A Scaleable Multi-document Centroid-based Summarizer

Dragomir Radev[1,2], Timothy Allison[3], Matthew Craig[2], Stanko Dimitrov[2], Omer Kareem[2], Michael Topper[2], Adam Winkel[2], and Jin Yi[2]
[1]School of Information
[2]Department of Electrical Engineering and Computer Science
[3]Department of Classical Studies
University of Michigan, Ann Arbor, MI 48109
{radev,tballiso,mwcraig,sdimitro,okareem,mtopper,winkela,jyi}@umich.edu

1 Introduction

We are presenting the most recent version of MEAD (v. 3.08), a large-scale public-domain summarizer that has been used in a number of applications, including the 2001 JHU summer workshop and the NewsInEssence project (`www.newsinessence.com`). A version of MEAD finished in first place on task 4 at DUC 2003 and finished in the top three on two other tasks.

In this demo, we will be showing several interfaces to MEAD, including a WAP-based cell phone interface, a Web-based interface, a command-line interface, and a Nutch-based interface.

1.1 Text summarization

Text summarization is the process of identifying salient concepts in text, conceptualizing the relationships that exist among them and generating concise representations of the input text that preserve the gist of its content.

One distinguishes between single-document summarization (SDS) and multi-document summarization (MDS). MDS, which our approach will be focusing on, is much more complicated than SDS in nature. Besides the obvious difference in input size, several other factors account for the complication, e.g.:

- Multiple articles might come from different sources, written by different authors, and therefore have different styles, although they are topically related. This means that a summarizer cannot make the same coherence assumption that it can for a single article.
- Multiple articles might come out of different time frames. Therefore an intelligent summarizer has to take care of the temporal information and try to maximize the overall temporal cohesiveness of the summary.
- Descriptions of the same event may differ in perspective, or even conflict with one another. The summarizer should provide a mechanism to deal with issues of this kind.

We also make the distinction between information-extraction- vs. sentence-extraction-based summarizers. The former, such as (Radev and McKeown, 1998), rely on an information extraction system to extract very specific aspects of some events and generate abstracts thereof. This approach can produce good summaries but is usually knowledge intensive and domain dependent. Sentence extraction techniques (Luhn, 1958; Radev et al., 2000), on the other hand, compute a score for each sentence based on certain features and output the most highly ranked sentences. This approach is conceptually straightforward and usually domain independent, but the summaries produced by it often need further revision to be more smooth and coherent.

1.2 Centroid-based summarization and MEAD

Centroid-based summarization is a method of multi-document summarization. It operates on a cluster of documents with a common subject (the cluster may be produced by a Topic Detection and Tracking, or TDT, system). A cluster centroid, a collection of the most important words from the whole cluster, is built. The centroid is then used to determine which sentences from individual documents are most representative of the entire cluster.

MEAD is a publicly available toolkit for multi-document summarization (Radev et al., 2000; MEAD, 2003). 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).

MEAD has an expansive architecture which allows end users to interface with its summarization capabilities through a Perl and Java API.

2 Demonstration of MEAD

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

MEAD toolkit to summarize and display (see Figures 1–4).

The Demo allows users to add documents by: selecting files from their computer, adding text in the text box or by supplying a URL to a specified web document. Documents can be plain text, HTML files or Microsoft Word files.

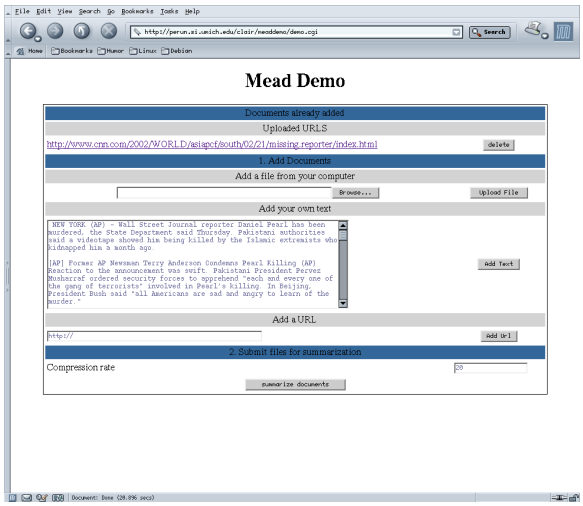

Figure 1: The user inputs a combination of files from their hard drive, text into the text box and document URLs.

