The Cornell TIPSTER Phase III Project

F. Ruth Gee
Office of Advanced Analytic Tools
Washington, D. C. 20505
E-mail: ruthfg@ucia.gov
Phone Number: (703) 613-8759

INTRODUCTION[1]

The overall objective of the Cornell University TIPSTER Project was to improve end-user efficiency in information retrieval systems by reducing the amount of text that the user must process [1]. The project focuses on high precision IR, near-duplicate detection and context-dependent summarization. The two main foundations of the research are the latest version of the Smart system for information Retrieval and the Empire system for natural language processing. Smart is an implementation of the vector-space model of information retrieval (IR). Its earlier purpose was to provide a framework to conduct IR research but current developments will make the system easier to use by non-researcher. Empire is a research-oriented system that uses machine learning methods to quickly perform partial parsing of sentences.

Cornell's integrated approach uses both statistical and linguistic sources to first identify relationships among important terms in the query or in the text. The integrated system then uses the extracted relationships to (1) discard or reorder retrieved texts (for high-precision IR); (2) locate redundant information (for near-duplicate document detection); or (3) generate summaries. A more detailed technical description about the research can be found in the Cornell University technical paper [2].

HIGH PRECISION INFORMATION RETRIEVAL

The goal of this effort is to give the users the capability of conducting a high-precision search. If the user selects the high-precision option, the system will attempt to retrieve fewer documents than it would in a normal search but within the returned hits list, most of the documents should be useful. Emphasis on high precision, however, extracts a penalty in terms of recall. That is, some of the relevant documents or passages that are available in the stored text collection might not be returned to the user since the system will retrieve fewer documents overall.

The high-precision option is optimized for users who have a specific information need and a limited amount of time. This option would provide the user with a few piece of highly relevant data quickly but would not necessarily provide all the data. Alternatively, the user may opt to de-emphasize high precision in favor of improved recall but might suffer the consequences that arise from having to process an increased number of irrelevant documents.

NEAR-DUPLICATE DETECTION

The goal of the research in this area is to devise a system that will reduce the amount of duplicated information that the user sees. Current retrieval system may return several versions of a document in which the differences results from changes made to the metafile, word deviations in the body (e.g. multiple authors using different words to describe the same event) or an update with new information added. While other retrieval-enhancement algorithms have been developed to identify exact and near duplicates, this research effort investigates methods for processing documents that are similar but do not necessarily contain the same terms. For example:

- If a user has seen document X and the first two paragraphs of unseen document Y only contain information that is the same as or similar to information in X, the system would process Y to "hide" the duplicated

[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.

data showing the user only the unique paragraphs. An operational version of this system could be integrated with an agency's retrieval system to remove exact and almost duplicates from the retrieval system's hits list. This application probably would benefit most Intelligence Community analysts.

- Other types of users may be interested in capturing all duplicates, near-duplicates and similar documents. In this scenario a user tasked to process a collection of document consistently would want to identify all duplicate and similar documents for processing in the same or a similar manner. If the near-duplicate detection effort is successful, the resulting system would provide the user with this identification capability.

CONTEXT-DEPENDENT SUMMARIZATION

The third research area continues the objective of reducing the amount of text that the user must read by presenting summaries of long documents in lieu of the full documents. The summarization software will either provide a short summary for each document in a collection or one summary for an entire group of related documents. If the collection contains disparate document, the Cornell approach uses a preliminary step to group related documents and then applies the summarization algorithms to each group.

Summarization will be done in the context of the query. The Cornell system will capture only those features relevant to the user's information need. This is distinguishable from a generic summary that would capture the salient items for the entire document without regard to any particular search query or information need. For example:

> Suppose the target document contains information on political profiles, military status, weapons proliferation issues and economic changes about a country of interest. A good generic summary would contain the essential elements on all four topics. If the user's only interest is the second topic, as would be reflected in the user-defined query, then a good context-dependent summary would contain only those elements that are relevant to military status.

In a combined approach to summarization, the users will have the options of generic summaries or query-based summaries for each document in a collection or a cross-document summary for the entire collection. (See [3] for details on generic summarization.)

DEMONSTRATION SYSTEMS

The Cornell TIPSTER efforts will result in several proof-of-concept demonstration systems, one for each of the three major tasks. These demonstration systems for high-precision information retrieval, near-duplicate detection, and summarization will be sufficiently developed to allow the sponsor to evaluate their performance against real or simulated user data. Further development will be needed, particularly in the area of human to computer interfaces, to demonstrate the systems' utility to end users in an operational environment.

REFERENCES

[1] Cardie, Claire and Buckley, Chris, "Improving End-User efficiency in a TIPSTER-Compliant IR System," Proposal to the Defense Advanced Research Program Agency, 1996.

[2] Buckley, Chris, C. Cardie, J. Walz, S. Mardis, M. Mitra, D. Pierce, and K. Wagstaff, "The Smart/Empire TIPSTER IR System," in Proceedings TIPSTER Text Program (Phase III), 1999, this volume.

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