TIPSTER Phase III Accomplishments

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

INTRODUCTION

The TIPSTER Text Program Phase III continued the sponsorship of research and development to advance state-of-the-art technologies for text handling and facilitation of cooperation among research and development components from industrial, academic and the U.S. Government.

The TIPSTER Phase III formally started with a Kick-off Workshop held in October 1996. Specific Phase III goals were to:

- Sustain the successes of Phase I and II in detection, information extraction, architecture, and formal evaluation;
- Push research in text processing technologies;
- Expand the architecture;
- Increase participation with Government agencies, researchers, developers, and the academic community in general;
- Expand multilingual and summarization efforts.

The overall purpose was to field a system for use within the operational elements of the Intelligence community and other Government agencies. Phase III Government participants included the Defense Advanced Research Projects Agency (DARPA), the Central Intelligence Agency (CIA), the National Security Agency (NSA), the National Institute of Standards and Technology (NIST), the Naval Research Lab (NRL), the Air Force Research Lab (AFRL), the Space and Naval Warfare Systems Command (SPAWAR) and the Defense Intelligence Agency (DIA)

RESEARCH

The 15 research projects sponsored by the Government for TIPSTER Phase III[1] built on the advances made in information extraction and detection, but also initiated research in text summarization. Furthermore, cross-technology issues played a bigger role among the research efforts of many Phase III participants. Short descriptions of the 15 research projects can be found in Figure 1. Additional details on most of these projects can be found in the Phase III papers included in this volume.

Participant research in **extraction** centered, in general, on three areas: accuracy, usability, and portability. In order to advance the state of the art, extraction researchers focused core technological efforts on developing algorithms to, for example resolve coreference and use machine learning or related techniques to acquire patterns semi-automatically. The ultimate goal was to push precision and recall in the scenario task to operationally usable levels.

The common pattern specification language (CPSL) was to be used to facilitate the porting of extraction systems or modules to new domains and languages. Although, , this objective was not fully realized, due to funding constraints, SRI implemented

[1] The 15 research projects referenced in Figure 1 do not include two projects that were selected but not funded by DARPA initially. The two projects, "Cross-Language Document Retrieval with Latent Semantic Indexing (University of Colorado) and " Multilingual Interactive Document Summarization (MINDS)" (New Mexico State University) were funded by ORD after TIPSTER Phase III began.

CPSL to develop a new extraction system called TextPro [1].

For the usability focus, some work focused on determining the optimal role of the user during operational deployment of the technology.

Detection research focused on advancements in the technology and usability. On the technology side, researchers pursued such topics as the appropriate role for Natural Language Processing (NLP) in detection, the usefulness of shallow extraction in indexing and retrieval, foreign language retrieval, combining different retrieval engines, and the use of machine learning and case-based reasoning.

On the usability side, Phase III detection participants investigated optimal query building approaches to capitalize on the role of the human in the concept of operations.

Usability issues also figured prominently in **text summarization**, the newest area of TIPSTER-sponsored research that had its beginning in Phase III. While the focus was on transitioning "enabling" technologies from detection and extraction, researchers exploring different strategies for identifying applicable analytic tasks, and assessing the near-term usability of various strategies for user-centric summarization.

Using both statistical and natural language processing techniques, summarization provides a systematic means to reduce the volume of a full text document without losing relevant content. This technology could be applied to a variety of tasks in order to assist an information searcher. In TIPSTER Phase III, the Government sponsored several research and development efforts, each with different approaches and potential uses for automatically produced text summaries.

Summarization, due to the multifaceted nature of its output and fluidity of definition, quite naturally employed a **cross-technology** approach. Phase III participants leveraged their entity-centered extraction and sentence-level detection methodologies in developing core summarization systems.

We witnessed other cross-technololgy advances. Detection research involved a more pronounced role for NLP, such as shallow extraction in indexing. In a similar fashion, extraction researchers explored the use of detection techniques, such as filtering to improve accuracy. We projected that, had the Architecture Capabilities Platform reached a sufficient level of maturity, the cross-technology approach would have garnered additional advances through the interchange of intermediate results between multiple engines and technologies.

ARCHITECTURE DEVELOPMENT

The Architecture Capability Platform.

