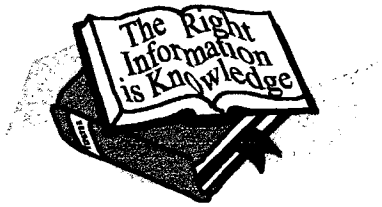
WELCOME to TIPSTER

TIPSTER is a Program that encourages the advancement of state-of-the-art technologies for text handling through the cooperation of researchers and developers in government, industry and academia. The resulting capabilities are being deployed within the intelligence community to provide analysts with improved operational tools.



Text Processing

The TIPSTER Text program is an inter-agency one, based on the common needs within the Intelligence Community and U.S. Government for automating the handling of textual data. It is jointly managed and funded by Defense Advanced Research Agency (DARPA), the Central Intelligence Agency (CIA) and the National Security Agency (NSA) in close collaboration with the National Institute of Standards and Technology (NIST), the Defense Intelligence Agency (DIA), and the Naval Command, Control, and Ocean Surveillance Center (NCCOSC). By cooperating closely in defining and procuring against these common needs, the Intelligence Community agencies have been far more successful than they could have been working separately. TIPSTER, through this focused and cooperative program management, has achieved considerable influence throughout the research sector, with substantial impact in the development sector and the commercial world as well. These sectors are now making substantial contributions to the Intelligence Community on text handling issues. There is a growing cadre of firms capable of using TIPSTER advanced technology in their applications.

The TIPSTER Text Program supports technology which addresses the problem of finding the information the user wants in machine readable text. Phases I and II of TIPSTER focused on two core text handling technologies, Document Detection and Information Extraction, which are central to the Intelligence Analyst's job. Document Detection, with an expression of the user's need, finds relevant text documents either from an incoming message stream or from a retrospective data repository. Information Extraction locates and identifies requested information within text documents, including such things as names of persons, places, organizations, equipment, activities, times, and relationships among any of these.

An enabling technology being developed by the TIPSTER Text Program is the TIPSTER Architecture. The architecture will provide a vehicle for delivering TIPSTER Text Document Detection and Information Extraction methods to the Intelligence Analyst and offer a convenient and efficient framework for research in the text handling technologies.

Fifteen TIPSTER systems are being developed for or have been deployed in operational environments to address real world problems. As operational prototypes, these systems are demonstrating the positive improvements in performance and accuracy which can be obtained by quickly identifying and implementing promising research methods. TIPSTER systems are in the forefront of those exploring the use of the most advanced retrieval methods.

Measuring Progress

The TIPSTER program encourages advances in technology by conducting frequent formal evaluations using realistic, well defined tasks that apply meaningful evaluation methods. The Message Understanding Conferences (MUC) provide a forum for assessing and discussing progress in the field of natural language processing. The most recent MUC increased the number of tasks (i.e., Named Entity, Information Extraction) used to compare systems. The Multilingual Entity Task (MET) debuted at MUC-6 permitting the assessment of systems on the same task in Spanish, Japanese, and Chinese. The Text REtrieval Conferences (TREC) use millions of articles, hundreds of queries and a variety of scoring techniques to evaluate the performance of Information Retrieval technologies. To further increase participation, TREC added more document collections, new topics and six special interest tracks (including ad-hoc, confused text, and foreign language). The conferences attract international participation from academia, business and government. The formal evaluation

of systems submitted by conference participants encourages continuous improvement in the accuracy and completeness of text handling technology.

Accomplishments

Phase I of TIPSTER resulted in the improvement of Document Detection technologies in three main areas: improvements in Recall of research systems from roughly 30% Recall to as high as 75% Recall; improvements in the processing of natural language queries (or statements of detection need); and the development of a variety of detection strategies to apply against different detection problems. Improvements in Information Extraction as a result of TIPSTER research are: increases in Recall and Precision from roughly 49% to 65% (Recall) and 55% to 59% (Precision); and increases in the ability to identify a wide spectrum of generic items such as personal and organizational names, dates, locations, equipment, times, phone numbers, and passport numbers.

