Summarization of Discussion Groups

Robert Farrell, Peter G. Fait-weather, and Kathleen Snyder
IBM T. J. Watson Research Center
PO Box 218 Yorktown Heights, NY 105
{robfarr, peterf, ksynder}@us.ibm.com

ABSTRACT
In this paper, we describe an algorithm to generate textual summaries of discussion groups. Our system combines sentences extracted from individual postings into variable-length summaries by utilizing the hierarchical discourse context provided by discussion threads. We have incorporated this algorithm into a Web-based application called IDS (Interactive Discussion Summarizer).

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *abstracting methods, linguistic processing.*

General Terms
Algorithms, Design, Experimentation, Human Factors, Theory

Keywords
Summarization, extract, discussion, discourse, text, hierarchical.

1. INTRODUCTION
The explosion of available textual information on the Internet has fueled the demand for automatic methods of text summarization. Existing approaches have primarily focused on summarizing documents such as news articles or technical papers. In this paper, we examine how to generate summaries of discussion groups.

Discussion groups are used widely to enhance remote asynchronous communication and collaboration. Perhaps the best known discussion groups are Usenet newsgroups [12], but many Groupware products (Lotus Notes), help desk software, and distance learning systems (Lotus LearningSpace, Blackboard) contain some form of asynchronous threaded messaging.

The literature suggests that summaries are a powerful tool for enhancing learning while reading [1][10]. Summaries organize material to be learned into categories that serve as schema to be tilled in as reading progresses [3]. Summaries of discussion group postings can help participants get an overview of the content in a discussion, "catch up" on what has happened since they last participated, evaluate others' contributions, identify expertise, or capture the different perspectives on a topic.

CIKM'01, November 5-10, 2001, Atlanta, Georgia, USA.

2. PROBLEM
We are investigating how to adapt sentence extraction techniques [4][6] to the problem of generating summaries of discussion groups to support multiple content characterization tasks. Discussion groups present two important challenges for an extraction-based summarization system. First, postings are too short and numerous to only offer document summaries of each one: summaries must span multiple documents. Second, multiple authors generate postings, significantly reducing coherence across extracted sentences.

Fortunately, discussion groups also offer features that can be exploited to improve text summaries. Postings often uniquely identify the author (e.g., using a directory or e-mail address). Second, postings are often explicitly linked to form discussion "threads" (see Figure 1).

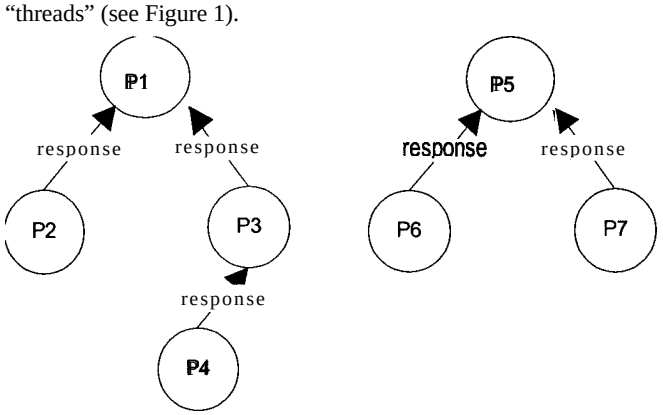

Figure 1: Discussion Threads

In this diagram, there are two discussion threads. P1 and P5 are *main topics.* The postings within a thread (the *responses)* are linked to the main topic through a set of transitive *response* relationships.

This paper explains how we can exploit the explicit discourse structure provided by discussion threads to generate improved text summaries. First we motivate our analysis, and then we describe a pilot experiment comparing human and machine summaries in this domain. Finally we describe our algorithm and implemented system.

3. MOTIVATION
The utility of discourse information for constructing summaries is well established. Studies of human summarization suggest that people construct a hierarchical discourse organization that organizes retrieval cues into memory and infer missing

information through reconstruction [2] [13]. Studies of professional abstractors indicate that they take a top-down strategy, exploiting discourse structure [5].