The Architecture Capabilities Platform (ACP) was a TIPSTER Phase III effort to support the evaluation, extension, and exploration of the evolving TIPSTER Architecture. The TIPSTER Program goal was that the ACP would provide an Internet-based toolbox of components for researchers and developers, and a test-bed for proposed Architecture changes. In addition, the ACP was to:

- Promote reuse of components and data developed during previous TIPSTER efforts, making research and demonstration projects, and evaluation efforts like TREC and MUC easier to obtain and integrate.
- Increase the viability of the TIPSTER Architecture beyond the current community.
- Provide a way to create distributed systems, without requiring changes to existing components. The ACP approach employs the Common Object Request Broker Architecture (CORBA), a commercial standard for distributing object oriented systems like the TIPSTER demonstration systems.
- Facilitate data exchange between TIPSTER systems and other Information Retrieval (IR) systems. The ACP pursued this goal by implementing software to allow TIPSTER and Z39.50 interoperability.
- Provide a platform for examining and evaluating proposed Architecture changes in a real-world setting.

Architecture Working Groups.

At the beginning of Phase III, there are many issues which needed resolution to refine and extend the TIPSTER Architecture to meet the needs of the growing range of applications. To address these

CONTRACTOR	DESCRIPTION OF EFFORT	Detection	Extraction	Human Machine Interface	Multi-lingual	Summariza-tion
BBN/Systems and Technologies	Merging & Anaphoric Resolution		■			
Carnegie Group, Inc.	Maximal Marginal Relevance					■
Cornell University	Duplicate Document Detection, Summarization	■				■
General Electric	Advanced NLP for Accurate and Flexible Indexing	■				■
Lockheed Martin	Coreference, TimeTool, and Document Intent		■			
New Mexico State University	Multilingual IR	■		■	■	
New York University	Enabling User IE Customization		■			
Queens College (City Univ. of NY)	Chinese IR & Evidence Combination	■			■	
Rutgers University	Multiple Information Seeking Strategies	■		■		
Systems Research and Applications Corp	Extraction & Customization by Machine Learning		■			
SRI International	Open Domains, Learning by Example, Coreference		■			
TextWise	Combination Retrieval	■				
University of Massachusetts	Merging, Routing, Filtering, and Topic Clustering	■				
University of Pennsylvania	Coreference Engine Summarization					■
University of Southern California	Text-Generated Summarization				■	■
		Architecture Development			**Architecture Support**	
Logicon, Inc.	Architecture Capabilities Platform	■	■	■		
Litton/PRC	System Engineering and Configuration Management				■	■

Figure 1. TIPSTER Phase III R&D Projects

issues, the Architecture Committee (AC) created a number of Technical Working Groups (TWGs) which included representatives of the Government, TIPSTER contractors and others involved in Tipster development. Four new working groups joined the Pattern Specification TWG, formed under Phase II of TIPSTER. Goals of the five TWGs are summarized below.

Pattern Specification: This TWG sought to develop a common notation to exchange information about patterns among information extraction developers.

Most information extraction systems operate through a process of pattern matching: successive stages of patterns are used to identify successively larger linguistic units. In the past, each contractor had used their own notation for these patterns and provided different pattern-matching capabilities which made it harder to achieve a "plug and play" architecture goal. A paper on the findings of this TWG can be found in this volume [2].

Annotation Standardization: The primary means by which text analysis components communicate in the TIPSTER Architecture is through annotations on documents. The Annotation Standardization TWG aims to define standard annotations
for document structure (title, source, author, date, body, etc.), for tagging names in documents, and for encoding information extraction templates as annotations.

Linking/Tagging: This TWG considered the mechanisms for linking together the copies of a document and for propagating particular attributes onto all the copies-- attributes needed for security classification or copyright, for example. This effort was eventually folded into the Annotation Standardization TWG.

Document Management: The architecture design developed under TIPSTER Phase II defined the functionality needed for document management
in single-user, single-process environments. When the Architecture was used in multi-process or multi-user applications, local extensions were made in such areas as protection and concurrency control. The document management TWG attempted to standardize these extensions.

Detection: This TWG sought to address capabilities that needed to be added to the detection part of the Architecture, such as the ability to view queries created by relevance feedback and automatic query generation. it also addressed new issues associated with the extension of the Architecture to use the Z39.50 standard for client-server communication in retrieval systems.

Architecture Mixed Results

The TIPSTER architecture in general and the ACP in particular did not achieve its intended goals. Part of this was due to the early demise of the TIPSTER Program and part was due to the Government's inability to enforce standards imposed by the TIPSTER software architecture. These negating factors were compounded by the fact that the architecture was not fully developed and the earlier versions were not fully supported by the developmental efforts. At the premature end of TIPSTER Phase III, the ACP was not sufficiently developed to truly test interoperability of software modules. In addition, the Government did not always insist on the TIPSTER architecture being implemented in the demonstration system. In some cases, it was not feasible to do so, especially for those projects that had begun before the architecture was sufficiently completed.

