MULTIPLE & SINGLE DOCUMENT SUMMARIZATION USING DR-LINK

Mary McKenna
Dr. Elizabeth Liddy
TextWise LLC
2-212 Center for Science and Technology
Syracuse, NY 13244
liz@textwise.com, mary@textwise.com
Phone: (315) 443-1989

Research Problem

Our Tipster Phase III research objective for the Summarization task is to produce a single summary across multiple documents returned from a search on an information retrieval system. An established set of metrics to evaluate the performance of our system is not available in this field at present, so this research is also developing a procedure to evaluate the summaries we create. We hope to uncover useful metrics and evaluation variables that can be used by others working in this area.

Automatic text summarization can mean many different things. A summary may be produced from the results of an information retrieval system query, or it may be created independent of any specified information need. A summary may represent a single document or a group of documents. A summary may be an extract of sentences or sections of text from the source documents, or it may use only small fragments or even none of the actual wording from the source documents. A summary may provide a general overview of document contents (indicative), or it may act as a substitute for the actual document (informative). Any evaluation methodology must take these variations into account, and clearly specify the type of summary the system is generating.

We have chosen to develop indicative, query dependent summaries for both single and multiple documents. We are using metadata, phrases, and sometimes representative paragraphs in our multiple document summaries. We have further refined our final evaluation framework by defining two applications for our multiple document summaries: query refinement summaries providing a thumbnail sketch of documents returned in response to a query, and topic overviews, supplying a much more detailed multiple document summary.

TextWise/Tipster Research Plan

This Tipster research project began October 15th, 1997. The goal of the research project as originally planned was to produce multiple document summaries, using the documents returned from a query using the DR-LINK information retrieval system. As this research project was part of the Tipster Phase III Text Summarization project, and Tipster did not have a multiple document summary evaluation track, our research plan was amended in January of 1998. We agreed to also create single document summaries in response to a query, in order to participate in the formal Tipster evaluation (SUMMAC) in February of 1998.

As of September, 1998, we are very close to completing this project. Today a user can run ad hoc queries on the DR-LINK/Tipster Summarization Project website. Users can display single document summaries for any document in a results set. For multiple documents summaries, two options are available. A Thumbnail Sketch (brief summary) of the top 30 documents is automatically provided at the top of the search results. The user may also select an option to create Detailed Summaries, specifying the number of documents (top-ranked 1-30) to be used in the summary.

Test Collection

SUMMAC administrators selected training and test queries for Tipster Phase III participants to use as a practice data set. The queries are all from the TREC collection. 200 of the top ranked documents associated with each query were also supplied to all participants. These 200 documents contain both relevant and nonrelevant documents in response to the query. There were four data sources used in the training set: the Wall Street Journal, Associated Press, Federal Register (FR), and Department of

Energy (DOE) documents. At the Tipster meeting in October 1997 it was decided that the FR and DOE documents were probably inappropriate for the summarization task at hand. At TextWise, we constructed 20 databases using the Tipster-selected data for system development. 10 of these databases contain all the documents from the retrieved sets, both relevant and nonrelevant, with FR and DOE documents removed. The second set of databases contain only relevant documents to a query, again with FR and DOE documents removed. The reason for the division is to determine what level of noise the nonrelevant documents are contributing to the summaries.

We did not use the lengthy narrative descriptions that are part of the TREC queries in the practice data set, as these are not representative of the length of queries used on our online service. We used only the Description sections of the queries as prepared by TREC.

DR-LINK Modules Used in Summaries

To compose multiple and single document summaries, we used the output of the DR-LINK system [1]. DR-LINK is a natural language information retrieval and analysis system which returns relevance ranked result sets in response to a query. DR-LINK document processing and indexing outputs comprise the components of the summary. These outputs include: complex nominals, which are selected noun/adjective phrases (information system, running shoes); proper nouns with associated categories (Country: India; Company: Analog Devices); subject fields which are metadata subject codes describing documents (Information Technology; Electricity/Electronics); and the selection of the most relevant section of a document in response to a query.

Selecting the 'Best' SFCs to Summarize a Group of Documents

Subject Field Codes (SFCs) were the first component of DR-LINK document tagging that were tested for use in summarizing multiple documents. SFCs are assigned when documents are indexed using the DR-LINK document processor. These are subject codes that describe what a document is about. There are about 900 possible Subject Field Codes. An SFC must have a certain value threshold to be used as a descriptor to represent a given document.

