DOCUMENT DETECTION SUMMARY OF RESULTS

Donna Harman

National Institute of Standards and Technology
Gaithersburg, MD. 20899

1. INTRODUCTION

This section presents a summary of the TIPSTER results, including some comparative system performance and some conclusions about the success of the detection half of the TIPSTER phase I project. For more details on the individual experiments, please see the system overviews.

Four contractors were involved in the document detection half of TIPSTER. Two of the contractors worked in English only (Syracuse University and HNC Inc.), one contractor worked in Japanese only (TRW Systems Development Division), and one contractor worked in both languages (University of Massachusetts at Amherst). The four contractors had extremely varied approaches to the detection task. TRW transformed an operational English retrieval system (based on pattern matching using a fast hardware approach), into a Japanese version of the same operation, with a special interface designed to facilitate work in Japanese. The University of Massachusetts approach involved taking a relatively small experimental system using a probabilistic inference net methodology, scaling it up to handle the very large amounts of text and long topics in TIPSTER, and modifying the algorithms to handle Japanese. Both Syracuse University and HNC Inc. built completely new systems to handle the English collection. In the case of Syracuse University, their system is based heavily on a natural language approach to retrieval, with many of the techniques traditionally used in document understanding applied to the retrieval task. HNC Inc. took a totally different approach, applying statistical techniques based on robust mathematical models (including the use of neural networks).

There were three evaluations of the contractors' work; one at 12 months, one at 18 months, and the final one at 24 months. In each case, the contractors working in English have made multiple experimental runs using the test collection, and turned in the top list of documents found. These results were first used to create the sample pool for assessment, and then were scored against the correct answers based on results from all runs (including TREC-1 runs for the 18-month evaluation and TREC-2 runs for the 24-month evaluation). Standard tables using recall/precision and recall/fallout measures were distributed and compared. The evaluation of the Japanese work took place only at the 24-month period.

2. 12-MONTH EVALUATION

The work done for the 12-month evaluation was mainly a scaling effort. Not all data was available so only partial results were completed. In particular, the University of Massachusetts turned in 7 runs using the adhoc topics, with experiments trying different parts of the topic to automatically create the query, and also adding phrases. Additionally they tried some manually edited queries. HNC Inc. turned in 4 runs using the adhoc topics, with experiments also using different parts of the topic to automatically generate queries. Additionally they tried various types of "bootstrapping" methodologies to generate context vectors. Syracuse University turned in no runs, but had completed the extensive design work as proposed in their timeline. The University of Massachusetts also did 4 runs on the routing topics, but the lack of good training data made this very difficult. In general the results for the systems was good, with the University of Massachusetts outperforming HNC Inc. on the adhoc runs, but it was felt by all that this evaluation represented a very "baseline" effort. For these reasons, no graphs of these results will be presented.

3. 18-MONTH EVALUATION

By the 18-month mark, the systems had finished much more extensive sets of experiments. The University of Massachusetts continued to investigate the effects of using different parts of the topic for the adhoc runs, but this time trying different combinations using the inference net methodology. Figure 1 shows three INQRY runs for the adhoc topics done for the 18-month evaluation. The plot for INQRYV represents results from queries created automatically using most of the fields of the topics. The INQRYJ results are from the same queries, but including phrases and concept operators. The INQRYQ results show the effects of manually editing the INQRYJ queries, with those modifications restricted to eliminating words and phrases, adding additional words and phrases from the narrative field, and inserting paragraph-level operators around words and phrases. As can be seen, the use of phrases in addition to single terms helped somewhat, but the results from manual modification of the queries were the best adhoc runs.

Figure 1 also shows the two HNC adhoc results for the 18-month evaluation period. These runs represent final results from many sets of bootstrapping runs in which HNC evolved techniques for completely automatically creating context vectors for the documents. The plot marked HNC2aV represents the results using these context vectors for retrieval and using automatically built queries from the concepts field of the topic. The HNC2aM results adds a rough Boolean filter to the process, requiring that three terms in the concepts field match terms in the documents before the context vectors are used for ranking. This Boolean filter provided considerable improvements.

Figure 2 shows the routing results for the 18-month evaluation. The three plots for the INQRY system represent a baseline method (INQRYA, same as INQRYJ adhoc approach), and two modifications to this method. The INQRYP results show the effect of manually modifying the query, similar to the method used in producing the INQRYQ adhoc results. The plot for INQRYR shows the first results from probabilistic query expansion and reweighting performed automatically using relevance feedback techniques on the training data. Both methods improve results, with the automatic feedback method results approaching the manual-modification method, especially at the high recall area of the curve.

The HNC routing results shown on figure 2 represent the use of two different types of neural networks. The plot marked HNCrt1 is the baseline result, created by using the adhoc methods similar to those used in run HNC2aV. The HNCrt2 results represent using neural network techniques to learn improved stem weights for the context vectors based on the training data. The HNCrt3 results come from using the training data to determine what type of routing query to use, i.e. an automated adhoc query (similar to HNC2aM), a manual query, or a query using the neural network techniques (HNCrt2). Clearly the neural network learning techniques significantly improve performance, with the per topic "customization" performance (HNCrt3) working best.

In terms of system comparison, the University of Massachusetts runs were consistently better than the HNC runs for the adhoc topics, whereas for the routing topics both groups were similar. The results were a major improvement over their baseline (12-month) results for both groups.