The first year of Phase II of TIPSTER, 1994-1995, resulted in the development and test implementation of the TIPSTER software Architecture. The purpose of this architecture is to allow Detection and Extraction technologies to be easily deployed jointly and to work synergistically together. The Architecture will also support component and module substitution, some software reuse and sharing among offices and agencies, and incremental growth within deployed systems. Version 1.5 of the Architecture was released in May 1995 and is being tested. Applications built using this Architecture for the text handling portions of their functionality will contribute to the overall pool of software available to other government users of the Architecture. These applications will themselves benefit by the option of using capabilities developed elsewhere under the TIPSTER Program at reduced cost and time.

The research sponsored by the TIPSTER Program has provided advanced capabilities leading to the commercialization of new or improved products. These products are being licensed for use in commercial and government organizations.

The National Performance Review has listed the TIPSTER Program as a National Reinvention Laboratory for many years. Recognizing its teamwork, its customer focus, and the fact it has broken down existing bureaucratic barriers. The Vice President of the United States personally presented to the TIPSTER Text Program the National Performance Review Hammer Award in recognition of its contributions to the reinvention of government. He noted that the advances made by TIPSTER research are not only "revolutionizing" the Intelligence Community, but are being made available through commercial offerings to non-government users as well.

Pursuing the Future

TIPSTER Phase III, starting in October 1996 will continue to support the evaluation conferences (MUC, MET & TREC) and to fund and encourage advanced research. The TIPSTER architecture will be enhanced and refinements added to position it better in an object oriented framework. The capabilities platform and continued support of operational prototype systems are also prime objectives for TIPSTER Phase III.

Vision

As the TIPSTER community continues to advance the state of the art in text processing technologies, the focus now turns more to the operational use of the technologies and tools being created. The Flexible Detection Scenario and The Advanced Analytical Scenario which follow provide insight into the world of future users and how they will benefit from TIPSTER's ongoing research and development.

Flexible Detection Scenario

This scenario is intended to describe how an analyst in the very near future might use technology that is not much advanced from today's capabilities, if this technology were properly available on the analyst's desk top. It is provided not only as a near term vision of how the technology can be made useful, but also as an indication of the kinds of processes that still need to be streamlined and made more automated, further into the future. It is intended specifically to illustrate that there is no one or single way of using detection for the analyst; the more flexibility provided to the analyst, the better she can pursue her intuitions, trains of thought, and questions. The analyst learns about her sources and her subject as she pursues her topic. The system she uses should be multi-faceted enough to allow her to feed that knowledge back into her search strategies.

Suzanne is an analyst following human rights issues and provincial politics in the People's Republic of China. She has been working with human rights issues for the past ten years, but in Africa and South Asia, and was transferred to the PRC account three months ago. She knows English and some French and Portuguese, but no Chinese. On Monday morning at 7:30 A.M. she reviews her mail. Her mail queue has 500 items in it. She has organized this queue into six categories, which correspond with four profiles she currently has active, plus a tasking category and a miscellaneous category.

The tasking queue accepts e-mail only from her boss, his boss, and their deputies, plus several customers she supports directly in other government organizations. This morning the tasking queue has one new item in it from a customer, sent late Friday evening. He needs a five page outline summary of major events/speeches in the PRC in the past six months which bear on human rights issues; this must be in his hands by 1:00 P.M. Monday to provide background support for a meeting he will attend at 3:00 P.M.. He simply needs a summary of the Chinese position and the occasion, nothing about international response. Suzanne prints the request, underlines the key phrase 1:00 P.M. in red and tapes the paper to her wall over her computer monitor.

She sighs and skips over the miscellaneous category which has 50 messages which have been accumulating since last Wednesday when she last cleaned it out. None are labeled priority either by the sender or by her profile sorter which she has configured to flag items from several important colleagues and items of certain types from computer support, personnel, and security.

The Global Human Rights issues queue has open source and classified material on human rights outside the PRC and targets only extremely high scoring items on a relevancy scale. Entries are organized by major geographic region of the material covered, and then relevancy ranked. This morning there are 30 articles and cables in the queue. Suzanne skims the titles, authors and sources and on that basis saves 12 of these to her regional files. She skims the text of one article, reporting on a U.N. conference in Cairo and saves it to a special U.N. file.

