The Michigan Single and Multi-document Summarizer for DUC 2002

Jahna C. Otterbacher[+] jahna@umich.edu
Adam J. Winkel[§] winkela@engin.umich.edu
Dragomir R. Radev[+§] radev@umich.edu

+ School of Information
§ Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109

Abstract

The MEAD summarization system, currently being developed at the University of Michigan, produces a summary of the user's desired length, based on one or more source documents. Recently, MEAD has been slightly adapted, and is now compatible for the summarization of this year's Document Understanding Conference (DUC 2002) articles. In addition, we have recently introduced an interactive, online news summarization system, NewsInEssence, which uses MEAD as the backend summarizer.

1 Introduction

For the single and multiple-document summarization tasks in DUC 2002, our approach is based on sentence extraction. In other words, we attempt to compress one or more documents by identifying their most important and informative sentences. In generating the summary, only such sentences are included, while the less important ones are omitted.

In the next two sections we describe the MEAD summarizer which we used for the DUC tasks, as well as the modifications that we made to the MEAD system in order for it to be DUC compliant. We then present the experiments and metrics we used in order to set the parameters of our system. Finally, we introduce our online news summarization system, NewsInEssence, which demonstrates the fruits of our research on multi-document summarization at the University of Michigan.

2 The MEAD Summarizer

The MEAD summarizer [Radev et al., 2000] [Radev et al., 2002], which was developed at the University of Michigan and at the Johns Hopkins University 2001 Summer Workshop on Automatic Summarization, produces summaries of one or more source articles (or a 'cluster' of topically related articles). In the initial versions of MEAD, a centroid-based approach was used for summarization via sentence extraction. For each cluster of related documents, a centroid was produced, which specified key words and their respective frequencies in the set of source articles. Given the input documents and a compression rate, the algorithm then chose sentences with a high number of the key centroid words, since it was thought that such sentences are central to the cluster's topic.

More recent versions of MEAD use a linear combination of three features to rank the sentences in the source documents. The first of the three features is the centroid score previously mentioned. The second is the position score, which assigns higher scores to sentences that are closer to the beginning of the document and lower ones to those further away from the beginning. Finally, the third feature, length, gives more weight to longer sentences. Using a linear combination of these three features, sentences are ranked by score and added to the summary until the user's desired length is attained. To avoid redundancy, MEAD also employs a 'sentence reranker' that ensures that the chosen sentences are not too similar to one another

as far as lexical items is concerned. A cosine similarity metric is computed for all pairs of chosen sentences. If this cosine is higher than the user's specified threshold, the later sentence is too similar to the earlier one, and is not included in the summary.

MEAD is publicly available at http://www.clsp.jhu.edu/ws2001/groups/asmd/. It is flexible in that it allows users to customize the sentence selection critera for various summarization tasks. Users may create almost any new feature to be used in sentence selection as well as different reranking algorithms. For example, a user may be interested in choosing sentences that are in an anaphoric relationship with another sentence. Customized features and reranking routines may be easily incorporated into the existing MEAD package.

3 Adapting MEAD for DUC 2002

The incarnation of MEAD system used for our DUC 2002 entry has been modified in a few important ways. As mentioned previously, in MEAD, the user typically specifies the length of the desired summary as a compression rate, or a percent of the length of the original documents, in sentences. In the adapted version of MEAD, a user can now specify the length of the output in words.

As is expected, the input document format required by MEAD differs from the DUC document format in a few subtle ways. Also, the extract files produced by MEAD needed to be massaged into the format required by the DUC evaluation system. To this end, we wrote several Perl scripts to perform these conversion operations. Another important way that MEAD's formats differ from DUC's is that MEAD's formats are all XML-based and DUC's are SGML-based. This presents parsing problems, because though SGML is a well-known standard, fewer tools are available to process SGML than are available for XML. Also, XML requires that all attributes be placed in single or double quotes, while SGML does not. Our final source of consternation was that some characters present in the DUC documents are illegal in XML, which doesn't allow any character greater than 0x7F.

We also adapted the MEADEVAL toolkit (see the next section) for compatibility with DUC-style documents. This required writing additional scripts to perform the conversion.

The previously described modifications to MEAD will be included in the next release, version 3.07, which will be made available towards the end of September 2002. We have also decided to integrate MEADEVAL into the MEAD distribution, so both MEAD and MEADEVAL will be compatible with DUC document formats.

4 Experiments and Training

In order to set the system parameters, we trained our adapted version of MEAD on the DUC 2001 data. Specifically, we used three clusters of last year's data in order to develop our summarization approach. The clusters in Table 1 were chosen because they represent different types of news stories that one finds in the press. For example, the first cluster about the junk bond trader Michael Milken can be categorized as biographical in nature. To contrast, the documents in the third cluster, which describe the Exxon Valdez oil spill and its cleanup, focus on a particular event and its subevents.

