FOCUS OF TIPSTER PHASES I and II

Patrick J. Altomari
Patricia A. Currier
U.S. Department of Defense
Ft. George Meade, MD 20755

Background of Phase I

The TIPSTER Program began in June 1989 just after the conclusion of the second Message Understanding Conference (MUC-2). The concept which the Advanced Research Projects Agency (ARPA) put forward at that time was based on the promising results of that conference and on the belief that the technologies being demonstrated at MUC for the automated handling of large volumes of text would be of great benefit to a variety of Government agencies.

Following a series of meetings, agreement was reached for sharing the planning, funding, and execution of the program. Several major concepts were also outlined and accepted as central to TIPSTER's progress. Initially two phases were planned: two years of research and development into advanced algorithms, followed by two years of development of prototype/demonstration systems. Within these two phases, there would be separate focus on detection (retrieval and routing) and on extraction (understanding). Portability with regard to domain and language would be emphasized; evaluation of complete systems would be encouraged and scheduled periodically; and, as part of the baseline for these evaluations, the Government would develop a large corpora for the training and testing of corpus-based techniques as well as for system development and evaluation. Proposals were solicited in June of 1990, and eventually three contractors were selected to investigate different approaches to detection and another three were selected for extraction research.

The Pre-Existing State Of The Art

The two language technologies which ARPA initially hoped to advance through the TIPSTER Program were Document Detection and Information Extraction.

Document Detection includes both routing, which involves running static queries against a stream of new data, and retrieval, which involves running ad hoc queries against archival data. Prior to TIPSTER, most analysts faced with finding essential information from large volumes of data used search systems based on Boolean keywords. These systems had been developed more than a decade earlier and typically had the following characteristics:

- the user loses an unknown quantity of useful information because the system is unable to retrieve many of the relevant documents (= low recall);
- the user must read a very large number of irrelevant documents which the system mistakenly retrieved (= low precision);
- the user must scan the entire list of retrieved documents because a good document is just as likely to be at the end of the list as at the beginning (= no ranking or prioritization);
- the user must generate by hand all variant spellings or alternate word choices because there are no built-in rules for adding variants;
- the user has to understand how the system works and the syntax of queries in order to use the system (= hand built queries).

Information Extraction is a technology in which pre-specified types of information are located within free text, extracted, and placed within a database. Notwithstanding prior ARPA and commercial support for the development of information extraction technology and MUC's positive impact, information extraction had been applied to the database update task as largely a manual procedure. This procedure was costly in terms of labor commitment and the training required and there was wide variance in the accuracy and consistency of the database content. The techniques typically only worked for English, and were difficult to port to new domains or even to extend within the current domains. Although a few automated information extraction systems had been deployed, they tended to be expensive both in terms of development and ongoing maintenance and were task specific with little reusability.

Review of Phase I

The TIPSTER Program officially began at a kickoff workshop held in September of 1991. 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 each of the workshops, uniform evaluations of system performance were conducted and were reported during the meetings. There were frequent exchanges of information among the government and contractors (with heavy use of e-mail) and, by the end of the two years, a sizable catalog of shareable resources had been developed.

Both the Message Understanding Conferences (MUC's), which preceded TIPSTER, and the Text REtrieval Conferences (TREC's) evaluated the state of the art and provided a major additional benefit in promoting text-processing research and development outside of the TIPSTER Text contracts, since they were organized by NRaD and NIST and advertised to a wider community. All of the TIPSTER Text contractors were required to participate in MUC or TREC, and MUC-5 and TREC-2, using TIPSTER evaluation techniques, were intentionally timed to coincide with the end of Phase I (the 24-month workshop) so as to provide a measure of the state of the art and identify good performers.

TIPSTER Phase II

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 successful concepts of Phase I were continued, with close cooperation among the government agencies and the contractors; regular workshops; and corpora for development and testing. Based on a variety of lessons learned and encouraged by the results of TIPSTER Phase I, a four part program began to informally take shape. While continuing its traditional focus on advanced research and metrics-based evaluation, the need for a supporting architecture was recognized, along with the realization that many of the techniques developed were now sufficiently stable to be applied in an operational environment as demonstration projects.

1. Research

Document Detection. The TIPSTER Phase II prototype systems gave the user Document Detection tools which feature the algorithms and technology developed in Phase I. There has been improved recall (higher recall of relevant documents) and improved precision (the user reads fewer useless documents in finding the ones he wants). Moreover, the system, not the user, can now automatically expand the queries to draw in more relevant documents (using concept based tools such as thesauri, or using a natural language description of the subject supplied by the user), and the documents are statistically ranked according to how well they match the query, thus improving the chances that the most useful documents will be near the top of the queue.

Information Extraction. As a result of algorithm development in Phase I, TIPSTER Phase II prototype systems were built with the following characteristics:

- increased extensibility within domain with reduced user involvement
- greater ease of portability to different domains
- language independence and portability to new languages
- task independence, solving multiple problems with reusable components
- user-focused maintenance with minimal system developer involvement

These systems provide the user with extraction tools which feature:

- accurate and consistent database content results
- minimal user intervention in reviewing extraction results

- initial cost expenditure with little maintenance cost
- flexibility in managing the amount of information to be extracted
- applicability to new tasks, such as indications/warnings, text tagging, and document detection support

2. Architecture Concept

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 would be the development of a common, open software architecture for the implementation of text-processing systems. This architecture was to 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 stresses functional and knowledge-based modularity and uses an SGML-like language for tagging text transferred between the modules.

This architecture was 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 included improvement of algorithms and research into combining the results of the application of diverse extraction and detection techniques.

3. Continuing Evaluation

During Phase II, TIPSTER became primary sponsor for both the Message Understanding Conferences and the Text Retrieval Conferences, based on the belief that these forums for evaluation of text-processing technologies are essential to continued success in TIPSTER research and development.

4. Demonstration Projects

A Broad Agency Announcement soliciting proposals for participation in Phase II R&D was issued in August 1993. Bidders were judged based on their ideas for research as well as their being potential sources for the demonstration projects. Individual agencies issued separate Requests for Proposals for each such project. For each project, a demonstration system based on the architecture and modules developed in the R&D tier was developed, installed and evaluated in the processing of actual "operational" data. Needs for architecture and algorithm improvements or additional research were fed back to the R&D projects.

A Successful Partnership

The continued close cooperation of multiple government organizations in formulating and implementing the TIPSTER program has been a major ingredient in its success. By regularly providing a forum for discussion, the Program has also fostered cooperation among an ever-expanding group of academic institutions and industry vendors, who have shared ideas and resources while pursuing different approaches to the problems of text processing. The use of these technologies will undoubtedly expand well beyond the prototypes and operational systems built during TIPSTER Phase II for a small number of Government agencies, as the world at large recognizes the need for document detection and information extraction.

Conclusion of Phase II

The 24-Month Workshop which concluded TIPSTER Phase II brought together a large number of researchers and developers to discuss their results and describe their progress since 1994 and to present their findings to a variety of potential customers. The papers summarizing the efforts sponsored by the TIPSTER Program are included in this volume.

Acknowledgments

Portions of the information above were originally in papers prepared by Roberta H. Merchant and Thomas H. Crystal and published in the TIPSTER Phase I Proceedings.

We also wish to thank Wiley Harris and Sharon Kaufmann for their suggestions and assistance in assembling and organizing all of the materials presented in this summary book for TIPSTER Phase II.