A notable success indicated that an architecture like TIPSTER's is workable, despite the setbacks. The University of Sheffield designed and implemented the General Architecture for Text Engineering (GATE) and used the TIPSTER architecture for its foundation. GATE is now in use extensively Europe. It was also used in the ACP so that the ACP could be delivered in a useful form at the end of Phase III, given the fact that TIPSTER ended early. GATE represents a success story for TIPSTER and illustrates one of the many examples of the program's impact on the commercial world.

A major lesson learned concerns that of the inability of a small Government-sponsored effort to influence industry standards. The focus should lie not in formally establishing architectures but in establishing Government business drivers and working with industry and commercial focus groups, where possible, to steer development in directions of benefit to the Government. See [3] in this volume for other lessons learned from TIPSTER architecture efforts.

EVALUATION

The Text Retrieval Conferences

Since the beginning of the TIPSTER program, there have been seven Text REtrieval Conferences (TRECs). The number of participating systems has grown significantly since TREC-1 and has, across the years, included many of the major text retrieval software companies and most of the universities doing research in text. A combined TREC roster from the seven past conferences contains participants from several foreign countries. The TIPSTER sponsors encouraged this international participation and worked toward the continuation of

the TREC resources, despite the formal end of the TIPSTER program. The diversity of the participating groups has ensured that TREC represents many different approaches to text retrieval, while the emphasis on individual experiments evaluated in a common setting has proven to be a major strength of TREC.

The test designs for the various TRECs have been similar. The participants ran the various tasks, sent results into National Institute of Standards and Technology (NIST) for evaluation, presented the results at the TREC conferences, and submitted papers for proceedings. The main test collection currently consists of over 1.6 million documents from diverse full-text sources, 300 topics and the set of relevant documents or "right answers" to those topics. This test collection supports the main TREC tasks of routing and ad hoc retrieval.

In addition to the main test collection, there are smaller test collections in Spanish and in Chinese. Also, TREC has sponsored several focused research tasks, called tracks. In TIPSTER Phase III, these have included

- Filtering Track
- Cross Language Information Retrieval (CLIR) Track
- High Precision Track
- Interactive Track
- Very Large Corpora (VLC) Track
- Spoken Document Retrieval (SDR) Track.

TREC has proven to be very successful, allowing broad participation in the overall DARPA TIPSTER effort, and causing widespread use of very large test collections. All conferences have had very open, honest discussions of technical issues, and there have been large amounts of "cross-fertilization" of ideas. TREC has received world-wide recognition as an evaluation resource for information retrieval systems. DARPA, NIST and other Government partners have continued their sponsorship beyond TIPSTER. See [4] for details of TREC-7, the last TREC sponsored by the TIPSTER program.[2]

[2] TREC, however, is continuing beyond the TIPSTER Program with TREC-8 scheduled for November 1999.

The Message Understanding Conference

The goal of the Message Understanding Conferences (MUCs) was to push information extraction systems toward improved accuracy and greater portability to new domains and to encourage basic research by providing evaluations of some basic language analysis technologies. There was a set of five evaluation tasks:

- **Named Entity** Task (NE): Recognition of entity names for people and organizations, place names, temporal expressions, and certain types of numerical expressions.

- **Coreference Task** (CO): Identification of coreference relationships among noun phrases..

- **Template Element Task** (TE): Information extraction about specified class of objects and filling of template for each instance of each such object.

- **Template Relationship Task** (TR): Information extraction about specified class of relationships between template elements and filling of template for each instance of each such relationship with pointers to template elements.

- **Scenario Task** (ST): This task combines the elements of the other four tasks and focuses on event-centered information extraction in a specific domain.

The first four tasks are independent of any particular domain. The last is equivalent to traditional information extraction. The NE and CO tasks entailed Standard Generalized Markup Language (SGML) annotation of texts. The template element, template relations, and scenario template are information extraction tasks where template slots are filled with extracted, categorized, or normalized information that might go into a database. See [5] for details on MUC-7, the last MUC sponsored by the TIPSTER Program. An Internet web site at www.muc.saic.com contains additional details on MUC-7.

Multilingual Evaluation Task

The Government sponsors of the second Multilingual Entity Task (MET) collected Chinese

and Japanese data for MET-2 Named Entity task. Each collection contained over 300 articles (including revised versions of MET-1 data) tagged appropriately for training data. Unfortunately, the Government group did not have sufficient staff to support timely data collection and preparation to continue the Spanish language thrust from MET-1 but some Thai data was provided for initial experimentation . See [5] for discussion of MET procedures and MET-1 results and [6] for details on MET-2.

MET-2 represented a somewhat richer variety of language patterns than the MET-1 data, which was collected from only a single newswire source in each language. The training collection included data from three Chinese and two Japanese sources. Whereas MET-1 training, dry run, and formal test data was retrieved using a single set of keywords, MET-2 used different keywords to select each data set. Consequently, participant systems were challenged to demonstrate greater portability in covering multiple text sources and domains.