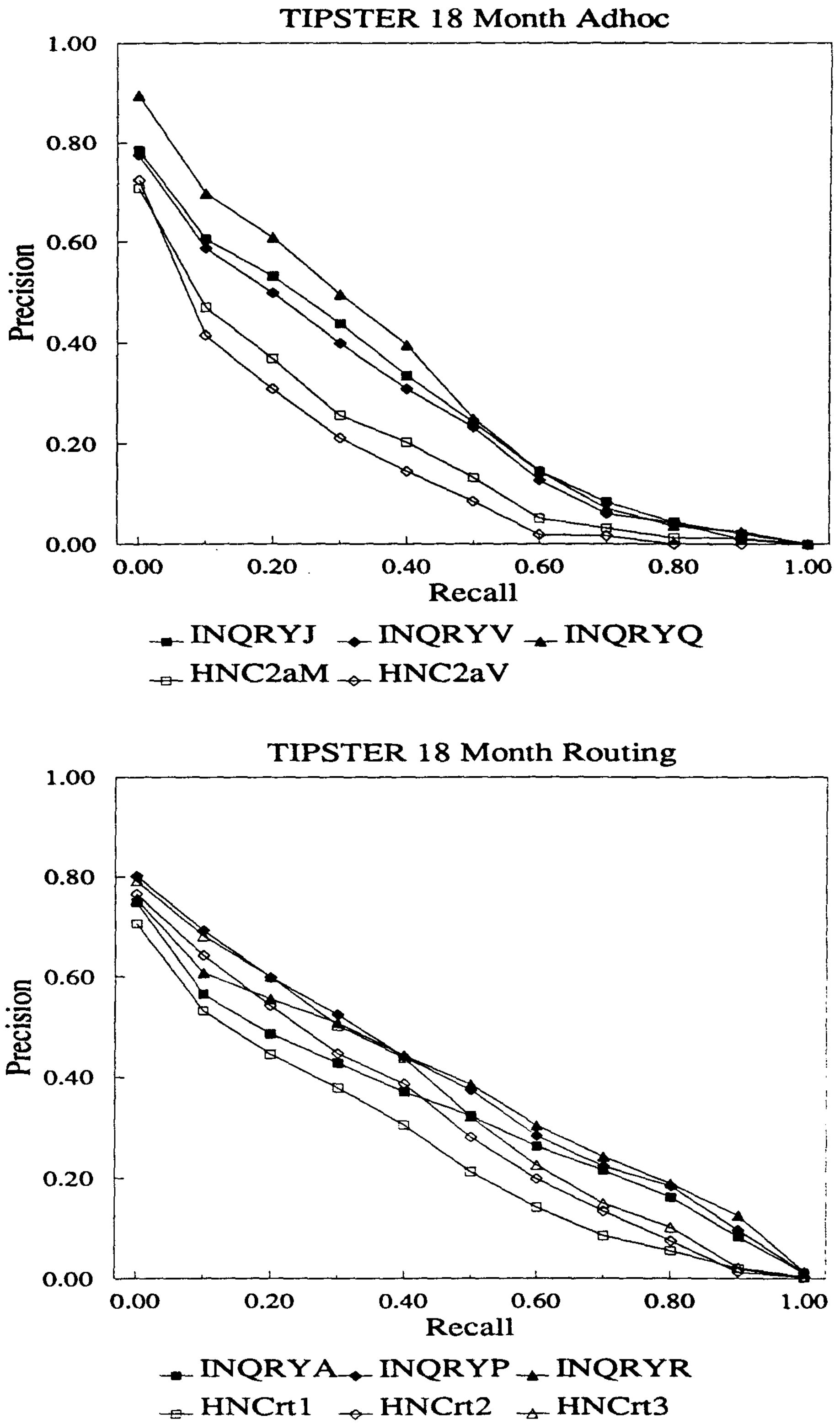
At the 18-month evaluation period, Syracuse University had the first stage of their system in operation and turned in results for the first time. The results for adhoc and routing are shown in figures 3 and 4. Since the results are for only a subset of the data used by the other contractors, they cannot be directly compared. Additionally since the results are only for the first stage of retrieval, which emphasizes high recall, they should not be viewed as the final results from the system.

Figure 3 shows four Syracuse runs on the adhoc topics. The documents used are the subset of the collection having the Wall Street Journal only. The first three plots, DRsfc1, DRpna1, and DRtsa1, represent three operations in the DR-LINK system. The first operation does a rough filtering operation on the data, only retrieving documents with suitable subject field codes. The next two operations locate proper nouns and look at document structure. There is a considerable improvement in performance between the first two operations. The fourth run (DRfull) used a manual version of the second stage to produce final results. These results are for only half the topics, so cannot be strictly compared to the first three runs, but they do indicate the potential improvements to precision that can be expected from the second stage.

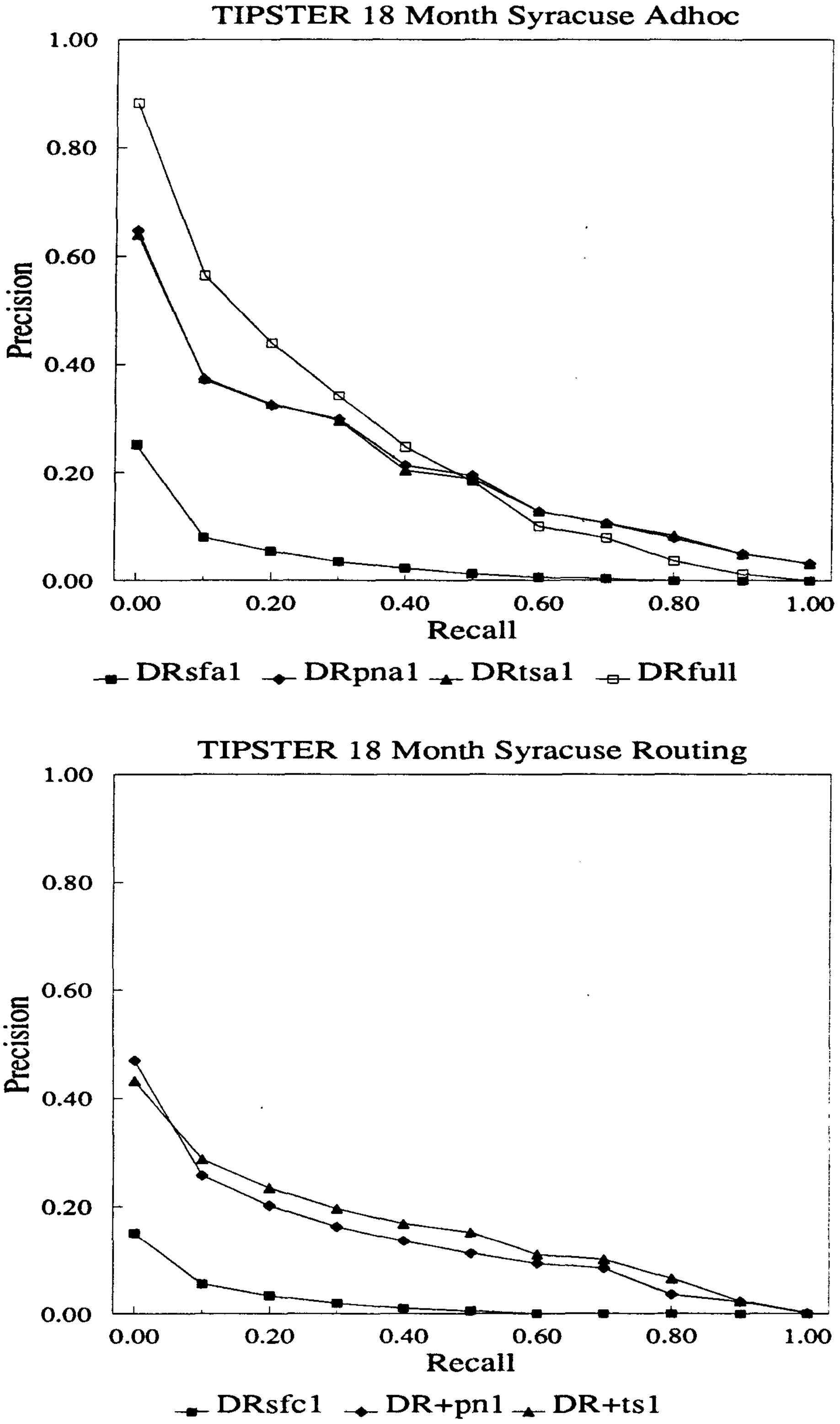
Figure 4 shows the same operations, and generally the same improvements, for the routing topics. In this case the subset of documents used was the AP newswire documents. The same three operations discussed above are plotted here. There was no second stage trial for the routing topics. These two graphs represent the baseline of the Syracuse system.