Several researchers have attempted to create summaries by parsing the text to find discourse relations and then selecting sentences for extraction based upon the inferred rhetorical structure [7] [8] [9]. However, relatively little work as been done to leverage discourse relationships when summarizing text generated by different authors, such as those found in e-mail exchanges, chat rooms, and discussion groups.

4. PILOT EXPERIMENT

To investigate discussion summarization, we had three IBM consultants read and summarize postings from a discussion database used by their colleagues. We asked them to think aloud while producing a six sentence free-form summary of each of two discussion threads. We then compared their summaries to those generated by *Textract* [3], a robust sentence-based summarizer that uses shallow linguistic processing and corpus statistics. Textract is the basis for the IBM Intelligent Miner for Text product. To create the Textract document summaries of each thread, we combined the text from the postings in the thread, including the main topic posting, in order of creation date, and generated a fixed-size summary (six sentences).

We found that while human summarizers extracted sentences and phrases from the discussion group postings, their choices were significantly different than the Textract algorithm, even after considerable tuning. Human summarizers found the sentence or sentences describing the key issue (e.g., problem, question) and then provided summaries of the responses relative to that issue. Temporal descriptions were also often relative (e.g., "The next day,") and repeated information was skipped. Most importantly, our human summarizers seemed to be using the structural discourse relationships between postings to guide their choice of summary sentences.

5. ALGORITHM

We have developed an algorithm *we* call ***hierarchical discussion summarization*** that performs sentence extraction and summarization recursively at multiple levels of a discussion hierarchy (see Figure 2).

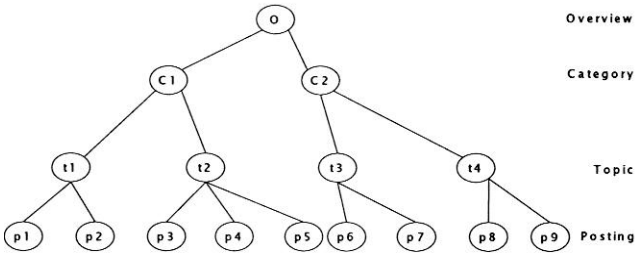

Figure 2: A Discussion Hierarchy

Our algorithm first selects the M most salient sentences from each posting. The salience of a sentence is computed from the salience of vocabulary items (single-token words, multi-word names, abbreviations, and multi-word terms excluding stop words) and the sentence's position in the document structure (proximity to the beginning or end of the posting). The item salience is given by the following inverse document frequency measure:

$$Salience(term) = \log_2 N_c/freq(term)_C AN_d/freq(term)_d$$

The next step operates over each of topics, combining the M posting-level summary sentences for each of P postings into a new topic-level synthesized document with P paragraphs of at most M sentences each, in date order. This process will generate at must M*P summary sentences for each thread (some postings may contain less than M sentences). Again, salience is computed for each term and proximity is used to prefer the summary sentences of initial and final postings in a thread. This results in at most M sentences for each topic.

After sentence extraction, a *context rule* is applied to each of the M sentences to check that at least Q sentences are included from its main topic. If none are included, the most salient main topic sentence is added, resulting in at most Q??M sentences. In practice, we set Q =I, but it we are experimenting with ways of determining Q automatically.

The same algorithm can be applied recursively up the discussion hierarchy, though position information is ignored at the Overview and Category levels. The result is a summary of the conversation that captures the salient exchanges while preserving context.

6. SYSTEM

We have incorporated hierarchical discussion summarization into a Web-based application called IDS (Interactive Discussion Summarizer). IDS provides the ability to query discussion databases by selecting document authors, topics, or categories. Optionally, the user may select a date range and desired summary document length (# of sentences). The resulting hierarchical summary is rendered as interactive web document using a style sheet (see Figure 3).