Example Document with Subject Field Codes

New Agent Orange Lawsuit Filed by Vietnam Veteran --- A Wall Street Journal News Roundup. 06/23/88 WALL STREET JOURNAL

A Vietnam veteran and his family filed a class-action suit in state court in Harris County, Texas, against seven chemical companies that manufactured the herbicide Agent Orange.

The suit seeks more than $15 billion in damages for veterans who didn't discover they had been injured by the herbicide until after the massive Agent Orange litigation was settled May 7, 1984.

The plaintiff, Ronald Hartman, served in Vietnam from December 1967 to December 1968. Last March, according to the complaint, Mr. Hartman was diagnosed as having lymphoma, a form of cancer that plaintiffs in the Agent Orange case allege is caused by exposure to dioxin, a byproduct of the herbicide.

The suit charges that the defendant chemical companies, Diamond Shamrock Chemicals Co.; Dow Chemical Co.; Monsanto Co.; Uniroyal Inc.; Hercules Inc.; Thompson Hayward Chemical Co., a unit of Harrisons & Crosfield PLC; and T.H. Agriculture & Nutrition Co. entered into a "willful conspiracy in calloused and complete indifference to the safety of those people who bravely served our country in Vietnam."

Michael Gordon, a New York lawyer who represents Diamond Shamrock, which has since split into several entities, dismissed the importance of the suit and contended those veterans whose symptoms developed after the settlement can also make claims against the 1984 settlement fund, which has grown over the years to more than $200 million.

Top 5 Subject Fields:
Laws/Court Proceedings
Business Practices
United States
Chemical Substances
Legal Decisions/Judgments

For our analysis, we listed the SFCs for the top 10, 20 and 30 documents in response to a query, and sorted by both frequency and an alphabetical listing. The frequency list was used to present DR-LINK's candidates for the best SFCs to represent a set of documents. The alphabetical list was used to provide the analysts with a list to choose the best SFCs to summarize the document set. We could have given the analysts the entire SFC list to choose from, but that would have been problematic on two counts: 1) the list is too long to make the selection a reasonable task, and more importantly, 2) the system has already selected a certain number of SFCs to represent the document sets, and our intention is not to rewrite the SFC module, but to determine if this module is helpful in summarizing a set of documents.

The senior researcher first performed the SFC coding task on a set of 30 documents to anticipate problems and provide written directions to the analysts. Three analysts participated in the task of choosing the 'best' SFCs to represent a group of documents. The analysts were asked to select the best SFCs to summarize the top 10 documents, the top 20 documents, and the top 30 documents. Analysts were not provided with the query; document SFCs are generated independent of a query. Analysts were given 30 full text documents for each of 8 topics selected from the test collection: 125 (A), 158 (B), 163 (C), 183 (D), 127 (E), 162 (F), 198 (G), 200 (H). In queries A-F, the top 30 documents contained both relevant and non-relevant documents. In queries G and H, the top 30 documents were all judged to be relevant to the query. While the n is admittedly very small here, relevant documents only for G and H were used in order to determine if analysts would find it easier to agree on SFCs for a relevant set of documents as opposed to relevant and non-relevant document sets.

Analysts were provided with lists of SFC candidates for the top 10, 20 and 30 documents, generated from the Tipster DR-LINK development site. The lists contained all the displayed SFCs used in a given set [1-10], [1-20], [1-30] of documents. SFCs were presented to the analysts in alphabetical order. The length of these lists varied. For the eight topics used:

Range of possible SFCs candidates for document set containing 10 documents: 10-16
Range of possible SFCs candidates for document set containing 20 documents: 14-26
Range of possible SFCs candidates for document set containing 30 documents: 17-31

Analysts were asked to choose from zero to ten SFCs that best represented a given group of documents. The limit of ten was imposed because the list was to be part of an indicative, not informative, summary. The analysts were asked to rank order these SFCs, although the rank order exercise was meant to provide possibly useful additional information, rather than being central to the effort. The purpose of this task was to see if the humans 1) agreed on what SFCs defined a set of documents, and 2) were the human selections similar to the frequency-ranked DR-LINK selections?

Intercoder Reliability Testing

The data from this exercise were analyzed using SPSS. The first test was a pairwise comparison of the analysts - did they choose similar SFCs to represent a given set of documents? The Kappa statistic [2] was used for this test. It is important to note that these are clearly somewhat subjective judgments on the part of the coders. We wanted to uncover the consistency of these judgments. To do this, we had to take into account the probability that agreements would happen by chance. That is why the Kappa statistic was used to judge intercoder reliability.

