Reflections of Accomplishments in Natural Language Based Detection and Summarization

Susan R. Viscuso
Office of Advanced Analytic Tools
Washington, D.C., 20505
E-mail: susanrv@ucia.gov
Telephone: 703-613-8749

INTRODUCTION[1]

In Phase III, the GE team focused on accurate context indexing of text documents, generation of effective search queries, extended statistical retrieval with constraints, and document abstracting and summarization

The common tie among these lines of research is that natural language processing techniques offer a way of overcoming the weaknesses inherent to purely statistical approaches. GE pioneered the large-scale use of natural language processing techniques in information retrieval. Standard statistical search methods use words, word fragments, and simple collocations to index documents. The GE work is unique in that it indexes texts using meaningful linguistic units such as phrases, relations, entities, and events.

Another area of strength was the creation and validation of new tools for automatic and user-assisted development of effective search queries as well as the rapid review of search results through automatically produced short informative summaries. The GE work took advantage of the growing supply of natural language processing tools (taggers, parsers, extractors)---many of which were developed under the Tipster Text Program.

TEXT RETRIEVAL

The GE natural language Information Retrieval System indexes and retrieves free text documents using advanced natural language processing techniques coupled with a state-of-the-art vector-space model (Cornell's SMART search engine). The retrieval system consists of multiple text-processing steps which are linked together into a Stream Model architecture---in other words, information from multiple sources flows into the system and is the combined to get the most accurate results.

The processing and indexing steps include a part-of-speech tagger, a phrase extractor, a proper name extractor, and a fast syntactic parser that extracts concept like head-modifier relations from the text. The system also includes an automatic query expansion system which forms queries from the user's initial request. Finally, another unique aspect of the retrieval work is that GE researchers incorporated summaries of documents retrieved in initial searches into subsequent queries against the same document collection. Including summaries improved retrieval accuracy.

SINGLE DOCUMENT SUMMARIZATION

The GE team developed a robust flexible text summarization system that produced highly readable abstracts of various types of documents. The user can choose from two types of summaries. Short indicative summaries give the main points of a document, while longer informative summaries include more information and could substitute for the original document. The summarizer also can produce a summary tailored to a user's query (topical summary). The single document summarizer performed well

[1] This material has been reviewed by the CIA. That review neither constitutes CIA authentication of information nor implies CIA endorsement of the author's views.

in the Tipster Text Program sponsored Summarization Evaluation effort and has been well received by a number of Government users.

Both the summarizer and the query expansion tool have been transferred within GE to a number of commercial applications. For example, GE Capital Services uses the system as part of the experimental "Five-Minute Briefing" internet sales and marketing support toolkit. The system tracks the latest news and derives summaries from documents of interest to the user.

CROSS DOCUMENT SUMMARIZATION

At the end of Phase III, the GE team began working on cross document summarization. Cross-document summarization means a summary produced across topically related documents. In other words, one summary is produced from a number of documents which incorporates all of the themes in the document collection. This work is continuing outside the umbrella of the Tipster Test Program through Government sponsorship.

The GE approach to cross-document summarization is that any system developed must be flexible and allow user interaction. This is because there is no "best" cross-document summary because the right summary depends on the task, the user, and the type and number of documents.

The current system provides one summary for all the documents (currently 20 to 30) in a group. The sources of information for the facts included in the summary can be traced. Information is not duplicated and similar issues are presented together in the cross-document summary.

The basic algorithm is as follows:
summarize each document, cluster like summaries together and then choose a representative summary for each group, order the representative summaries to form the final summary.

Currently, the cross-document summarizer produces topical indicative summaries from news texts from any number of documents. The similarity of document clusters can be changed interactively by the user. The user can also examine similar documents in a grouping and also view the source documents from which various parts of the summary were constructed.

Currently, the Government is conducting an informal user-evaluation of the initial system. Some of the changes we expect to recommend will be improved clustering using cross-document coreferences, performing the initial clustering using full-text documents, clustering using geographic and temporal information as well as improved source selection and organization of the final summary. We also plan to extend the system to agency-specific documents as the next step in preparing the system for operational deployment.

OPERATIONAL USES OF SUMMARIZATION

Operationally, text summarization is an aid for people who deal with large amounts of text and want a tool that will allow them to determine what information exists in a text or a text collection and whether they have to read the entire document. Also, succinctly representing similarities and differences across a group of documents is useful because sometimes the analyst wants to focus on facts mentioned in only one or a few sources. For example, it may be that one article may contain a key fact that is not contained in the other documents in the collection. Interviews of analysts have shown that often the presence of a fact in only one or a few sources may be significant.

REFERENCES

Strzalkowski, T., Stein, G., and Wise, G.B., 1999. "Enhancing Detection through Linguistic Indexing and Topic Expansion", Proceedings Tipster Text Program (Phase III), this volume.

Strzalkowski, T., Stein, G., and Wise, G.B., 1999. "A Text-Extraction Based Summarizer, Proceedings Tipster Text Program (Phase III)", this volume.