The Chinese Internal Human Rights queue also carries open source and classified information. It is categorized by date of authorship of the article or message, and ranked under each date according to the relevancy score. There are 150 items. Ten are marked as highly similar or identical to items already in her personal archives. Twenty are duplicates of documents which also hit on her other profiles and are in her other queues. Fifteen more are duplicates or near duplicates of items in this queue and are clipped together with the similar documents. This leaves 100 items that she has not seen before. She has five that are marked as priority. She skims the text of all these. One of the priority items concerns an arrest of a film director which she forwards with comments directly to her manager. She skims the titles, sources, and authors of the rest of the items for anything that strikes her as probably irrelevant, opens the text to check her hunch and is able to discard 15 items in this fashion. Then she copies the entire remaining queue into a daily workspace. She has configured this workspace into four geographic regions with a timeline for each region. The timeline shows months for the past year, with previous months available by scrolling. Documents from her queue are automatically sorted by date and location of the events described in the documents, and attached to the correct month and geographic region. When she brings material from her archive or other searches, it will be automatically ordered in the same manner.

Suzanne's two other mail queues are Chinese External Human Rights and Chinese Provincial Politics. The External Human Rights queue is organized like the Internal Human Rights one; the Provincial Politics queue is organized by Province and by key person names. She reads them and disposes of items in a similar fashion, except that she copies the Provincial Politics items directly into her personal archives as she does not anticipate working on the material that day.

Then she goes to her morning staff meeting, where she tells her manager about her paper due that afternoon. She will have to get her material to him by noon for review. He tells her about two more projects he wants her to start. By the end of the week he needs a summary outline of changes in the Chinese official statements on Human Rights in the last five years, tied where possible to specific Chinese domestic and international events. In addition, he wants her to start tracking a new topic - Potential for Domestic Opposition to Current Chinese Leadership. She is to have some initial material ready in two weeks, when a determination will be made how to proceed further on this topic. This is a topic which is new to this work group as well as to Suzanne, and she has not been responsible before for material like this for other countries.

Suzanne starts with her paper due at noon to her manager. She searches her predecessor's archive, which includes open source, classified material, and his own papers. She uses as a detection need four articles from her mail queues, two from internal and two from external, one each classified and open source. She has her returns sorted by date of occurrence of the events described in the text and constrains the search to return only items concerning events in the past year. A quick scan of the titles, sources and authors of the top 25 documents reveals mostly relevant documents but shows none were authored by her predecessor; so she does another search for documents he authored within the past year on Human Rights issues. She gets lucky. He did a summary of the type she needs only a week before he left. A review of that document shows that it constitutes half of the material she needs already completed for a different customer.

She now searches his files for documents covering the two weeks between the last data in the paper and the time he left, using her first detection need. She runs the same detection need against her own internal and external archives, constraining both searches to documents dated from two weeks before she came on the job. She has the returns integrated and sorted by date of events described in the text, and then by date of authorship. She has her returns list include the length of the documents, as well as the titles, sources, and authors. From this returns list she concludes there were four major events which have occurred in the last three months, which agrees with her memory of the time since she arrived on the job.

Suzanne then picks the best item for each event based on her knowledge of the sources and the length of the item. She chooses reasonably long items since she wants comprehensive material on each event. Since she is providing only a paragraph on each event without much commentary or evaluation in this case, she is able to abstract the material quickly from the selected articles. She merges her material with her predecessor's material. After editing, she uses each section (concerned with summarizing a different event) as a detection need and runs it against her and her predecessor's archive with a date constraint of a year. She then reviews the text of the top three articles for each topic and the titles, sources, etc. of the top 30 articles for each topic looking for additional or contradictory information. When she is satisfied that her material is accurate, she passes it on to her supervisor.

The piece is passed to the customer at 12:45 and he immediately calls with questions. Suzanne has kept a working file of all the documents she looked at, as well as the top 30 documents from each of her searches. As she talks to her customer on the phone she is able to scan these lists for material to answer his questions. She forwards him a copy of an FBIS translation of a People's Daily editorial that bears on two of the events she summarized.

Next Suzanne looks for a paper by her predecessor similar to the one her supervisor needs by the end of the week. She constrains the search to those papers he authored in the past three years with China, human rights, and change or policy shift in the title or first or last paragraph. She finds a very lengthy paper he wrote three years before discussing Chinese human rights policy, without focusing specifically on policy shifts. She skims that, copying out the relevant passages. She realizes that she does not have the background she needs to determine quickly when major policy shifts occurred and create a timeline of these shifts.