All Kappa results in this report were statistically significant - $p < 0.01$

Average Kappa value for selecting the SFCs that best summarize the top 10 documents:
Pairwise comparisons between humans 1, 2, & 3, and then averaged:
n=125 .49 .63 .48 Avg .53

Average Kappa value for selecting the SFCs that best summarize the top 20 documents:
Pairwise comparisons between humans 1, 2, & 3, and then averaged:
n=165 .56 .59 .57 Avg .57

Average Kappa value for selecting the SFCs that best summarize the top 30 documents:
Pairwise comparisons between humans 1, 2, & 3, and then averaged:
n=195 .59 .61 .57 Avg .59

To interpret this result, we need to consult the guidelines first presented by Landis and Koch [3]

0.0-0.20 = slight agreement
0.21-0.40 = fair agreement
0.41-0.60 = moderate agreement
0.61-0.80 = substantial agreement
0.81-1.00 = almost perfect agreement

We found that our agreements among coders, selecting the best SFCs to summarize document sets, is right on the line between Moderate Agreement and Substantial Agreement. There are studies that suggest Kappa may be interpreted differently depending on the complexity of the coding task - that is, a lower Kappa may be signaling strong agreement if the task is very complex. However, as we are not able to judge the 'complexity' of this coding task, we will be using the Landis and Koch guidelines for our results interpretation. Another way of interpreting this statistic is that a Kappa of, for example, 0.50, shows that there is a 50% agreement between coders above what could have occurred by chance [4].

The levels of agreement did not significantly change for codes assigned to relevant and non-relevant document databases (A-F) versus relevant document only databases (G,H). However, this may be due to the small sample size.

Our results comparing human coders with the computer (DR-LINK generated output) required some data correction to account for the fact that in the initial coding scheme, the computer was forced to a code of one (i.e., the computer always coded for the presence of the SFC). In the corrected data scheme, the computer output was recoded so that all occurrences of SFCs with a frequency of one were changed to zero.

Average Kappa value for selecting the SFCs that best summarize the top 10 documents:
Pairwise comparisons between humans 1, 2, & 3, and the computer, then averaged:
n=125 .45 .33 .36 Avg .37

Average Kappa value for selecting the SFCs that best summarize the top 20 documents:
Pairwise comparisons between humans 1, 2, & 3, and the computer, then averaged:
n=165 .29 .50 .49 Avg .43

Average Kappa value for selecting the SFCs that best summarize the top 30 documents:
Pairwise comparisons between humans 1, 2, & 3, and the computer, then averaged:
n=195 .28 .40 .48 Avg .39

It is clear that the level of agreement between the human coders and the computer do not match the level of agreement between humans. However, it is encouraging to note the levels of agreement between humans and the computer are still highly significant, or to use Landis & Koch's language, a fair to moderate agreement. These results reflect how much agreement was reached among humans, and comparing human selections with the with the automatically produced SFC selections. These results do not demonstrate the extent to which the list of subject field codes, selected by either the humans or the computer module, actually represented the set of documents.

From the human perspective, as the task became more difficult as more codes were introduced, the level of agreement with the computer showed little change (a better statistical analysis of this would require even larger document sets).

Most Relevant Paragraph Selection

For the next module investigation, we examined the DR-LINK selection of the Most Relevant Section (MRS) of a document in response to a query. DR-LINK processing divides documents into logical sections. The section that is most similar to the query, as chosen by a selection algorithm, is presented as the Most Relevant Section of the document to a user. In order to adjust this algorithm to be used for multiple document summaries, we chose a single paragraph within the Most Relevant Section to serve as the summary text of the document. We have named this selection the Most Relevant Paragraph (MRP). We select the Most Relevant Paragraph from the Most Relevant Section by simply using a list of query terms and a stopword list to determine the most appropriate paragraph within the MRS. The MRS algorithm has already performed most of the work, the MRP algorithm just refines this a step further.

We use Most Relevant Paragraphs in both single and multiple document summaries. For the multiple document summaries, we did not want to include relevant paragraphs that were duplicates or near duplicates; we wanted to avoid using repetitious information as much as possible. We removed duplicate and near duplicate paragraphs using a very simple algorithm that computes similarities among substrings in the MRPs. We require only one substring overlap to be found in order to declare a match. Duplicate paragraphs are noted, although not displayed, in the summary. A link is provided to the full text of the document containing the duplicate paragraph should a user want to investigate that document (duplicate paragraphs may be from a different source or a later/earlier edition of a story, so it is important to retain the link to the duplicate paragraph document.)