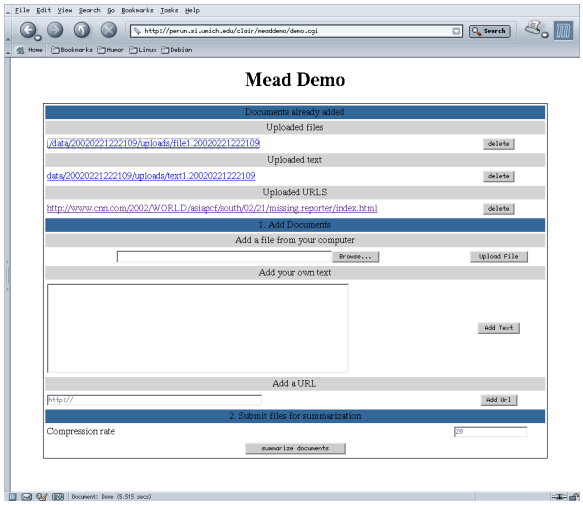

Figure 2: After user has added all of the documents to be summarized, they select the compression rate and then submit for summarization.

3 Demonstration of WapMead

WapMead (Figure 5) is a WAP (Wireless Access Protocol) interface to MEAD to access IMAP-based email

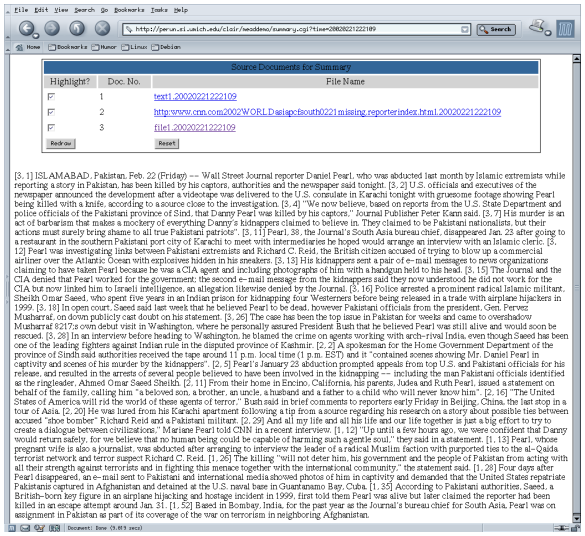

Figure 3: The summary page displays the summary as well as provides links to each document the user has submitted.

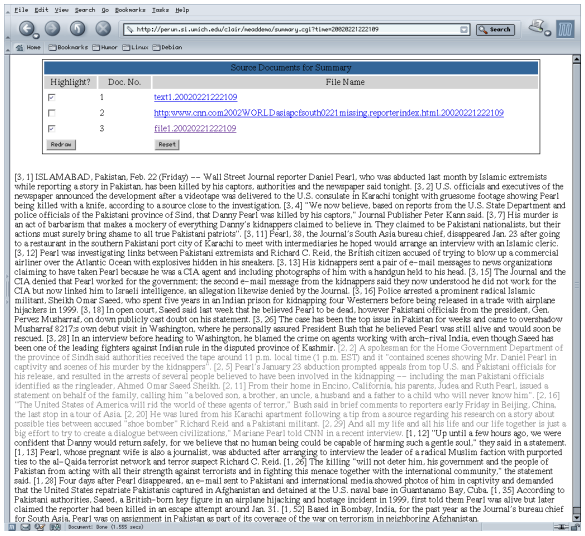

Figure 4: The summary page also allows users to select specific documents from the summarization to be highlighted.

mailboxes from a cell phone or other WAP-enabled device. WapMead has two modes: a mailbox view, in which a user can search for an email message and a summary view, in which a summary of a message is displayed. Summaries are displayed hierarchically, first showing the most salient sentences in the entire message and then (on a need basis) showing in greater detail particular areas of the message.