Yet she needs to do this first, since once she has the major shifts identified she can summarize each one reasonably easily, using the same searching, abstracting and checking techniques she used for her events summary in the morning. She also believes that her predecessor's material may not be very complete in this area so she decides she will have to search corporate text data bases for the material.

After a little thought, Suzanne decides that FBIS, which follows the human rights issues pretty thoroughly, has probably not missed any major changes in policy in the last five years. In addition, she has found valuable a number of the reports coming out of other U.S. Agencies with a presence in the major East Asian capitals. So she targets one search to the FBIS archives, China coverage, containing the words "human rights", constraining the search to the last 6 years and to translations of editorials. She has the returned material categorized by source, and then date of publication, with the People's Daily on top, the rest of the sources listed alphabetically. When these are returned, she asks for copies in Chinese of the major People's Daily editorials and uses these as detection needs to run against a database of Chinese newspaper articles, again constraining the search to the past six years and to editorials. These are returned to her with translated titles and bibliographic information and with some keywords highlighted and translated; again she has the return list sorted by source and date. She looks for editorials that are different in date from her major People's Daily editorials and from any other sources. Several look interesting and she sends them to be machine translated. They will not return until after she has left work so she puts off their review until the next day, when she will determine if they signal any policy changes she has not yet uncovered.

Her second search she constrains to other U.S. Government reports from several capitals, specifies the date range, and requires the words "human rights". She has the returns sorted by date. She scans material that does not coincide with the dates from the People's Daily editorials, but can find no indications of any policy shift not already uncovered by her FBIS search. At this point she feels confident she has the major policy shifts identified and has a good start on the material she needs, and she can proceed over the next several days using a similar strategy to the paper she just completed that morning.

However, Suzanne is very worried about her new topic. She has never done a report on potential for opposition before and she is not sure where to begin. So at 3:30 she puts aside her unfinished work on the Chinese Human Rights Policy and tries to think about what to do. She posts an electronic note on the analyst's bulletin board she used at her previous job working on African countries. Then she thinks she remembers a paper from several years before concerning the potential for opposition in Nigeria which might have a general outline and methodology that she could learn from. She accesses the North Africa work group's server and uses their data base of documents. She puts in a query using the words Nigeria and opposition; she constrains the search to five years, to material over 10 pages in length, and to material produced by her agency. She finds the document, but it is less helpful than she hoped it might be.

Next Suzanne accesses her own work group's document data base. She tries a search using the phrase "opposition to the government in China" with a constraint that China is required. This draws a lot of hits but all have to do with Chinese external affairs, with student and popular unrest, and with discussions of long past revolutionary movements. She is supposed to be looking for information on the possible formation of true opposition groups with power within the existing political system. She tries writing a paragraph describing in some detail what she wants and uses this as her detection need. None of the material returned has to do with the PRC; however, there are several newspaper and magazine articles ranked high on the list which concern opposition parties in South Korea, Japan, Malaysia, and Singapore. Since she is looking among other things for general ideas, background and methodology which would help her structure her research and her report, she decides to pursue this lead. She uses the best of these articles as a detection need and runs it against a multilingual database of East Asian and Southeast Asian material. The returns have translated titles and bibliographic citations. Among these, she finds a major article from a Japanese magazine which appears, from the highlighted and translated hit terms, to concern the development of opposition parties in Vietnam. She sends this out for translation, with the hope that it will provide her with some methodology or framework on which to proceed with her analysis.

She knows the Japanese article will not return until morning, but she has a little more time before she has to go home. She decides to try something she hasn't done before. She knows her retrieval system uses a query expansion capability and that she can access the system query and modify it. She expands her "opposition to the government in China" detection need to be "opposition to the government in China, but excluding anything about foreign affairs or student unrest". Then she opens up the resulting system query and looks at the expansions for "opposition to the government". These include many words concerning revolutionary and protest activities. She eliminates the more extreme words. Then she returns to her detection need and adds some phrases she hopes will pick up the more subtle kinds of activities she believes might be possible: "meetings between provincial leaders", "caucuses at the national legislature", "disagreements", "resolution of differences", "political faction", "struggle for leadership", and so forth. She then weights returns from non-Chinese sources higher than those from China, believing that the type of event she is interested in is most likely to be discussed in the International and Far Eastern English press, in the Japanese press, or by U.S. Government reports. She constrains the search to the last three years and runs it against the entire corporate on-line data base. This strategy proves modestly successful. Before she leaves for the day, she has sitting in her workspace several items discussing issues on which the Chinese central government seemed to have encountered some concerted, low-level domestic opposition to its policies. These items she can follow-up on the next day by using her multi-lingual database browser to search for details in the native language holdings for her work group.