The SUMMAC Evaluation

As noted above, our original project goal and focus was to develop multiple document summaries. However, the Tipster SUMMAC evaluation did not include the evaluation of multiple document summaries. Therefore, in order to participate in the SUMMAC evaluation, we briefly diverted all efforts to create single document summaries as required by the formal evaluation process.

We participated in two of the three tasks for the SUMMAC evaluation: Ad Hoc - produce indicative summaries to convey what the document is about; Question & Answer - produce informative summaries that would serve as substitute for the original document. We did not participate in the categorization task because the task was query independent and our current system is built around the use of a query.

We submitted results for 10% fixed length Ad Hoc summaries (summary could be no more than 10% of original document size), 'best' length Ad Hoc summaries, and Question & Answer summaries (limited to 30% of original document size). Our submission for Ad Hoc 'best' and Q&A were both limited to 30% of the document size. Our submission for Ad Hoc Best and Q&A were identical. We participated in the Q&A task just to get an idea how well our admittedly 'indicative' summaries fare when being judged as 'informative' summaries. (The evaluation results made it clear that our indicative summaries cannot serve as informative summaries!)

For the 10% (Brief) summaries we used the Most Relevant Paragraph, and if we needed to cut off in mid-sentence, an ellipses was used. If no MRP was chosen, then we defaulted to displaying the lead paragraph. If still under 10%, we then included the top 3 Subject Field Codes for that document. If still under 10%, we used the top 5 complex nominals with a frequency greater than 1. Finally, if still under 10%, we listed as many proper nouns with a frequency greater than 1 from the People, Places, and Company categories as possible.

For the 30% (Best) and Q&A summaries, we included lead paragraph in addition to Most Relevant Paragraph when these were not identical. We used the top 3 subject filed codes (sorted by frequency), all complex nominals with a frequency greater than 1 (again sorted by frequency), and all proper nouns from the People, Places, and Companies categories, provided length limitations did not prohibit their use. Finally, we formatted the summaries using descriptive labels to explain to the user what is contained in the summary. These labels were Lead Paragraph:, Most Relevant Paragraph: , Subject Areas Mentioned:, and Frequent Phrases Mentioned:.

Below is an example of summaries submitted for a query 257, used in both the Q&A and the Ad Hoc task. The query asked about cigarette consumption in the U.S. (We were not allowed to use titles in any summaries.)

Example - 10% Adhoc Summary #AP880521-0203

Since 1981, total U.S. cigarette consumption has dropped more than 10 percent, and last year Americans consumed 575 billion cigarettes, or ``pieces," the fewest since 1972.

Subject Areas:	*Common Phrases:*
Smoking/Tobacco	*last year*
Commerce/Trade	*tobacco support operations*
Government Powers	*tobacco lawmakers*
	tobacco program
	trade bill

Example - 30% Adhoc Summary #AP880521-0203

Lead Paragraph: The golden leaf of tobacco is offering little shade from the heat generated by U.S. Surgeon General C. Everett Koop's declaration that nicotine in cigarettes is addictive.

Most Relevant Section: Since 1981, total U.S. cigarette consumption has dropped more than 10 percent, and last year Americans consumed 575 billion cigarettes, or ``pieces," the fewest since 1972.

Subject Areas Mentioned :
Smoking/Tobacco
Commerce/Trade
Government Powers
Governmental Institutions

Frequent Phrases Mentioned:
last year
tobacco support operations
tobacco lawmakers
tobacco program
trade bill
net outlays

People:	*Places:*
Verner Grise R-R.I.	*District of Columbia*
John Chafee D-Ky.	*North Carolina*

Companies: Commodity Credit CORP

SUMMAC Evaluation Results

For the Ad Hoc task, participants were all very close to one another in performance. TextWise was the only participant in the best quadrant for the F-score by Time on the fixed length summaries. We believe this is due to the fact that we were the only participants that used lists and sentences in our summaries; all other participants used only sentence extracts. Lists can be viewed very quickly. Since the point of summaries is to save time, a well selected list can save a lot of time.