4. 24-MONTH EVALUATION

For the 24-month evaluation, all groups turned in many runs. The runs were much more elaborate, with many different types of parameters being tried. The University of Massachusetts tried 7 experiments with the adhoc topics, using complex methods of combining the topic fields, proximity restrictions, noun phrases, and paragraph operators. Additionally an automatically-built thesaurus was tried. They also did 15 runs with the routing topics, trying various experiments combining relevance feedback, query combinations, proximity operators and special phrase additions. HNC Inc. did 4 adhoc runs using various types of learned context vectors. Additionally they tried a simulated feedback query construction run. For routing they did 5 runs, trying multiple experiments in different combinations of adhoc and neural net approaches. Syracuse University turned in 10 runs for their



Figures 1 and 2: Adhoc and routing performance at the 18-month evaluation period (using full collection)


TIPSTER 18 Month Syracuse Adhoc

TIPSTER 18 Month Syracuse Routing

Figures 3 and 4: Adhoc and routing performance at the 18-month evaluation period (using WJS or AP only) "upstream processing module" (3 adhoc and 7 routing), trying various types of ranking formulas. Additionally they did 13 runs using the full retrieval system (4 adhoc and 9 routing). Full descriptions of these runs are given in the system overviews.

Figures 5 through 12 show the results from the 24-month evaluation. Figures 5 and 6 show some of the adhoc results for the full collection, and figures 7 and 8 show some of the routing results. The results from Syracuse University on a smaller subset of the document collection are shown in figures 9 through 12.

Figure 5 shows the recall/precision curves for the adhoc topics. The three INQRY runs include their baseline method (INQ009), which is same as the baseline method INQRYJ developed at the 18-month evaluation period. The first modification (INQ012) uses the inference net to "combine" weights from the documents and weights from the best-matching paragraphs in the document. The second modification (INQ015) shows the new term expansion method using an automatically-built thesaurus. Both modifications show some improvements over the baseline method.

The three HNC runs shown on figure 5 include a baseline (HNCad1) that is similar to their best 18-month adhoc approach (HNC2aM), but that uses a required match of 4 terms rather than 3. The HNCad3 results show the effects of using a larger context vector of 512 terms rather than only 280 terms for the baseline results. This causes a slight improvement. The HNCad2 results are using some manual relevance feedback.

The University of Massachusetts results are better than the HNC results, but there were improvements in both systems over the 18-month evaluation. Figure 6 shows the recall/fallout curves for the best runs of these two systems. Both plots show the same differences in performance, but it can be seen on the recall/fallout curve that both systems are retrieving at a very high accuracy. At a recall of about 60 percent (i.e. about 60 percent of the relevant documents have been retrieved) the precision of the INQRY results is about 30 percent. The fallout, however, is about 0.0004, meaning that most non-relevant documents are being properly screened out. This corresponds to a probability of false alarm rate of 0.0004 at this point, in ROC terminology.

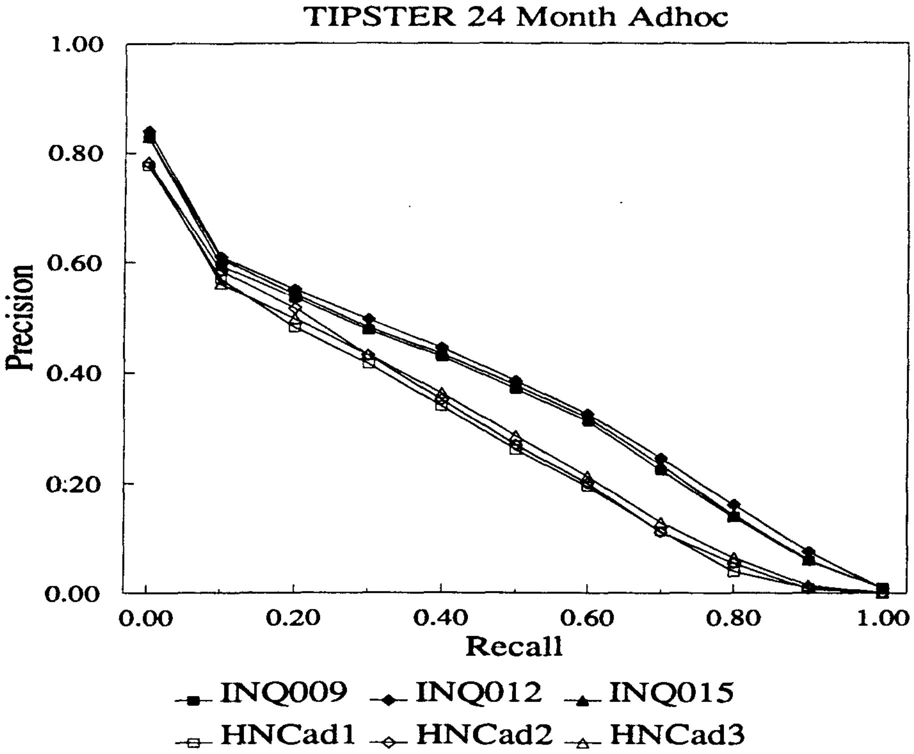
Figure 7 shows the routing results for both groups. The run marked INQ026 is the baseline run of the INQRY system and uses the same methodology as the adhoc INQ009 run. The other two runs add some type of relevance feedback using the training documents. The plot marked INQ023 uses both relevance feedback and proximity operators to add 30 terms and 30 pairs of terms from the relevant documents to the query. The most complex run, INQ030, constructed the queries similarly to run INQ023, but additionally weighted the documents using a combination method similar to adhoc run INQ012. These runs represented the best results from many different experiments, and the relevance feedback gives significant improvement over the baseline runs.