Although the multilingual task was confined, as in MET-1, to Named Entity extraction, texts were selected according to their suitability for future Template Element and Scenario Template applications.

The Government component of TIPSTER began a campaign to acquire newly available resources for the community in support of the multilingual information extraction tasks. In particular, since MET-1 the Government group has acquired two online part-of-speech tagged Chinese lexicons, the larger of which differentiates 39 morpho-syntactic categories in glosses of over 100,000 terms.

Because segmentation (finding word-boundary) has proven to be a bottleneck problem for IE tasks in various non-Roman languages, the Government group developed a second segmentation tool to help identify proper names, technical terms, newly coined words, etc., that may be missing from the lexicon. This tool utilizes a core lexicon of only 5000 terms, selected for their high-frequency occurrence in newspaper text.

In addition, TIPSTER industrial and academic partners contributed generously to help improve the existing capabilities of tools that support the labor-intensive process of data collection and mark-up. For example, a revised version of the NMSU Chinese segmenter was made available.

The Government group played a key role in advancing participants' technical capabilities by serving as a clearing-house for basic multilingual text processing resources such as segmenters, dictionaries, and tagging tools and by encouraging participants to share basic techniques, tools, and data to support the multilingual extraction effort.

Summarization Analysis Conference

The first Government sponsorship of summarization evaluation occurred in Phase III and took the form of the Summarization Analysis Conference (SUMMAC). SUMMAC included several tasks intended to judge the utility and appropriateness of the generated summaries and to provide a way to measure improvement consistently. The tasks focused on the relevancy of user-directed summaries, as compared to similar relevance judgments using the full text of a document.

The growth in the Internet and in World Wide Web use has resulted in a dramatic increase in electronically available information. This same information explosion is duplicated in office environments. The sheer magnitude of the information overload has forced information managers to investigate alternative means of data presentation. Summarization technology, applied at different steps in a traditional text processing flow, has the potential to effectively and accurately reduce the volume of information presented to a user by as much as 60-80%.

If summarization evaluation continues beyond TIPSTER, additional tasks are needed to address the ability of systems to extract specific items of information in a "question and answer" scenario.

DEMONSTRATION SYSTEMS

TIPSTER Phase III participants delivered many R&D systems that have been used by the Government sponsors to showcase advances in the detection, extraction and summarization technologies. Many Government agencies, building on the successes and lessons learned from Phase II and III, now have TIPSTER-enhanced systems deployed in an operational environment. See Section C of this

volume for discussion on a few of these systems from the Government's perspective. Other papers in this proceedings also will contain information on TIPSTER -sponsored systems for Phase III.

THE PROGRAM ENDS

The formal sponsorship of TIPSTER ended with the final program workshop on 15 October 1998 but collaboration continues among many of the Government, industrial and academic partners. Beyond the end of the program, we will continue to track the impact of TIPSTER by documenting the commercial products and Government deliverables that have roots in the TIPSTER Program research and development.

The work started in TIPSTER has recently expanded to an increased multilingual focus in the new DARPA sponsored program, Translingual Information Detection, Extraction and Summarization (TIDES).

ACKNOWLEDGEMENTS

We thank the TIPSTER program founders for their foresight in addressing the information overload problem through the sponsorship of this program. We thank the many Government, industrial and academic participants who made all phases of this program a success. Despite the early end of TIPSTER Phase III, the R&D efforts exemplified by the papers in the remainder of these proceedings--and the contributions by participants to the advancements in the state-of-the art in text processing--are legacies in which all the TIPSTER Government, industry and academic partners can take great pride.

REFERENCES:

[1] Steven Maiorano, "The SRI TIPSTER III Project", Proceedings TIPSTER Phase III, 1999 (this volume).

[2] Douglas E. Appelt and Boyan Onyshkevych, "A Common Pattern Specification Language", Proceedings TIPSTER Phase III, 1999 (this volume).

[3] Harold Corbin and Aaron Ternin, "TIPSTER Lessons Learned: The SE/CM Perspective", Proceedings TIPSTER Phase III, 1999 (this volume).

[4] Ellen M. Voorhees and Donna Harman, "The Text Retrieval Conferences (TRECs)", Proceedings TIPSTER Phase III, 1999 (this volume).

[5] "Multi-lingual Entity Task", Section H, Proceedings TIPSTER Text Program (Phase II), September 1996.

[6] Elaine March, "TIPSTER Information Extraction Evaluation: The MUC-7 Workshop", Proceedings TIPSTER Phase III, 1999 (this volume).

[7] TIPSTER Spring 1997 Brochure