Evaluation of Automatic Text Summarization Across Multiple Documents

Mary McKenna & Elizabeth D. Liddy
TextWise LLC
2-212 Center for Science and Technology
Syracuse, NY 13244
(mary@textwise.com; liz@textwise.com)

Abstract

This paper describes an ongoing research effort to produce multiple document summaries in response to information requests. Given the absence of tools to evaluate multiple document summaries, this research will test an evaluation method and metric to compare human assessments with machine output of newstext multiple document summaries. Using the DR-LINK information retrieval and analysis system, components of documents and metadata generated during document processing become candidates for use in multiple document summaries. This research is sponsored by the U.S. Government through the Tipster Phase III Text Summarization project.

TextWise is a participant in the Tipster Phase III Text Summarization project funded by the U.S. Government. Our research objective is to produce high quality multiple document summaries. An established set of metrics to evaluate the performance of our production of multiple document summaries is not available at present. Therefore, this research effort is also concerned with developing a procedure to evaluate the summaries we create. We hope that we will uncover useful metrics and evaluation variables that can be used by other research efforts in this area.

The lack of automatic summarization evaluation tools is directly connected to the need for a comprehensive description of the different types of summaries possible. Automatic text summarization can mean many different things. The summary may be addressing a need of an information seeker (query dependent summary) or it may be independent of any specified information need (generic summary). The summary may represent a single document (single document summary) or a group of documents (multiple document summary). The summary may be an extract of sentences or pieces of text from a document (extract summary) or it may not use any of the actual wording from the source documents (generated text summary). Finally, the summary may provide a general overview of document contents (indicative summary), or it may act as a substitute for the actual document (informative summary). This terminology will be used though out this report in an attempt to clarify and define the various possible outcomes of automatic text summarization.

The combinations from the variations suggested above produce the following set of possible automatic text summarization outcomes:

1. Query dependent, single document, extract, indicative summary.
2. Query dependent, multiple document, extract, indicative summary.
3. Query dependent, single document, generated text, indicative summary.
4. Query dependent, multiple document, generated text, indicative summary.
5. Query dependent, single document, extract, informative summary.
6. Query dependent, multiple document, extract, informative summary.
7. Query dependent, single document, generated text, informative summary.
8. Query dependent, multiple document, generated text, informative summary.
9. Generic, single document, extract, indicative summary.
10. Generic, multiple document, extract, indicative summary.
11. Generic, single document, generated text, indicative summary.
12. Generic, multiple document, generated text, indicative summary.
13. Generic, single document, extract, informative summary.
14. Generic, multiple document, extract, informative summary.
15. Generic, single document, generated text, informative summary.
16. Generic, multiple document, generated text, informative summary.

Of course, there are many variations on the list above such as using phrases, proper nouns, etc. from a document instead of extracting sentences. A summarization system may also use metadata created in the course of processing and indexing a document collection. Metadata accompanying the source document or collection will likely be candidates for use in the summarization representation.

Noting the variety of possible summaries, it is impossible to come up with a single evaluation methodology or metric that is going to cover all variations of automatic text summarization. An evaluation should measure the degree to which the desired outcome has been achieved. The evaluation method chosen must be able to measure the extent to which the system produces the desired outcome. Ideally, we need a set of evaluation measures that will accommodate the different outcomes of text summarization systems noted in the list above. We also need measures that can judge the quality of summarization within a single system, and measures to compare systems.

There are many variables to be considered in measuring the performance of a text summarization system. Some variables are more important than others - the type of text summarization a system is producing will dictate what variables are most important. The length of the summary will vary depending on the length of the document in single document summaries, or the number of documents used to create multiple document summaries. Should informative summaries be expected to be longer than indicative summaries? Is 20-25% of the original document size the 'best' length for a summary? Is the time the user needs to review a single document or a multiple document summary a good judge of performance? Is cohesion essential? Is the summary accurate, comprehensive, and/or useful? Is narrative prose necessary? Multiple document summaries bring up evaluation variables that are not necessary relevant or as important in single document summaries: Information Source and Information Repetition. The source of the information presented in a multiple document summary is extremely important to convey. Repetition of information, while of minor concern in a single document summary, becomes extremely problematic in multiple document summaries, particularly using query dependent data. Display of summary information is another important part of the process of multiple document summarization. Should the summary display be interactive? Will using graphics improve the information representation of multiple documents containing textual information? Finally, can humans agree on what constitutes a good summary of a document (Salton et al. 1997) or set of documents?