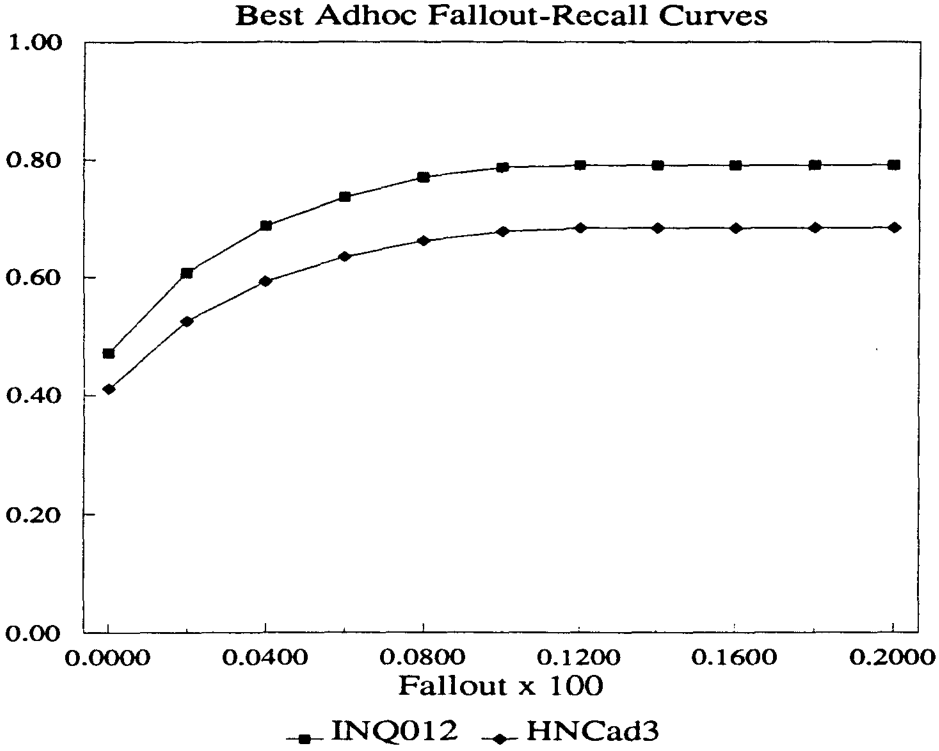
The HNC routing results also represent the best of many experiments. The results for HNCrt5 show the neural network learning using stem weighting, similar to HNCrt2 at the 18-month evaluation. The second two sets of results represent data fusion techiques, with HNCrt1 being fusion of four types of retrievals, using the same combinations for all topics, and HNCrt2 using different combinations for different topics. The data fusion combinations both work well, but the per topic combination works the best, just as the less sophisticated version of this run worked best at the 18-month evaluation.

Again the University of Massachusetts results were better than the HNC results, but with major improvements in both systems over the 18-month evaluation. Figure 8 shows the recall/fallout curves for the best runs of both groups.

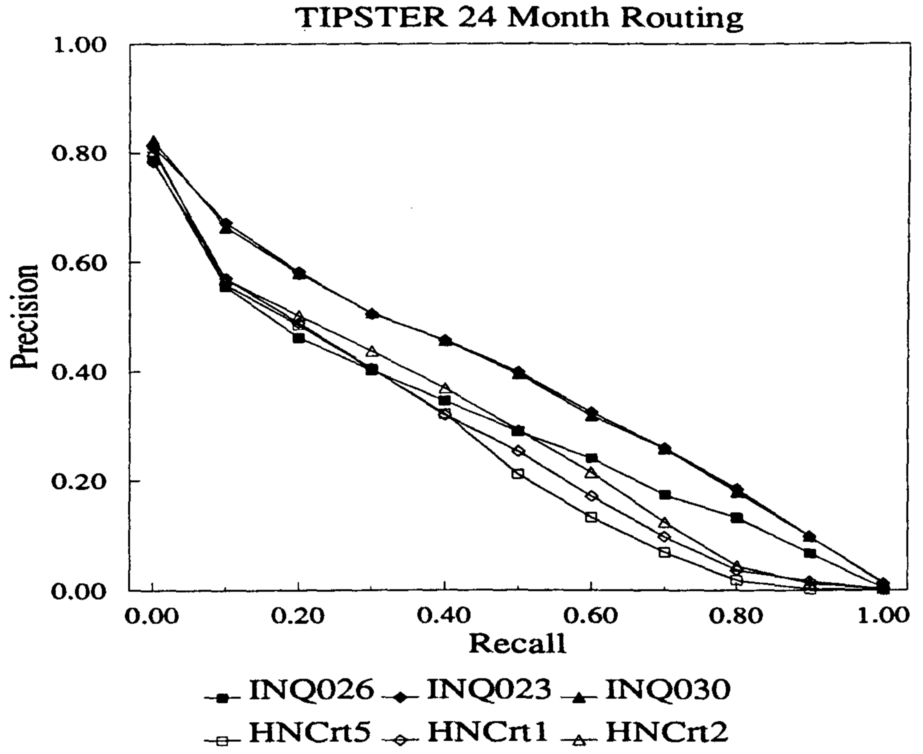
The Syracuse runs were on a subset of the full collection so are not directly comparable. However they also showed significant improvements over their 18-month baseline. Figure 9 shows three first-stage Syracuse runs, the results of trying different complex methods of combining the information (subject field code, text structure, and other information) that is detected in the first-stage modules. The results of this combination are passed to the second stage (figure 10). Note that due to processing errors there were documents lost between stages, and these official results are therefore inaccurate. Additionally only 19 topics (out of 50) are shown in figure 10. The improvements that could have been expected do not show because of these problems.

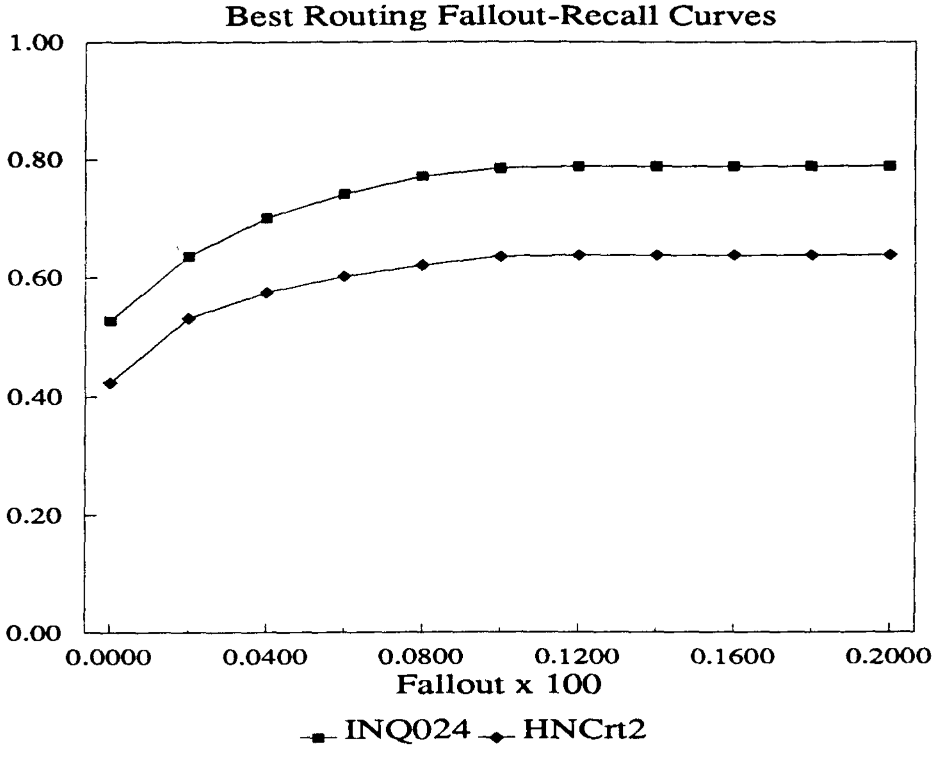
Figures 11 and 12 show the Syracuse routing runs. The first stage runs show not only the combinations from the adhoc, but also additional ways of integrating the data. Again there were processing errors with the second stage results, and therefore no improvement is shown using the second stage.