Suzanne picks up her briefcase, and then shakes her head in amazement. She has forgotten to eat her lunch again, and her homework for Chinese class tomorrow morning at 6:30 is still undone. And it's only Monday.

Advanced Analytical Scenario

Analyst not too long after the Year 2000

Note: the tasks described here are fictitious; however, they describe a method of working which would be applicable to a number of real tasks faced by Government analysts.

Ada is a technology analyst, with a background in materials and mechanical engineering. Her only language is English. She is currently tasked to track new developments in materials, design, manufacturing techniques, and performance of a complete list of non-motorized sports equipment. She writes a weekly report on current trends in technologies. Each week she focuses on several different kinds of sporting equipment and protective gear, for example, tennis rackets, kayaks, bicycles, skis, in-line skates, and helmets, boots, safety straps and so forth; additionally, whenever there is a major shift in technology trends for any of these product categories, she publishes a special report. Ada covers her technology for only Africa and South Asia. She is part of two work groups, one in which she collaborates with other analysts following all sporting equipment technology throughout the other geographic areas of the world, the other in which she collaborates with analysts following related materials and manufacturing technologies in Africa and South Asia. Members of these work groups review each other's work, jointly produce reports several times a year, and cover for each other on short term tasking. She communicates and collaborates with both work groups via voice and computer links, since not all the analysts' workspaces are contiguous. All of these analysts interface with their computers using a combination of voice, keyboard, and touch, depending on the task. They have common workspaces where they build their collaborative reports and have the ability to share with or prohibit their own work spaces to whomever they wish among their colleagues.

Ada and her colleagues receive most of their information from a variety of text publications, some received digitally and a few in hard copy. They have full access to several commercial on-line services which provide data bases and abstracts or full text of newspaper and journal

articles in a wide range of technical and non-technical domains. The office downloads, indexes, and distributes the full collections from these services. Additional electronic bulletins and newsletters arrive periodically from commercial industry sources. These are handled by the same indexing and distribution system which deals with newspaper and journal materials. Advertising is a particularly useful source of information to Ada and the scientific literature she reads contains a large number of diagrams and chemical and mathematical formulae. The indexing system the office uses allows Ada to view all graphics and photographs on the page as they were presented in the original publication. She can retrieve information based on queries which include mathematical expressions and specifications of visual content, such as color or shapes. A number of the analysts correspond with experts in their fields around the world via an international e-mail system. At their discretion, useful material is also indexed along with published materials.

Hard copy materials, most of them in foreign languages, arrive from a number of different sources. The copy quality of these is highly variable. The office has a scanner and OCR system which can read most foreign scripts. Thus most materials, including graphics, advertising and format information, can be captured in digital form allowing this material to be indexed and distributed in the same manner as material received digitally. There is some degradation for very poor quality originals, such as material printed on extremely cheap paper or poor reproductions. The system identifies and routes those which the OCR cannot clean up adequately to a system administrator, who can check the document's importance and determine how it should be stored. Some important documents are manually corrected and fed into the system as digital text documents. Others are stored as images and retrieved through a somewhat less powerful search capability than that used on the digital text collection. Every query sent against the digital text collection also has a version automatically sent against the document image collection, and imaged documents are available to the analysts through the same interface they use for the rest of their work.

Many of the sources of information for these analysts come in languages other than English. Everything is stored in its original form. All queries from analysts, although developed in English in most cases, are transformed by the routing and retrieval system to retrieve documents in all languages requested by the analyst. While Ada knows only English, several of her colleagues are fluent in relevant languages and will edit the queries the system develops for those languages. Documents in languages other than English are returned to Ada with English citations (titles, authors, source, date) and with English summaries which focus on material related to her many sports technologies. She can retrieve documents with key portions run through a quick-and-dirty machine translation system. She can then route documents of interest to her either to a more fully capable machine translation system or, in cases where nuances are of importance to her, she can consult with a linguist about the meaning of certain passages. In rare instances, she has documents translated in full by a linguist.

