TIPSTER PROGRAM OVERVIEW

Roberta H. Merchant

US Department of Defense
Ft. Meade, MD 20755
rhmerch@afterlife.ncsc.mil

1. TIPSTER PHASE I

The task of TIPSTER Phase I was to advance the state of the art in two language technologies, Document Detection and Information Extraction.

Document Detection includes two subtasks, routing (running static queries against a stream of new data), and retrieval (running ad hoc queries against archival data).

Information Extraction is a technology in which prespecified types of information are located within free text, extracted, and placed within a database.

2. THE STATE OF THE ART IN DOCUMENT DETECTION BEFORE TIPSTER

Before TIPSTER users searching large volumes of data and using many queries had few information retrieval tools to use other than the boolean keyword search systems which had been developed more than a decade earlier. The characteristics of these boolean systems are:

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

3. DOCUMENT DETECTION DELIVERABLES IN PHASE II

As a result of algorithm development in Phase I, during TIPSTER Phase II, prototype systems will be built, giving the user Document Detection tools which feature the technology developed in Phase I:

- improved recall (comparative evaluation of systems in TIPSTER and TREC [1] has demonstrated higher recall of relevant documents)
- improved precision (the user will read fewer useless documents in order to find the ones he wants)
- ranked retrievals (the user reviews documents 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)
- query expansion (the system, not the user, automatically expands 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 to generate queries)

4. THE STATE OF THE ART IN INFORMATION EXTRACTION BEFORE TIPSTER

Notwithstanding ARPA and commercial support for the development of information extraction technology and the positive impact of the series of Message Understanding Conferences, before TIPSTER, information extraction had been applied to the database update task as largely a manual procedure. Manual extraction is characterized by:

- wide variance in the accuracy and consistency of the database content
- heavy labor commitment
- continuing cost expenditure and training demand

- difficulty of porting to new domains
- difficulty of extending within current domain

The deployment of information extraction systems was rare for both commercial and Government applications. Such systems have been characterized by

- lack of extensibility within domain
- lack of portability to new domains
- language dependency (English only)
- task dependency, solving a single problem with little reusability
- high development cost with only system developer maintenance

5. INFORMATION EXTRACTION DELIVERABLES IN PHASE II

As a result of algorithm development in Phase I, during TIPSTER Phase II, prototype systems will be built with the following characteristics:

- increased extensibility within domain with reduced user involvement
- greater ease of portability to new different domains
- language independence, 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

[1] The Text REtrieval Conferences (TREC's) are described in the Document Detection section.