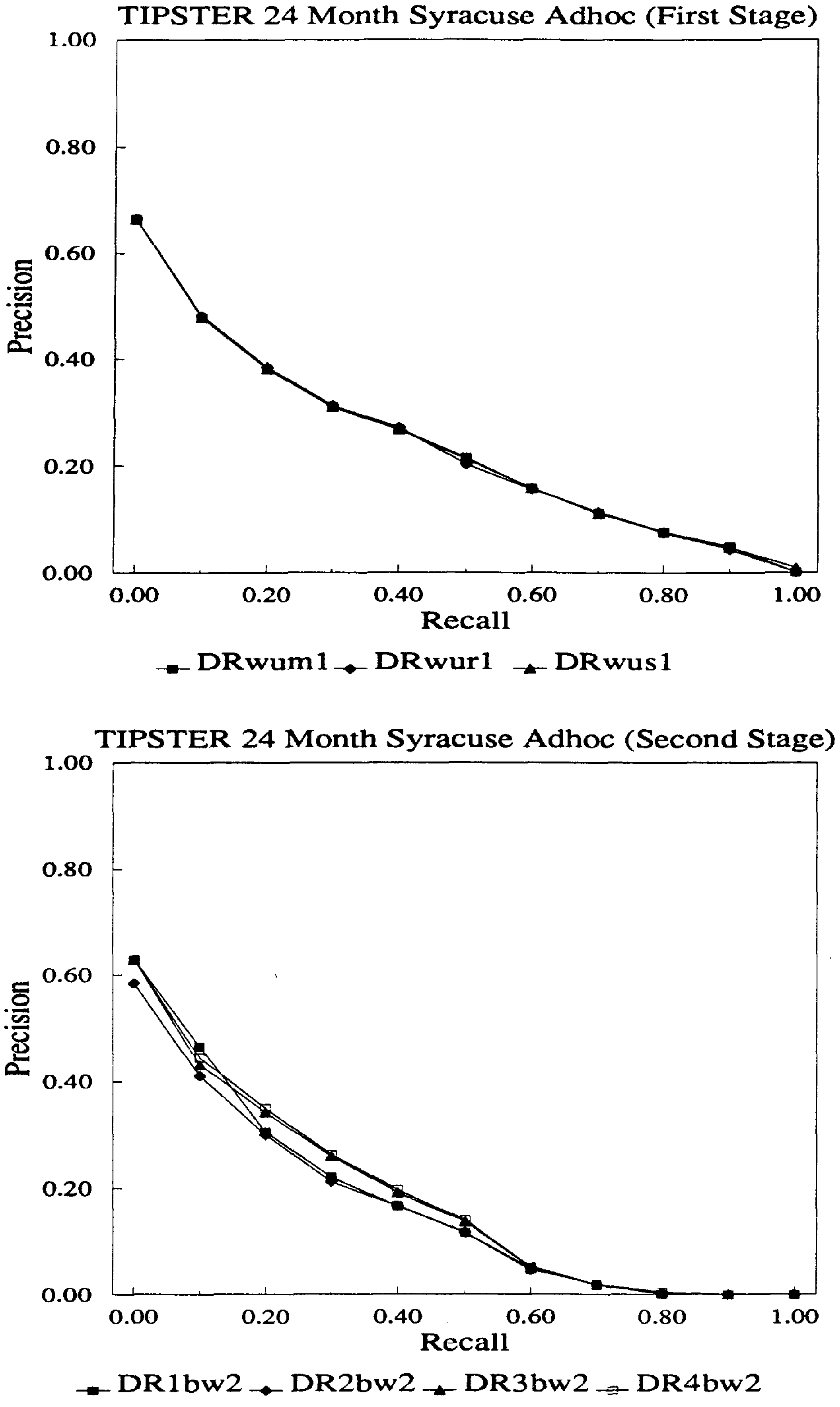


Figures 5 and 6: Adhoc performance at the 24-month evaluation period (using full collection)

