TIPSTER PROGRAM HISTORY

Thomas H. Crystal

Advanced Research Projects Agency
3701 N. Fairfax Drive
Arlington, VA 22203
crystal@arpa.mil

The history of the TIPSTER Text program has multiple threads. And, as preparation of this report marks the end of Phase I of a two-phase program, part of the history is planning for Phase II, maintaining the successful threads.

One of the threads is the close cooperation of six government organizations in formulating and implementing the program. This included not only sharing in the program formulation and funding, but also sharing in the definition of the tasks, in the preparation of large text corpora used for development, and in the development and implementation of evaluation methodologies.

The second thread is the cooperation of the contractors in sharing ideas and resources while pursuing different, competitive approaches to the problems of text processing.

The third thread is the sponsorship of the international Message Understanding Conferences (MUC's) and Text Retrieval Conferences (TREC's). These conferences, which evaluated the state of the art and promoted text-processing R&D outside of the TIPSTER Text contracts, were organized by NRaD and NIST. MUC-1 and MUC-2 preceded and set the stage for TIPSTER, before the sponsorship of these conferences became part of the program.

Formulation of the Program.

The concept of the TIPSTER Text program was developed at ARPA beginning in June 1989, following the end of MUC-2. The promising results of that conference, along with an appreciation of the need for automated handling of large volumes of text, led to the formulation of a text-processing technology-development plan. ARPA approved funding for the program which, in taking some risks in developing the technology, could result in substantial benefits for facing the government's ever-growing need for sorting and analyzing large volumes of text.

Beginning in January 1990, a succession of meetings were held among government agencies interested in the development and use of text-processing technology. From this came the agreement for sharing the planning, funding, and execution of the program.

Significant decisions included having a program with (1) two-phases: two years of R&D into advanced algorithms, followed by two-years of development of prototype/demonstration systems; (2) separate focuses on detection (retrieval and routing) and on extraction (understanding); (3) emphasis on domain and language portability; (4) periodic evaluation of complete systems; (5) development by the government of large corpora for training and testing of corpus-based techniques and system development and evaluation.

In June of 1990, a BAA was published soliciting proposals for participation in the program. This led to the selection of three contractors for investigating different approaches to detection and another three for extraction. Clarification of the proposals, selection of sources and negotiating the contracts took most of fiscal 1991.

Phase I of the Program

Prior to the beginning of the contractors' work, the government began an intensive, two-year effort into the acquisition and preparation of annotated corpora for the project, as described in the separate sections on detection and extraction.

The virtual start of TIPSTER Text Phase I occurred at a kickoff workshop held in September of 1992. The government reviewed the framework, objectives and plans for the following two years of work. The contractors described their specific approaches to detection and extraction and laid the groundwork for the future sharing of ideas and of software and data resources. The workshop included parallel working sessions for discussion of specific issues in the different areas of research, including details for addressing the different domains and different languages and the government's preparation of the data.

The workshops were repeated at 6-month intervals for the duration of Phase I. Selected researchers from other ARPA Human Language Technology (HLT) programs were also invited. In connection with the 12-month, 18-month and 24-month (final) meeting, uniform evaluations of system performance were conducted and reported at the meetings. Between meetings, there were frequent exchanges of infor-

mation among the government and the contractors (with heavy use of electronic mail) and, by the end of the two years, a sizable catalog of shareable resources had been developed.

The availability of dual-use funding permitted the addition of an additional detection and an additional extraction contractor for the final year of Phase I.

TIPSTER Text contractors were required to participate in MUC or TREC. MUC-5 and TREC-2, using TIPSTER evaluation techniques, were held to coincide with the final Phase I evaluation so as to provide a measure of the state of the art and identify good performers. Phase I of the TIPSTER program concluded with the 24-month workshop.

Phase II Planning

During the last year of Phase I, the government began planning Phase II. Scenarios were developed to indicate the variety of actual applications of the systems to be developed. A two-tiered program of (1) continued algorithm development and (2) transfer of technology into demonstration projects was defined

The management of Phase II will follow the successful threads of Phase I. There will be close cooperation among the government agencies and the contractors. There will be regular workshops, corpora for development and testing and periodic evaluations. MUC and TREC will be continued and there will be increased interaction with the ARPA HLT community. Contractors will share software using a license developed by that community.

The design of the Phase I systems and analysis of the scenarios indicated the complementary nature of detection and extraction operations and the desirability of supporting both capabilities within a single system. There also appeared to be many similar modules in the diverse systems. From this, it was determined that an initial activity of Phase II will be the development of a common, open software architecture for the implementation of text-processing systems. This architecture will also facilitate sharing of the development tasks, transferring technology to actual applications, future R&D into improved algorithms and continuous upgrading of systems which use the architecture. The architecture will stress functional and knowledge-based modularity and will use an SGML-like language for tagging text transferred between the modules.

This architecture will be developed as part of Phase II R&D through the cooperative efforts of multiple contractors, coordinated by an independent Systems Engineering/Configuration Management contractor. The R&D contractors will then be tasked to fit their system modules into the architecture they have designed. Phase II R&D will also include improvement of algorithms and research into combining the results of the application of diverse extraction and detection techniques.

A BAA soliciting proposals for participation in Phase II R&D was issued in August 1993, with responses due from the bidders in October. Selection and negotiation is planned to be completed so that work may begin early in the spring of 1994. Bidders will also be judged as potential sources for the demonstration projects. Individual agencies will issue separate RFP's for each such project. For each project, a demonstration system based on the architecture and modules developed in the R&D tier will be developed, installed and evaluated in the processing of actual "operational" data. Needs for architecture and algorithm improvements or additional research will be fed back to the R&D projects.

Procurement and award of the demonstration projects will be coordinated with the development of the architecture.