The TIPSTER Text Program Overview

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

INTRODUCTION

These TIPSTER Phase III Proceedings bring to a close a program that had significant impact on information technology. Since 1991, the TIPSTER Text program has fostered the advancement of state-of-the-art technologies for text handling through the efforts of researchers and developers in the U.S. Government, industry and academia. The resulting capabilities are being deployed throughout the intelligence community to provide analysts with improved information processing tools.

The TIPSTER Program focused on research and development in three technology areas: Document detection, information extraction and text summarization. In addition, the development of multilingual or cross-lingual capabilities in each of these areas became a vital component of the program. Metrics-based evaluations also constituted a critical program element. For example, the Text Retrieval Conference (TREC), TIPSTER's metrics-based evaluation for document detection is now recognized as the premier source of ground-truth data upon which information retrieval developers can test their systems. The TREC collection is a sizable one and provides the foundation for developers to test scalability of retrieval systems. The Message Understanding Conference (MUC), the Multilingual Entity Task (MET) and the Summarization Analysis Conference (SUMAC) also are TIPSTER evaluation mechanisms to help the Government gauge technology progress in relevant areas.

A summary of the accomplishments for the first two phases of the TIPSTER program is provided in the sections below. For more details, please see the proceedings for TIPSTER Phase I [1] and TIPSTER Phase II [2], respectively. The next article in this volume [3] provides the highlights from Phase III, the last phase of the TIPSTER program.

PHASE I ACCOMPLISHMENTS

The focus of TIPSTER Phase I was to advance the state of the art in two text-processing technologies, Document Detection and Information Extraction. Document detection included two subtasks: Routing (running static queries against a text stream) and ad hoc retrieval (running ad hoc queries against archival data). Information Extraction is a technology in which pre-specified types of information are located within free text, extracted and placed within such structured forms as templates which can be considered as pseudo-databases.

The algorithm development in detection and extraction during Phase I resulted in improvements in the technologies. As a result of TIPSTER advances, detection users had:

- Improved recall (the system retrieves more of the relevant documents available in the applicable document collection)
- Improved precision (the system returns to the user a higher percentage of relevant documents in the "hits" list, meaning that the user will read fewer documents in order to find the one he wants)
- Ranked retrievals (the user reviews documents statistically ranked according to how well they matched the query, thus improving the chances that the most relevant documents will be at the top of the hits list)
- Query expansion (the system would attempt to automatically expand queries to draw in more relevant documents by using concept based tools such as thesauri)
- Automatic query generation (the system uses a natural language description of the subject

supplied by the user, including example text, to generate queries)

The TIPSTER Program continued Government sponsorship of information extraction research. The extraction efforts, begun with the DARPA-sponsored MUC in 1987. sought to reduce the burden of tasks largely characterized by manual procedures and large resource investments both in terms of people and corresponding dollars. As a result of extraction algorithm development in Phase I, systems could be developed with:

- Increased scalability within a domain with reduced user involvement
- Increased ease of portability to new domains and languages (Phase I focused on two domains and two languages.)
- Greater task independence, solving multiple problems with reusable components.

In addition, the advances improved the ability of systems to users with

- Accurate and consistent database content results
- Minimal user intervention in reviewing extraction results
- Initial cost expenditures with reduced maintenance costs
- Flexibility in managing the amount of information to be extracted
- Applicability to new tasks such as text tagging and document detection support.

PHASE II ACCOMPLISHMENTS

The Government continued its sponsorship of information technology in TIPSTER Phase II. The participating agencies defined a two-tiered program of continued algorithm development and transfer of technology into demonstration projects. The Government, industry and academia continued their close cooperation and, based on Phase I experiences, crafted a four-part program. (See Figure 1.) While continuing the traditional focus on advanced research and metrics-based evaluation, the program supported the development of a common software architecture and applications of the technologies to help solve operational problems.

While advances were continuing in the technology areas, there was a growing need for interoperability among the diverse systems. The impetus for the architecture came from an analysis of the designs of Phase I systems and analysis of operational scenarios that indicated the complementary nature of detection and extraction operations. The Government sponsors wanted an architecture that would support both technology

TECHNOLOGY RESEARCH Research on the two underlying technology areas: - Document detection (and the more general category of information retrieval) - Text extraction For Phase III of the program, 15 research contracts were executed.	**METRIC-BASED EVALUATIONS** Development of evaluation methodologies and creation of data collections, with ground truth, to serve as the testbed for software systems in the three technology areas. The evaluations forums were: - Text Retrieval Conference (TREC) - Message Understanding Conference (MUC) - Multilingual Entity Task (MET)
TIPSTER ARCHITECTURE A framework to enable sharing and interchangeability of software modules developed under the TIPSTER program. The architecture documents provided standards for basic software components and specifications for interfaces between two components.	**DEMONSTRATION PROJECTS** Projects funded independently by various partner agencies to evaluate the technologies for transfer into the workplace. A total of 15 projects were executed in Phase II and several continued into Phase III.

Figure 1: Four Parts of the TIPSTER Program.

areas. An architecture working group (AWG), consisting of TIPSTER Phase II R&D contractors, was formed to address the issues of developing a common, open architecture. This architecture would provide the framework for interoperability between detection and extraction systems and for plug-and-play flexibility. The AWG, with the support of an independent System Engineering/Configuration Management contractor, drafted initial versions of a TIPSTER architecture.

To test the feasibility of applying the TIPSTER-developed algorithms to operational environments, individual Government agencies sponsored separate development projects. For each of these projects, a demonstration system based on the architecture and modules developed in the R&D tier was developed. Some 15 systems were developed and tested in operational environments at several Government agencies as a result of this effort. Identified needs for architecture and algorithm improvements or additional research were fed back to the R&D projects.

The research and development efforts of the TIPSTER Program Phase II included improvements of algorithms and research into combining the results of the application of diverse extraction and detection techniques. There were improvements in detection recall and precision. Automatic query generation and relevance ranking spread beyond the TIPSTER Program and began to be common features in commercial search engines. Extraction technology advanced to the point that for at least one task, named entity extraction, machine performance was nearly the same as human performance. Extraction robustness had improved to allow operational users to test systems on real-world problems to determine that automatic population of databases was indeed possible with this technology.

The program continued its primary sponsorship of both the Message Understanding Conferences and the Text Retrieval Conferences. This sponsorship was based on the belief that these forums for evaluation are essential to technology advances, synergistic interactions of conference participants and the continued success in TIPSTER research and development. The combined efforts of TIPSTER Phase I and Phase II effectively set the scene for the third—and final—phase of the TIPSTER Text Program.

REFERENCES:

[1] "Proceedings, TIPSTER TEXT PROGRAM (PHASE I)", Morgan Kaufmann Publishers, Inc., September 1993

[2] "Proceedings, Advances in Text Processing, TIPSTER PROGRAM PHASE II April 1994 – September 1996", Morgan Kaufmann Publishers, Inc., September 1996.

[3] F. Ruth Gee, "TIPSTER Phase III Accomplishments", Proceedings TIPSTER Text Program (Phase III), 1999, this volume.