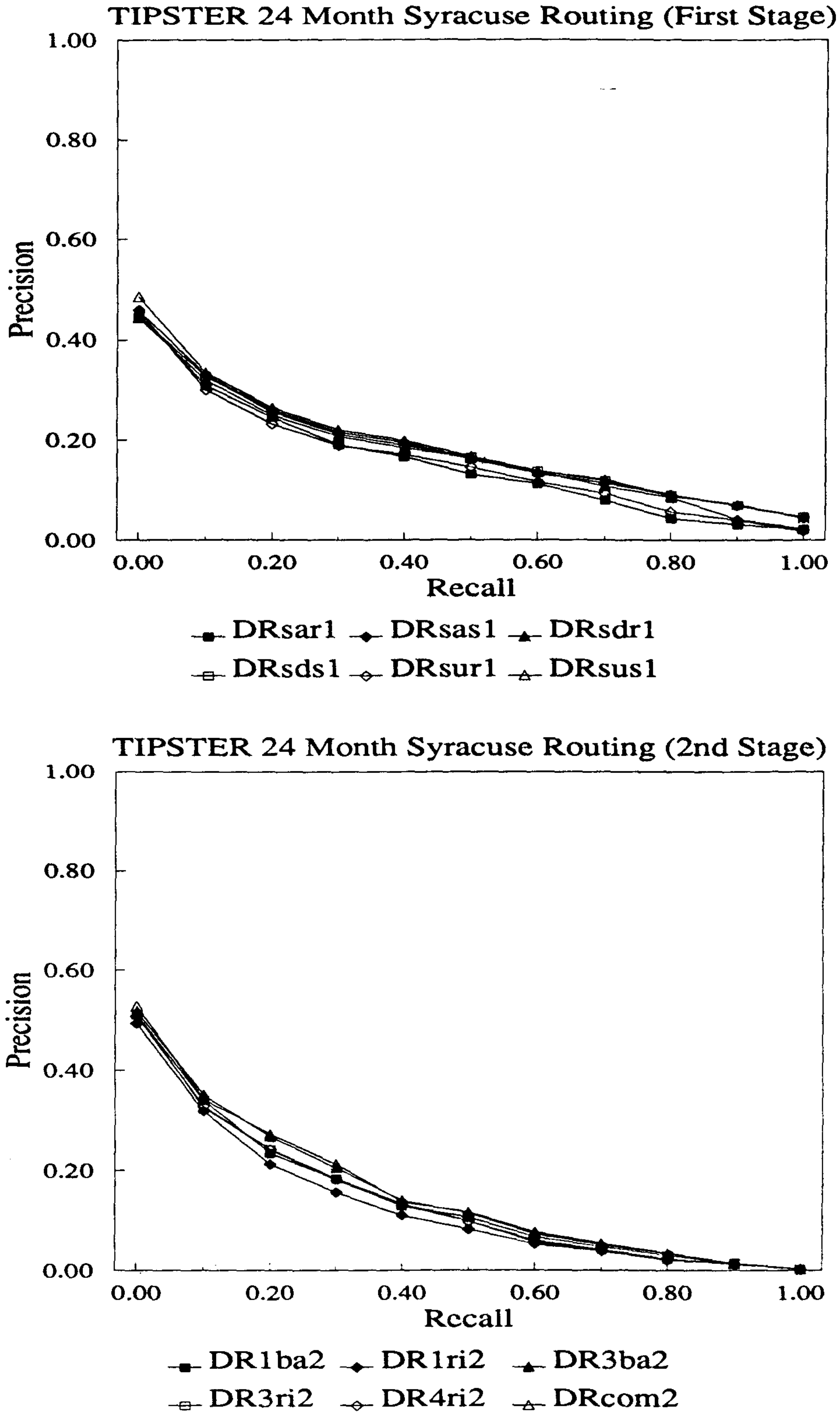




Figures 7 and 8: Routing performance at the 24-month evaluation period (using full collection)



Figures 9 and 10: Adhoc performance at the 24-month evaluation period (using WJS only)



Figures 11 and 12: Routing performance at the 24-month cvaluation period (using SJMN only)

5. COMPARISON WITH TREC RESULTS

How do the TIPSTER results compare with the TREC-2 results? Two of the TIPSTER contractors submitted results for TREC-2 and these can be seen in Figures 13 and 14. These figures show the best TREC-2 adhoc and routing results for the full collection. More information about the various TREC-2 runs can be found in the TREC-2 proceedings [1]. The results marked "INQ001" are the TIPSTER INQUERY system, using methods similar to their baseline TIPSTER INQ009 run. The "dortQ2", "Brkly3" and "crnlL2" are all based on the use of the Cornell SMART system, but with important variations. The "crnlL2" run is the basic SMART system, but using less than optimal term weightings (by mistake). The "dortQ2" results come from using the training data to find parameter weights for various query factors, whereas the "Brkly3" results come from performing statistical regression analysis to learn term weighting. The "CLARTA" system adds noun phrases found in an automatically-constructed thesaurus to improve the query terms taken from the topic. The plot marked "HNCad1" is the baseline adhoc run for the TIPSTER 24-month evaluation. The TIPSTER INQUERY system is one of the best performing systems for the TREC-2 adhoc topics.

The routing results from TREC-2 (shown in figure 14) exhibit more differences between the systems. Again three systems are based on the Cornell SMART system. The plot marked "crnlC1" is the actual SMART system, using the basic Rocchio relevance feedback algorithms, and adding many terms (up to 500) from the relevant training documents to the terms in the topic. The "dortP1" results come from using a probabilistically-based relevance feedback instead of the vector-space algorithm. These two systems have the best routing results. The "Brkly5" system uses statistical regression on the relevant training documents to learn new term weights. The "cityr2" results are based on a traditional probabilistic reweighting from the relevant documents, adding only a small number of new terms (10-25) to the topic. The "INQ003" results also use probabilistic reweighting and add 30 new terms to the topics. The "hnc2c" results are similar to the HNCrt1 fusion results for the 24-month TIPSTER evaluation.

These plots mask important information as they are averages over the 50 adhoc or routing topics. Whereas often the averages show little difference between systems, these systems are performing quite differently when viewed on a topic by topic basis. Table 1 shows the "top 8" TREC-2 systems for each adhoc topic. The various system tags illustrate that a wide variety of systems do well on these topics, and that often a system that does not do well on average may perform best for a given topic. This is an inherent performance characteristic of information retrieval systems, and emphasizes the importance of getting beyond the averages in doing evaluation. Clearly systems that perform well on average reflect better overall methodologies, but often much can be learned by analyzing why a given system performs well or poorly on a given topic. This is where more work is needed with respect to analyzing the TIPSTER and TREC results.

Tables 2 and 3 show some preliminary analysis of two of the topics with respect to the TIPSTER contractors. Table 2 gives the ranks of the relevant documents retrieved either by the HNCrt1 run or the INQ023 run. Clearly the HNC run is better for this topic, providing much higher ranks for most of the relevant documents. Note that five of the relevant documents were not retrieved by either system.

Table 3 shows a slightly different view of the same phenomena, but for topic 121. There were a total of 55 relevant documents for this topic, with only 13 of them found by the TIPSTER systems. Table 3 lists those 13 documents, the rank at which they were retrieved, and the "tag" of the system retrieving them. Note that for this topic the INQUERY system is performing better than the HNC system. These tables illustrate the varying performance of different methods across the topics. A major challenge facing each group is to determine which strategies are successful for most topics, and which strategies are successful only for some topics (including how to identify in advance this topic subset).

6. JAPANESE RESULTS

By the 24-month evaluation, only 7 topics were ready for testing. Both TRW and the University of Massachusetts ran these topics successfully, and the results are discussed in their system evaluations. No comparison of the results is possible between the two systems because of the preliminary nature of having only 7 topics. However, the University of Massachusetts (who did both English and Japanese) reported that minimal effort was necessary for porting their English techniques to Japanese, especially given the availability of the JUMAN Japanese word segmentor. Additionally the new TRW Japanese interface was judged a major success by the beta site tests.

7. CONCLUSIONS

What are some of the conclusions that can be drawn from the many experiments performed in the TIPSTER and TREC evaluations, and equally important, what is the lasting value of this two-year project?

First, the statistical techniques (using non-Boolean methods without any formal query language) that were used on the smaller test collections DO scale up. The simplest example of this is the consistently high performance of the Cornell SMART system in TREC. This very basic system relies on the vector-space model and on carefully crafted term weighting to produce their high results. A more complex example of the successful use of statistical techniques is the University of Massachusetts INQUERY system, which uses the more sophisticated inference network approach to achieve their high performance. This system has been very successful throughout the TIPSTER project, and has achieved this success using variations on their original system rather than having to completely revise their techniques.