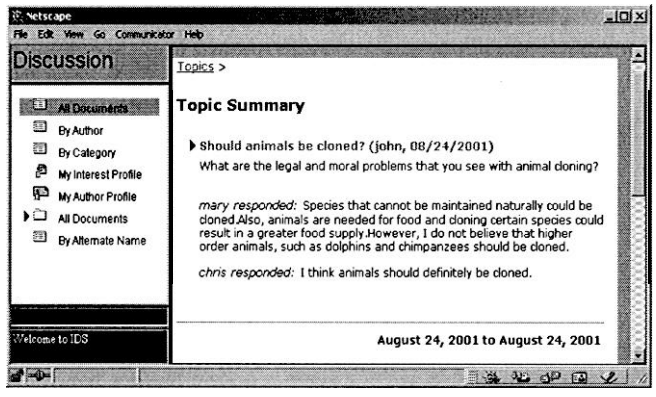

Figure 3: Topic discourse summary

The initial view lists the authors, topics, or categories that matched the user's query with associated descriptive plus a concise discourse summary of each. In Figure 3, the user selected a single topic ("Should animals be cloned") and the resulting summary consists of a single sentence from the main topic posting ("What are the legal and moral problems...?"), a three sentence summary of a posting from Mary ("Species that cannot...") and a

single sentence summary from Chris ("I think animals . .."). There is no attempt to give equal treatment to all contributions, but any topic may be expanded to uncover a listing of all the responses along with a short, equivalent-length summary of each. Summaries of postings can be further expanded into full documents with summary sentences highlighted.

7. CONCLUSION

We have developed a novel hierarchical discussion summarization algorithm and have applied the algorithm to several discussion databases. We are now evaluating its performance against a larger set of human summarizers. We anticipate that it will generate summaries for discussion groups that are closer to human summaries than previous statistical sentence extraction methods. We are also currently testing IDS as a tool for undergraduate teachers to use in reviewing student contributions to discussions in distance learning classes.

8. ACKNOWLEDGMENTS

Our thanks to Mary Neff and Roy Byrd, who allowed us to build on the Textract system, Nagaraj Ramarao for implementation, and Norma Baker, Dick Lam, and Richard Vigilante for access to discussion databases.

9. REFERENCES

[1] Anderson, R. C. & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *Psychology of Learning and Motivation.New* York Academic Press, 9, pp. 89-132.

[2] Bartlett, F. C. (1932). Remembering. Cambridge: Cambridge University Press.

[3] Bogurev, B. and Neff, M. (2000) Discourse segmentation in aid of document summarization. HICSS-33. Hawaii, IEEE.

[4] Edmundson, H.P. (1969) New methods in automatic abstracting. *Journal of the ACM,* 16(2):264-285.

[5] Endres-Niggermeyer, B. *(1998). Summarizing Information.* Springer-Verlag, New York.

[6] Luhn, H.P. (1959) The automatic cdretion of literature abstracts. IBM J. Res. Develop. 2(2), 159-165.

[7] Mann, W., and Thompson, S. (1988) Rhetorical structure theory: *Text* 8(3):243-281

[8] Marcu, D. (1999) Discourse trees are good indicators of importance in text. In I. Mani and M.T. Maybury eds., *Advances in Automatic Text Summarization.* Cambridge, MA: MIT Press.

[9] Miike, S., Itho, E., Ono, K., and Sumita, K. (1994) A full text retrieval system with a dynamic abstract generation function. In SIGIR'94, 152-161.

[10] Rothkopf, E.Z. (1996). Control of mathemagenic activities. In D.H. Jonassen (Ed.), *Handbook of research for educational communications and technology. New* York: Simon & Schuster Macmillan, 879-896.

[1 1] Rumelhart, , D.E. (1975) Notes on a schema for stories. In D.G. Bobrow and A. M. Collins (Eds) *Representation and Understanding,* New York: Academic Press.

[12] Usenet [see http://www.faqs.org/faqs/top.html]

[13] vanDijk, T.A. (1979) Recalling and summarizing complex discourse. In W. Burchart and K. Hulker, eds., *Text Processing,* Water de Gruyter, Berlin.