Automatic Text Summarization in TIPSTER

Thérèse Firmin
Department of Defense
9800 Savage Road
Fort George G. Meade, Maryland 20755-6000
tfirmin@romulus.ncsc.mil

Inderjeet Mani
The MITRE Corporation, W640
11493 Sunset Hills Road
Reston, Virginia 22090
imani@mitre.org

Automatic Text Summarization was added as a major research thrust of the TIPSTER program during TIPSTER Phase III, 1996-1998. It is a natural extension of the previously supported research efforts in Information Extraction (IE) and Information Retrieval (IR). There is considerable interest in automatically producing summaries due, in large part, to the growth of the Internet and the World Wide Web. The TIPSTER program sponsored seven research efforts into text summarization, all with different approaches to the problem.

- **Carnegie Group Inc** and **Carnegie Mellon University** teamed up to investigate a technique called "Maximal Marginal Relevance" or MMR, which produces summaries of documents by identifying key relevant, non-redundant information found within the document, intended primarily for very long documents.
- **Cornell University** and **SabIR Research, Inc.** used the document ranking and passage retrieval capabilities of the SMART IR engine to effectively identify relevant related passages in a document.
- **GE Research & Development** identified the discourse macro structure for each document and selected the passages from each component that scored well using both content and contextual clues.
- **New Mexico State University** also used information about the document structure combined with part of speech and proper name recognition to weight and select sentences to be included in their summaries.
- **Textwise LLC** primarily concentrated on summaries of multiple documents. They assigned subject field coding of documents as an initial indicator of document content and identified the most relevant paragraphs, combining statistical information about term frequency with linguistic information.
- **University of Pennsylvania** used co-reference resolution as the basis for their summaries, finding information within a document that is naturally linked together by referring to the same individual, organization, or event and extracting that related information to generate a summary.
- **USC Information Sciences Institute** (ISI) used a multi-faceted approach, including the optimal position of a sentence within a text, which varies based on text type, and building thematic representations of texts based on external ontologies.

Each of the systems is described in detail elsewhere in the proceedings.

In coordination with the various research efforts, DARPA sponsored an evaluation of text summarization systems. This evaluation was conducted in two phases. In September 1997, a dry-run was held to validate the evaluation methodology and establish a baseline for performance. Participation in the dry-run was limited to TIPSTER researchers only. A formal evaluation was conducted in May 1998 and was open to all interested parties.

The summaries can be characterized and evaluated by many features, including:

- *Coverage* – A summary can cover a single document or a group of related documents.
- *Focus* – A generic summary captures the main theme(s) of a document, whereas a user-directed summary is geared towards a particular topic of interest indicated by the user.
- *Intent* – An indicative summary provides a quick overview of the content of the full text, but is not intended to serve as a substitute. An informative summary should capture enough relevant information to be a replacement for the full text document.

The evaluation consisted of four tasks designed to assess performance of automatic summaries used in real world tasks and to leverage off of previous evaluations in IR and IE, the Text REtrieval Conferences and Message Understanding Conferences, respectively. It primarily addressed indicative summaries of single documents.

In the *adhoc* task, analysts read the output of the summarization systems intermixed with full text and baseline (lead sentence) summaries and assessed the relevance of each text presented to them with respect to a given topic (each system summary was a user-directed summary).

In the *categorization* task, analysts reviewed a different set of summaries, full text and baselines and had to determine which topic, among a set of five related topics, best represented the theme of the text (each summary was a generic summary).

For the *question-and-answer* task, each user-directed summary was evaluated to determine if it included the answers to 4-5 questions considered essential for a document to be relevant to a given topic. The percentage of questions answered was compared to that of a full text document. A summary that successfully captures all of the relevant concepts in a document could be considered a good informative summary of that document.

The *acceptability* task simply asked the analysts to read each of 30 summaries and the corresponding full text and determine if the summary was a good summary for the document. This task was not conducted during the dry-run.

For the adhoc and categorization tasks, participants submitted two summaries. One was restricted to 10% of the length of the document, the other could vary in length and was intended to capture the best summary a system could produce. The question-and-answer and acceptability tasks used only these best summaries.

TREC documents and relevance data were used for all tasks. This provided a good basis for the initial evaluations, however future evaluations should evolve beyond single document summarization of newspaper-style text.

The formal evaluation was held in the spring of 1998 and in addition to the seven TIPSTER contractors included participants from around the world:

- British Telecomm's ProSum
- Center for Information Research (Russia)
- Intelligent Algorithm's Infogist
- IBM Thomas J. Watson Research
- Lexis-Nexis
- National Taiwan University
- SRA International
- University of Massachusetts Center for Intelligent Information Retrieval
- University of Surrey

There was not a large difference in performance between the various systems in either evaluation, however the results do show some encouraging trends. The analysts were able to process the best summaries more quickly than the full text without a significant loss in accuracy, and they preferred reading documents that were shorter in length than the typical long full text articles. The results from the dry-run were discussed during the TIPSTER meeting in October 1997 and are documented in [Firmin and Chrzanowski 1998]. The results of the formal evaluation were discussed in May 1998 and are documented in [Mani et al., 1998].

Interest in summarization continues to grow. Two frequently mentioned applications include summarization of the results of an IR query (e.g. from Altavista or Infoseek) or combining summarization with a traditional text processing application such as Microsoft Word. Research efforts are moving towards summarization across documents and summarization in languages other than English. As this research continues and more applications come on the market, it will be useful to have a benchmark against which to evaluate the utility of the systems. The DARPA evaluation was a first step in that direction.

The dry-run and formal evaluations were conducted by the Department of Defense, SPAWAR Systems Center and The MITRE Corp. under DARPA sponsorship through the TIPSTER program.

References

[Firmin and Chrzanowski 1998] Firmin, T. and Chrzanowski, M. J., An Evaluation of Automatic Text Summarization Systems, in Mani, I. and and Maybury, M. *Advances in Automatic Text Summarization*, MIT Press, 1998.

[Mani et al., 1998] Mani, I., Firmin, T., House, D., Chrzanowski, M., Klein, G., Sundheim, B., Hirschman, L., Obrst, L. (1998), The TIPSTER Text Summarization Evaluation: Final Report, http://www.tipster.org/.