For our Tipster project on summarization, we are producing query dependent, multiple document, extract, indicative summaries (number two in the list above). We will begin by perfecting our query dependent, single document, extract, indicative summaries which will provide a solid base for our multiple document summaries. Our 'extracts' will contain sentences, pieces of text, and metadata created during document processing. We will be constructing summaries using the top 10, top 20, and top 30 documents retrieved in response to a query.

To compose multiple document summaries, we will be using the output of the DR-LINK system (Liddy et al. 1994). DR-LINK is a natural language information retrieval and analysis system which returns relevance ranked result sets in response to a search request. DR-LINK document processing and indexing outputs are used for the components of the summary. These outputs include noun phrases (information system, running shoes), proper nouns with their categories (Country: India; Company: Analog Devices), subject fields (Information Technology; Electricity/Electronics), text structure (Consequence; Prediction), and the most relevant section of a document in response to a query.

DR-LINK is a web-based information system. Multiple document summaries will be another information analysis feature we will offer. For our Tipster project, we have agreed to develop a summarization system for news text, creating multiple document summaries using the top 10, top 20, and top 30 documents returned in response to a query. (If time allows, we may add the ability to summarize a group of user-selected documents.) We hope that our system will be extensible to other domains, but we will not know from these experiments.

Our research will examine the suitability of the DR-LINK document components described above one at a time. For example, the first component we are testing is the subject field. Subject tags define what a document is about. Subject tags are metadata created when the document is processed and indexed. When we collect all subject tags from multiple documents and then sort them by frequency, we have a list of subject areas that define the subject domains of the multiple documents set. Our testing will determine to what extent do these subject frequency lists represent the document collection. How many subject fields are necessary/appropriate to represent a given set of documents?

For our evaluation procedure, three human analysts will review eight queries (selected from the TREC queries (Harmon 1996) and provided as training queries for the Tipster Text Summarization project). These queries will be run against the Tipster collection of news text articles, which include documents from The Wall Street Journal and Associated Press. Analysts will be provided with lists of all possible subject fields for the top 10 documents, the top 20 documents, and the top 30 documents for each query. The analysts will select the 'best' set of subject fields, and rank them in priority order. Analysts are provided with written guidelines for selecting and ordering subject fields. They have been asked not to confer with one another about this task; any questions should be directed to the project leader.

After collecting the data from the analysts and entering it into a spread sheet, we will use the Kappa statistic (Carletta 1996) to calculate the intercoder reliability measure. Do the analysts agree on the 'best' subject fields to represent a document set? Is there any pattern discernible as to which subject fields have been chosen, and how many subject fields have been selected? We will then compare the results of this analysis to the system output. We will then adjust our automatic summarization module if the analysis of results suggest a modification to the algorithm.

This procedure will be repeated with each of the outputs mentioned above: noun phrases, proper nouns with their categories, text structure, and the most relevant section of a document in response to a query. We will add each component to the automatic summary representation, and then adjust the algorithm as suggested by the comparison with the analysts' selections.

The final evaluation will be to have the analysts assess how well the automatic summaries represent query dependent, multiple document, indicative summaries. Analysts will assess the comprehensiveness, accuracy, coherence, and usefulness (keeping in mind time saved and ideal summary lengths) of the multiple documents summaries. We will also ask the analysts to address two specific issues: repetitive information in the summaries, and the system's ability to easily inform the user of the source(s) of the information. Also, all analysts will be interviewed immediately after their assessments to elicit the criteria they used to make their assessments. Finally, the three analysts' assessments of the multiple document summaries will also be reviewed for the intercoder agreement level - to what extent did the analysts agree on what a useful, accurate, comprehensive, etc. multiple document summary might be.

Our research started in mid-October 1997. We are scheduled to be finished with this project in mid-October 1998. We hope this project will achieve two goals: useful multiple document summaries from documents processed by DR-LINK, and an evaluation method that will reliably measure the performance of a system producing query dependent, multiple document, extract, indicative summaries.

References

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*. 22(2):249-254.

Harmon, D.K. 1996. Overview of the Fourth Text REtrieval Conference (TREC-4). In Proceedings of The Fourth Text REtrieval Conference (TREC-4), 1-23. Gaithersburg, MD: National Institute of Standards and Technology Special Publication.

Liddy, E.D., Paik, W., Yu, E.S. & McKenna, M. 1994. Document Retrieval Using Linguistic Knowledge. In Proceedings of the RIAO 94 Conference, 106-114. Paris, France: JOUVE.

Salton, G., Singhal, A., Mitra, M. & Buckley, C. 1997. Automatic Text Structuring and Summarization. *Information Processing & Management*. 3(2):193-207.