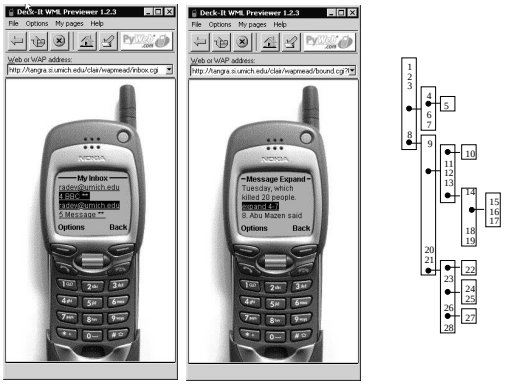

Figure 5: WapMead: (Left) mailbox view, (Middle) summary view, (Right) hierarchical view.

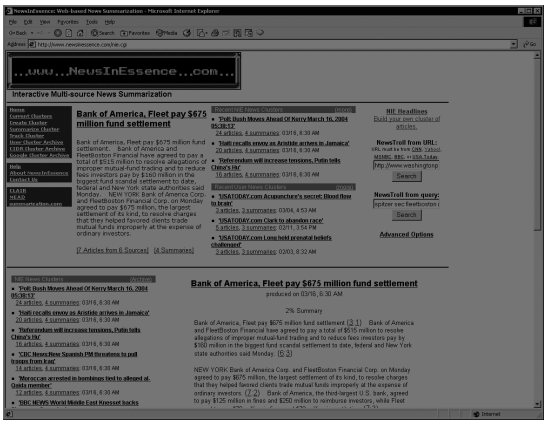

Figure 6: NewsInEssence (NIE) interface.

4 Other interfaces

We will be also showing the Java-MEAD interface (Figure 7) implemented in Nutch (a public-domain search engine from www.nutch.org), an older interface, NewsInEssence (Radev et al., 2001; Radev et al., 2002) (Figure 6), as well as the MEAD command-line interface.

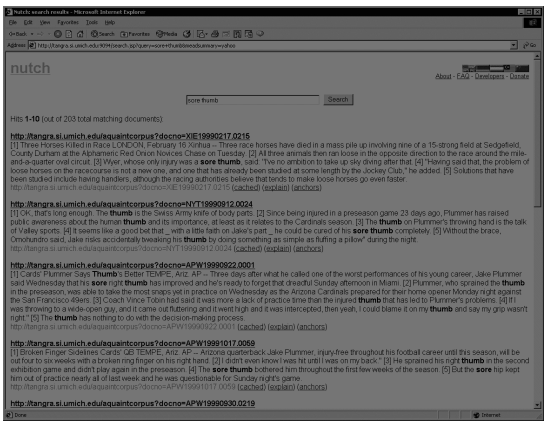

Figure 7: Nutch interface.

5 Acknowledgements

This work is partially supported by grant ITR 0082884 from the National Science Foundation (NSF). The authors would like to thank the MEAD team (Sasha Blair-Goldensohn, Simone Teufel, Horacio Saggion, Wai Lam, Arda Çelebi, John Blitzer, Hong Qi, Elliott Drabek, and Danyu Liu) for their hard work on various versions of the MEAD system.

References

H.P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165.

MEAD. 2003. Mead documentation. WWW site, URL: http://www.summarization.com/mead.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, September.

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 R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. 2001. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Human Language Technology Conference*, San Diego, CA.

Dragomir R. Radev, Michael Topper, and Adam Winkel. 2002. Multi Document Centroid-based Text Summarization. In *ACL Demo Session*, Philadelphia, PA.