Ada has her system configured to do a considerable amount of processing at night while she is not there. When she arrives in the morning, all material routed to her during the previous twenty-four hours has been processed together. In fact the only material she looks at during the day, as it comes in, is high priority material. On her computer desktop at 8:00 A.M. is a summary of all material from the previous day. This is in bulletized format in an outline that she has designed. Each bullet summarizes what the material has said about developments in the different aspects of sporting equipment technology she is following. Differences between what this new material has said and the key elements of material in her personal archive and in the work group archives are highlighted for each bullet. For each bullet she is able to access immediately the number, names, dates and origins of new and old sources which support or contradict this bullet. She can access any of these source documents directly from the summary material. The system is sometimes inaccurate in the way it has interpreted material, but Ada is able to quickly remove inappropriate documents from the queues or categories in which they have been placed and put them in the right categories. The system immediately queries her about the reasons for her moving the document using a simple error report which she fills out. From these reports the system gradually learns to respond to her information needs more accurately. Reprocessing to adjust indexes to this increased level of accuracy is accomplished at night. Ada can display numerical

data about the summary material, such as dates or numbers of sources, and geographic data, such as origins of material, on appropriate time lines and maps, so that she can easily see, without herself having to do any further analysis, the numbers of sources supporting a summary bullet by month or by week, for as much as the past five years.

One of Ada's work groups maintains a formatted data base of information on scientists around the world working in their materials and manufacturing disciplines. This data base is fed automatically by a system which receives documents directly from the document routing system. The data base is in English, although many of the documents feeding it are in other languages. Additional multi-media information, from radio and TV news broadcasts and video clips, is available through this data base, automatically linked to the scientist or facility where he or she works. Data which comes in concerning scientists is usually in free text and must be extracted automatically into the data base. Since the work group likes to keep this data base as a corporate resource of reasonably well verified and deconflicted material, they have a system for regularly reviewing material. The extraction system itself does the first level of review, checking data base records to see if material already exists about the subject being input. In the case of contradictions of new material with the data base, the system checks the original source material to determine whether the new or old source is more reliable. Some expert rules have been built in to resolve reliability issues. However, in some cases, new and old records, with their corresponding texts are clipped together and flagged as unresolved. Additionally, the extraction system flags material which it is unsure of in any record. All material flagged by the system is reviewed by a junior level analyst who is trained to resolve certain issues and to route the rest to more senior analysts for resolution, according to their areas of specialization and/or languages. The junior analyst works part time on this task, and the senior analysts spend less than an hour per week reviewing questioned records. All changes are documented to the system which learns from these changes. In addition, the system automatically does checks on the consistency of word, phrase, discourse, and grammar use to check for drifts in language use or other anomalies which suggest the need for further checking of records. The data administrator carries out various random checks of set numbers of records several times a year to ensure that the accuracy level of the data is within the target limits.

Ada also has at her disposal a quick reaction extraction system. This sits on her computer desktop and is closely coupled with a data base management tool. Through the interface she can build a data base record and rules to fill the fields, from text examples. She uses this system herself on English text, but must get a linguist to help her should she want to use it on non-English text. With this tool she can build a new data base schema with fill capability in about two hours. She uses this tool at least once a month when special requests come in that are not covered by the topic/bullets of her normal work routine or when she needs to understand a import/export trend or a pattern of technology transfer in greater depth than normally.

Both of Ada's work groups have clustering systems which run constantly off-line looking for new combinations of words and concepts to bring to the attention of the analysts. The analysts in her work groups participate in a rotating "watch" for new developments, using the clustering systems' input and their fast browse capability as their starting points. For at least four hours out of every day, there is an analyst on "watch" for each work group. Ada's share comes to two hours one day per week. During her watch, Ada begins by reviewing the material the system has designated as possible new clusters and unexplained or unusual material. Document clusters are displayed graphically on the screen in a schema that she has configured for her own particular visual work style. She can indicate clusters of interest, have them reclustered, indicate documents of interest and jump from document to document on the basis of a number of different document characteristics. "Watch" work is not done at the analyst's desk but at a special terminal. At this terminal, Ada can view, group, browse and annotate documents; she can also route documents to herself or colleagues who need to see them. However, she does not use this terminal for her other regular analytical duties.