Second, the results obtained by the best systems in TIPSTER and TREC are at a level of performance that is generally accepted to be superior to the best current Boolean retrieval system. More importantly, this performance is achieved from simple natural language input, allowing consistently superior retrieval performance without exhaustive training or experience. These systems are clearly ready to be tested in fully operational environments.

Third, the use of a large test collection has shown some unexpected results. Techniques that should have brought improvements have not done so. The use of phrases instead of single terms has not resulted in significant improvements; the use of proximity or paragraph-level retrieval has not shown especially good results; and the use of more complex NLP techniques have not worked well yet. Conversely, techniques that have not been successful before such as using types of automatic thesaurii for topic expansion have had unexpected success. These unexpected results using a large test collection are reopening research on old discarded ideas and starting research in new areas. It is much too early to draw firm conclusions on any of these techniques. Often poor performance that is attributed to one problem may be the result of lack of balance in parameter adjustment, e.g., the lack of improvement from phrases may be caused by the difficulty in balancing the weights of these phrases and the weights of single terms.

What is the lasting value of the document detection half of the TIPSTER phase I project? The first contribution in my opinion has been the development of a large test collection and the wide acceptance of its use via the TREC conferences. The lack of a large test collection has been a major barrier in the field of information retrieval and its removal allows an expansion of research by many groups world-wide.

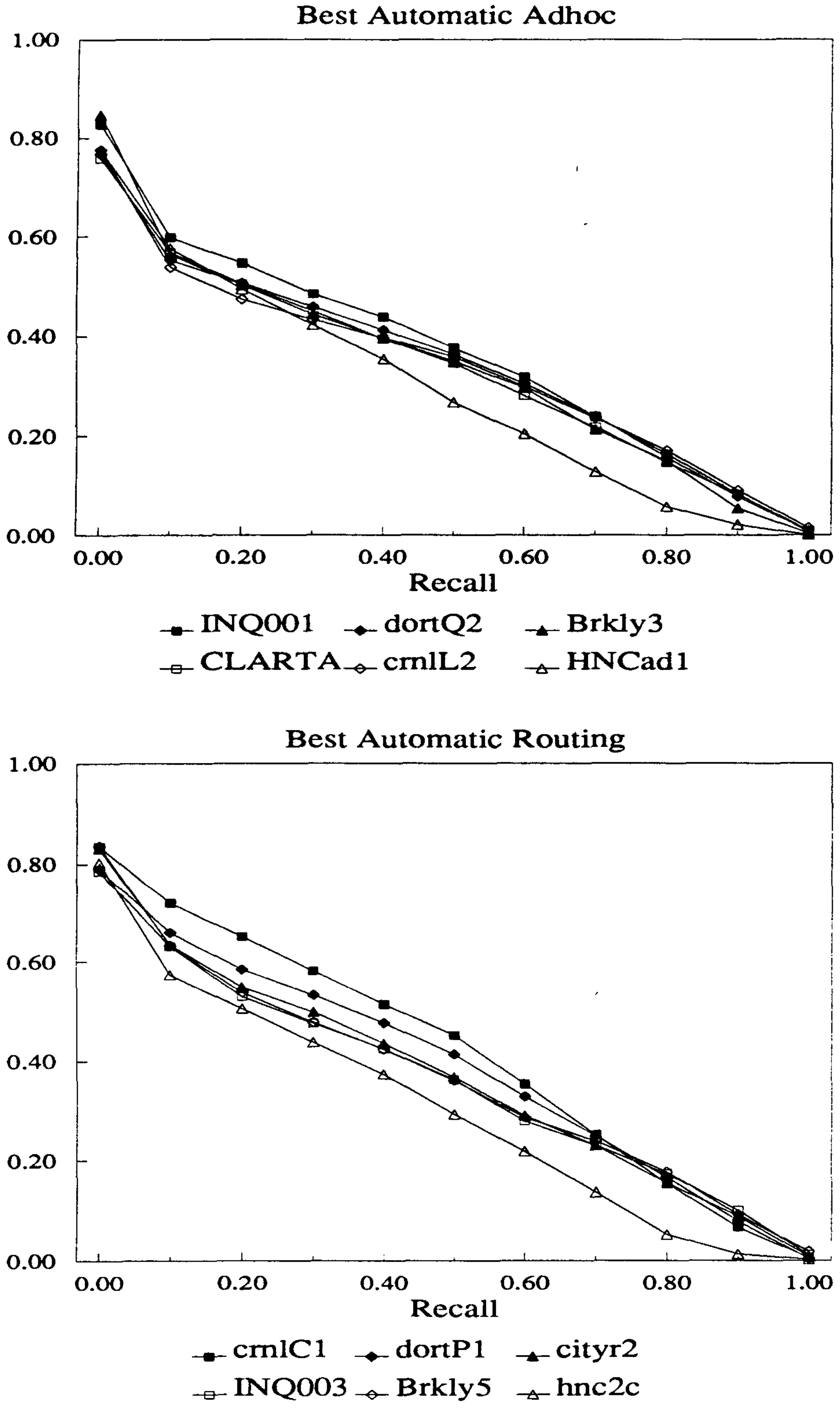
The second lasting value is the demonstration of the feasibility of using the non-Boolean, statistically-based retrieval systems both in the ARPA community and in the broader commercial sector. Not only have well-established small-scale research groups braved the scaling effort, but at least four new commercial products have used the TIPSTER/TREC program as launching pads. The TIPSTER program has caused the establishment of two major new retrieval research groups; both Syracuse University and HNC Inc. have built systems during the TIPSTER project that are approaching the power of the best of the TIPSTER/TREC systems. Additionally many of the TREC systems are either new groups in the information retrieval research arena or are older groups expanding their small programs to tackle this major retrieval experiment.

The final lasting value of the TIPSTER project has been the joining of the NLP community and the information retrieval (IR) community in the project. This has led to the high expections for combining these disjoint technologies in phase II and has helped cement the important collaboration of two diverse groups of researchers.

These three lasting contributions are not only of value individually, but will lead to a resurgence of research in the information retrieval area. The combination of the large test collection, the growing demand for improved retrieval products, and the increased collaboration between the NLP and IR communities will result in new techniques that will finally achieve the breakthrough in performance that is TIPSTER's goal.

8. REFERENCES

[1] Harman D. (Ed.).*The Second Text REtrieval Conference (TREC-2).* National Institute of Standards and Technology Special Publication 500-215, in press.