Cluster	Topic	# Documents
d02a	Michael Milken	11
d35f	Cancer and smoking	7
d52i	Exxon Valdez	9

Table 1: Training Clusters from DUC 2001

Three human judges were asked to read the 27 documents from the three chosen clusters. They were asked to assign a score to each sentence, indicating how important and relevant it was to the

general topic of the overall cluster. The scores ranged from 0, indicating that the given sentence was completely irrelevant to the cluster, to 10, which meant that it was quite central or crucial to the topic. We call these scores 'utility judgements.'

Once we collected the above data, we were able to generate summaries based on the three clusters of news stories and to evaluate them. Specifically, we used the four evaluation metrics below in assessing our summaries:

1) Interjudge agreement: Expresses to what extent judges' ratings agree.
2) Expected random utility: The average of the utility of all possible system outputs at a given summary length. (E.g. The expected value of the utility of a summary of a given length made up of randomly chosen sentences.)
3) Relative utility: The sum, over all judges and all sentences in the summary, of the ratio of the assigned utility score to the maximum score assigned by any judge.
4) Normalized relative utility: A relative utility score that restricts the system performance to be between 0 and 1. If the system performs no better than random, the normalized score is 0.

These metrics allowed us to set upper and lower bounds on the system's performance for a given cluster of news documents during the training process. This assessment was done using our evaluation package, MEADEVAL, which is available at http://perun.si.umich.edu/clair/meadeval.

5 MEAD Parameters

The main parameters that we needed to determine were related to the three features used by MEAD, the centroid, position and length scores. The default parameters are as follows:

Parameter	**Value**
Centroid	1
Position	1
Length	9
Cosine Similarity (Reranker)	0.7

Table 2: MEAD Default Parameters

Length is a 'cutoff feature' in MEAD, such that if the length parameter is set to 9, sentences less than nine words long will not be considered for inclusion in the summary. The values of centroid and position are used in the linear combination of the features to obtain scores for the sentences. For example, the default parameters have the following interpretation in the sentence ranking algorithm, so long as the length of the sentence is greater or equal to nine words:

Score (sentence) = {1*Centroid + 1*Position}

If the sentence is less than nine words long, it receives a score of zero. The cosine similarity parameter is the threshold used by MEAD to determine if a given sentence is too similar lexically to another sentence of a higher rank. In other words, the default value of 0.7 means that when each pair of candidate sentences are compared, if the cosine between the two text strings is greater than 0.7, the lower ranked sentence is excluded from the summary.

In our experiments with the DUC 2001 data, we tested many different combinations of parameters. We originally thought that we would find that various configurations performed differently in the DUC tasks. We also felt that we might develop several configurations of our system for use with the different types of news clusters, for instance biographical versus event or sub-event news articles. However, in the end we did not find other configurations that performed as consistently as

did the default MEAD system. In retrospect, this was not that surprising given that the default parameters of MEAD were determined after lots of experimentation during the system's development.

6 Results

The Michigan team did not do as well as we had hoped to in this year's DUC competition. Most of our energies this year were devoted to adapting the MEAD summarizer to accommodate the DUC 2002 data and the competition's tasks. As mentioned previously, we needed to modify the system to accept compression rates expressed in number of words rather than as a percentage of the number of sentences in the source documents. Additionally, we need to convert SGML to XML for use in our system. In future competitions we hope to commit more of our energies to the modification of our algorithms rather than the MEAD system itself.

Another thing to note is that we found a discrepancy in our sentence ranking scripts, such that the feature length may have been included as a standard feature as well as a cutoff feature during the DUC competition. This means that length would have been the dominant feature that determined the scores of the sentences, which we did not anticipate. We believe this may be partially responsible for our rather poor performance in the DUC tasks.

7 Conclusions

The MEAD summarizer has been adapted and is now compatible for use with DUC 2002 data, such that it first converts documents in SGML to XML format. Users may specify a compression rate in terms of a percentage of the number of sentences in the source articles, or in terms of the desired number of words. In addition to creating summaries of one or multiple related news documents, MEAD offers the user an evaluation package which attempts to quantify the utility of a given summary.

Recently, we have also implemented MEAD in an online news summarization system, NewsInEssence (NIE). It is now available publicly at http://www.newsinessence.com. NIE allows the user to choose either an existing set of articles to summarize or to specify an online article to be used as a seed. If the user does not know the URL for a seed article, the 'Findnews' feature in NIE can be used to find news articles of interest. Next, NIE searches the web for articles that are related to the seed document. Finally, the user may indicate the compression rate of the desired summary. This work represents one way in which automatic text summarizers such as MEAD can assist users in finding the information they want to read on the web.

References

[Goldstein et al., 1999]

Jade Goldstein, Mark Kantrowitz, Vibhu O. Mittal, and Jaime G. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Research and Development in Information Retrieval*, Berkeley, CA, 1999.

[Radev et al., 2000]

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

[Radev et al., 2001]

Dragomir Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Celebi, Hong Qi, Daniu Liu, and Elliott Drabek. Evaluation challenges in large-scale multi-document summarization: the MEAD project. Johns Hopkins University CLSP Workshop Final Report, 2001 .