As mentioned above, we did not do well in the Question and Answer task by submitting identical summaries for both the Ad Hoc task and the Question and Answer task. The Question and Answer task clearly required more informative summaries, which our system was not designed to create. We would need to do development work in this area to actually produce informative summaries.

For future development, we will continue to develop mixed summaries, i.e., using both lists of proper nouns, phrases, and subject areas, as well as summary sentences. Improving our precision selecting lists and sentences to summarize documents will be the aim of any further development of our single document summaries.

Proper Nouns & Complex Nominals Used in Multiple Document Summaries

After completing the SUMMAC evaluation work, we returned to the creation of multiple document summaries. The next DR-LINK modules selected for use in the multiple document summaries were proper nouns (PNs) and Complex Nominals (CNs) (noun/noun or adjective/noun phrases). The DR-LINK document processing module tags proper nouns and assigns one of 50 descriptive categories for each proper noun in our indexes. Complex nominals are bracketed as well.

For the selection exercise, we ordered PNs and CNs using frequency of occurrence among a given document set. We used only PNs and CNs with a frequency greater than one; using a frequency of one to summarize a set of documents makes little sense. If a word or phrase is mentioned only once in a set of ten, twenty, or thirty documents, it is hardly exemplary of the document set.

We provided directions and materials for our three analysts, and asked them to choose the most appropriate PNs and CNs to represent a given set of documents, using our eight query test set. Analysts were not provided with any lists to choose from. The PN and CN selection was done after the analyst read each set of documents.

We then compared the analysts selections. There was very little intercoder consistency among analysts, except for the very top frequently occurring PNs and CNs. We chose to use only this high frequency group of PNs and CNs in our final summaries. Frequency Cutoff figures are noted below. The ranges below are not absolute; a different query may present a higher or lower number of PNs or CNs for any given frequency in a document set.

	Frequency Cutoff	Example of Ranges from Test Queries
PN range for 10 documents:	3	5-32
PN range for 20 documents:	4	10-38
PN range for 30 documents:	5	7-42
CN range for 10 documents:	3	3-24
CN range for 20 documents:	4	4-37
CN range for 30 documents:	4	7-47

An interesting outcome of this experiment was the varying numbers of PNs and CNs each analyst used to represent a set of documents, even though they were each presented with an identical set of directions and documents. To summarize a set of 20 documents, on average, Analyst One used 23 CNs and 25 PNs. Analyst Two used 18 CNs and 15 PNs. Analyst Three used 6 CNs and 5 PNs. The analysts briefly described the task they had in mind when they were composing their lists. The variance apparent among the three analysts' interpretations of the word 'summary' is not unlike the wide range of interpretations of what the word 'summary' may mean to any group of users.

As a direct result of this observation, we have decided to implement two types of multiple document summaries. A Thumbnail Sketch, consisting of the five most frequent PNs, CNs, and SFCs, allows for a quick check on the results. A second, more comprehensive summary, the Detailed Summary, uses SFCs with a frequency of 3 or greater, PNs and CNs with a frequency of 2 or greater, and the most relevant paragraphs that do not contain duplicate information.

Final Evaluation

The final evaluation for this project is a qualitative assessment of the usefulness of both types of multiple document summaries. The final evaluation is still underway as this paper goes to press. We are again using three analysts for the evaluation. They are being asked to assess the following for both the Thumbnail Sketches and the Detailed Summaries: Were the summaries useful given the application description? Did the summary make sense? Did the summary allow the user to accomplish the information gathering task in a more efficient manner? Was there too much repetitive information in the summary? Were the most important ideas or themes included, while trivial details excluded? We are also soliciting general feedback as to each evaluator's opinions of the summaries - what's missing, what other applications are appropriate, what application are unmet, etc.

The purpose of the final evaluation is to assess our current system in order to direct future efforts. What can be said about our automatic summaries at the completion of this project? What areas need more development effort? What new directions might be pursued should we continue work in this area?

Conclusion

This research has produced single document summaries, and two types of multiple document summaries, using the DR-LINK system. We have discovered little agreement in the research community regarding definitions of summaries or the evaluation of summaries, although the Tipster project has certainly brought both issues to the fore. With this research project, we hope to have made a useful contribution to the early body of research on the creation and evaluation of both single and multiple document summaries.

References

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

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

[3] Landis, J.R. & Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*. 33, 159-174.

[4] Krippendorff, K. (1980*) Content Analysis: An Introduction to its methodology*. Sage: Newbury Park