users to quickly and easily find news articles that they would like to use with NewsInEssence. FindNews parses the front page of popular news sites and adds an icon to each top news story. By clicking the NIE icon next to a news story, the user is forwarded to the NewsInEssence page with that story automatically added as the seed article.

4 Mead Demo

The Mead Demo is a web-based demonstration of MEAD. Users are able to add multiple documents

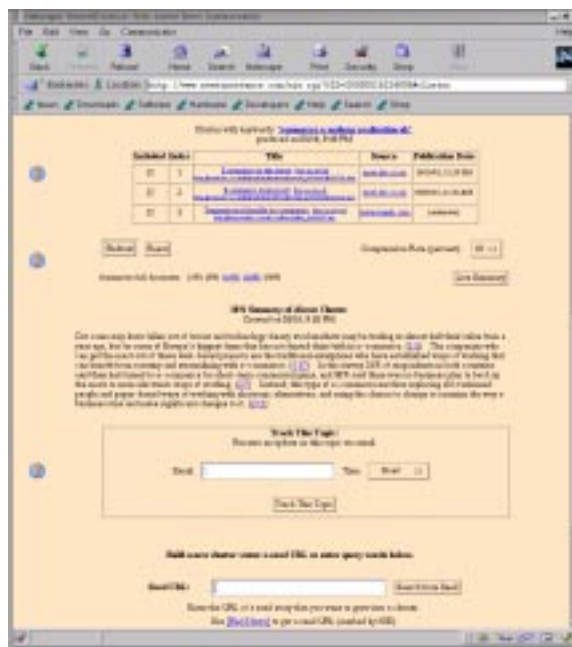


Figure 1: NewsInEssence cluster and summary

for the MEAD toolkit to summarize and display.

5 Credits

This work was partially supported by the National Science Foundation's Information Technology Research program (ITR) under grant IIS-0082884. We are grateful to Sasha Blair-Goldensohn for his work on an earlier version of this system.

References

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April.

Dragomir Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Celebi, Hong Qi, Dan Liu, and Elliott Drabek. 2002. Evaluation challenges in large-scale multi-document summarization: the mead project. submitted to SIGIR 2002, August.