Figures 13 and 14: Best TREC-2 adhoc and routing performance using full collection

Topic	Top 8 Systems							
101	rutcomb1	VTcms2	crnlV2	INQ002	dortQ2	pircs3	Brkly3	CLARTM
102	crnlL2	crnlV2	VTcms2	siems3	dortL2	INQ002	siems2	CLARTM
103	siems3	siems2	schau1	citri1	crnlV2	lsiasm	HNCad2	HNCad1
104	dortQ2	CLARTM	CLARTA	pircs4	pircs3	dortL2	HNCad2	lsiasm
105	citri2	lsiasm	citri1	siems2	siems3	crnlV2	schau1	crnlL2
106	VTcms2	INQ002	INQ001	TOPIC2	pircs4	pircs3	CLARTM	dortL2
107	CnQst1	CnQst2	rutcomb1	TOPIC2	VTcms2	INQ002	rutfmed	CLARTM
108	citri1	dortQ2	siems3	VTcms2	siems2	HNCad2	schau1	dortL2
109	dortL2	crnlL2	dortQ2	CLARTA	CLARTM	pircs3	crnlV2	pircs4
110	INQ002	INQ001	Brkly3	dortQ2	nyuir3	nyuir2	cityau	siems2
111	CLARTA	CLARTM	INQ001	dortQ2	Brkly3	siems2	siems3	pircs4
112	INQ002	INQ001	VTcms2	nyuir2	nyuir3	HNCad1	HNCad2	CnQst2
113	VTcms2	crnlL2	dortL2	crnlV2	nyuir1	siems2	CLARTM	INQ002
114	INQ002	cityau	VTcms2	INQ001	siems3	siems2	lsia1	TOPIC2
115	nyuir2	nyuir3	nyuir1	siems2	dortL2	crnlV2	siems3	crnlL2
116	VTcms2	CLARTA	HNCad2	HNCad1	siems3	siems2	CLARTM	Brkly3
117	citri2	citri1	dortQ2	INQ001	TMC8	lsiasm	gecrd2	schau1
118	nyuir2	nyuir3	nyuir1	TOPIC2	citymf	dortQ2	CLARTA	INQ001
119	nyuir1	nyuir2	nyuir3	INQ002	INQ001	dortQ2	citymf	VTcms2
120	citymf	nyuir2	nyuir3	nyuir1	CnQst2	CnQst1	VTcms2	erima2
121	TOPIC2	CLARTM	VTcms2	Brkly3	nyuir1	prceo1	INQ002	rutfmed
122	siems2	siems3	INQ002	INQ001	dortQ2	Brkly3	CLARTM	crnlV2
123	nyuir1	nyuir2	nyuir3	CLARTA	INQ001	INQ002	CLARTM	pircs4
124	nyuir2	nyuir3	nyuir1	dortL2	dortQ2	INQ001	Brkly3	TMC9
125	crnlV2	Brkly3	crnlL2	CLARTM	siems3	CLARTA	pircs4	pircs3
126	siems3	crnlL2	siems2	Brkly3	crnlV2	INQ002	CLARTM	INQ001
127	cityau	Brkly3	CLARTA	HNCad2	INQ001	INQ002	siems2	siems3
128	VTcms2	CLARTA	siems3	siems2	CLARTM	TOPIC2	citri1	lsiasm
129	INQ001	INQ002	cityau	CLARTM	siems2	Brkly3	crnlL2	CLARTA
130	INQ002	INQ001	dortQ2	crnlL2	pircs4	CLARTM	dortL2	pircs3
131	TOPIC2	VTcms2	HNCad1	HNCad2	siems3	Brkly3	siems2	INQ002
132	dortL2	INQ001	INQ002	citri1	citri2	dortQ2	HNCad2	crnlL2
133	CnQst2	CnQst1	rutcomb1	pircs4	INQ002	pircs3	cityau	INQ001
134	crnlL2	dortL2	nyuir1	nyuir2	nyuir3	INQ002	INQ001	dortQ2
135	nyuir2	nyuir3	nyuir1	Brkly3	INQ001	INQ002	siems3	siems2
136	VTcms2	CnQst1	CnQst2	CLARTM	pircs4	CLARTA	dortQ2	TOPIC2
137	CLARTA	nyuir2	nyuir3	Brkly3	siems2	siems3	CLARTM	nyuir1
138	nyuir2	nyuir3	rutfmed	rutcomb1	nyuir1	schau1	gecrd2	citri1
139	nyuir2	nyuir3	nyuir1	VTcms2	dortL2	HNCad2	dortQ2	HNCad1
140	nyuir2	nyuir3	nyuir1	dortQ2	dortL2	INQ002	siems3	siems2
141	VTcms2	INQ002	CnQst2	INQ001	Brkly3	dortL2	dortQ2	CnQst1
142	dortQ2	siems2	crnlL2	VTcms2	siems3	CLARTM	crnlV2	Brkly3
143	INQ002	INQ001	siems2	siems3	crnlL2	crnlV2	nyuir2	nyuir3
144	VTcms2	Brkly3	citymf	crnlV2	siems3	lsiasm	siems2	HNCad2
145	crnlL2	crnlV2	dortL2	CLARTM	nyuir1	siems3	siems2	dortQ2
146	Brkly3	siems3	siems2	lsiasm	crnlV2	schau1	CLARTM	citri1
147	HNCad2	HNCad1	VTcms2	citri1	INQ002	INQ001	citymf	CLARTA
148	lsiasm	crnlL2	crnlV2	siems2	siems3	Brkly3	dortL2	dortQ2
149	nyuir1	CnQst2	TOPIC2	CnQst1	CLARTA	rutfmed	Brkly3	rutcomb1
150	crnlL2	dortQ2	CLARTM	siems3	INQ002	INQ001	crnlV2	siems2

Table 1: The TREC-2 system rankings (using average precision) on individual topics

Relevant Documents	HNCrt1	INQ023
AP900115-0233	-	-
AP900410-0212	-	-
AP900905-0174	-	-
SJMN91-06059027	-	-
SJMN91-06072107	-	-
AP900910-0280	1	360
AP900914-0252	3	192
AP901018-0234	4	288
AP900822-0232	8	499
SJMN91-06034021	10	23
SJMN91-06063161	54	140
AP900818-0028	208	549
AP900924-0260	-	245
AP900906-0203	-	322
AP900903-0137	559	555
AP900816-0111	849	-
AP900829-0239	904	668

Table 2: Ranks of retrieved relevant documents for topic 89

Relevant Documents	Rank	Run Tag
AP880214-0002	80	INQ010
AP880223-0008	26	INQ013
AP880622-0070	59	INQ010
AP880815-0056	40	INQ010
AP890522-0036	5	INQ010
AP891004-0223	39	INQ013
AP891130-0147	49	INQ010
AP891206-0043	23	INQ013
WSJ870325-0156	51	INQ013
WSJ900801-0135	124	INQ011,INQ012
ZF08-270-494	105	HNCad2
ZF08-305-768	3	HNCad1 (all)
ZF08-386-296	57	HNCad2

Table 3: Relevant Documents for topic 121 (55 total relevant, 13 found TIPSTER 24-month)