Design and Evaluation of the CLARIT–TREC-2 System

David A. Evans[1,2] **and Robert G. Lefferts**[2]

[1]**Laboratory for Computational Linguistics** **and** [2]**CLARIT Corporation**
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213-3890

Suite 200A, 319 South Craig St.
Pittsburgh, Pennsylvania 15213-3726

1 Introduction

The CLARIT team used the opportunity of the TREC-2 evaluations to explore several facets of the CLARIT system. In particular, given the performance of the CLARIT system on TREC-1 tasks (Evans *et al*. 1993), we focused our attention on evaluating

1. *fully-automatic processing* of topics and potentially-relevant documents and
2. *topic/query augmentation* using CLARIT thesaurus-discovery techniques.

All of the results we report in this paper follow from straightforward applications of base-level CLARIT processing, utilizing essentially the same CLARIT components that were employed in the CLARIT–TREC-1 system. The general improvements we observe in CLARIT–TREC-2 processing are attributable to modifications (especially simplifications) in processing steps and in the settings of system variables.

In the following sections, we describe the CLARIT–TREC-2 system, report our official processing results, and offer a brief analysis of performance. In addition, we report on several subsequent experiments we have conducted on the TREC-2 collection that test the parameters of the CLARIT–TREC-2 system and identify sources of immediate improvements in processing.

2 CLARIT–TREC-2 System Description and Processing Method

The CLARIT–TREC-2 system reflects a re-organization of the tools and techniques employed in the CLARIT–TREC-1 system. One of our principal goals was to streamline CLARIT processing and to establish a baseline method that is amenable to parameterization and analysis. As a consequence, the flow of data in the CLARIT–TREC-2 system is simple, straightforward, and efficient; furthermore, all CLARIT processing is fully automatic.

2.1 Changes from TREC-1

The essential differences between the CLARIT–TREC-1 and TREC-2 systems are in the preparation and evaluation of queries (TREC-2 "topics") and the automation of steps designed to identify and process potentially relevant documents for use in query augmentation. The following summaries highlight these points.

- **One-Pass Querying.** The CLARIT–TREC-1 system employed a two-step process to retrieve documents—a first pass for partitioning ("evoking") and a second pass for final ranking ("discrimination"). This has been eliminated in the CLARIT–TREC-2 system. Querying takes place in one step over the entire collection using vector-space-retrieval methods.
- **Automatic Query Creation.** The CLARIT–TREC-1 system was categorized as a "manual" system, though the required manual intervention was minimal. In particular, users were expected to assign an *importance coefficient* (with possible values "1", "2", or "3") to the CLARIT-parsed terms in a topic statement and possibly also to add terms to or delete terms from the CLARIT-generated list. In the CLARIT–TREC-2 system, the importance coefficient is assigned automatically by simple heuristics (described below). While users are still free to modify coefficients or terms, such intervention is not required. All "CLARTA" results reported in this paper reflect processing in which queries were fully automatically prepared by the CLARIT system, without review or modification.
- **Automatic Retrieval Refinement.** When processing ad-hoc queries, the CLARIT–TREC-1 system required that the user evaluate a few of the top-ranked retrieved documents. User-nominated documents were processed to identify terms for use in supplementing the source query. In the CLARIT–TREC-2 system, user evaluations are not required. Initial querying is accurate enough to support the automatic processing of the highest-scoring retrieved documents without 'inspection'.

- **Sub-Document Processing.** The CLARIT–TREC-1 system treated all documents as whole texts; retrieval 'scores' were calculated over full documents. The CLARIT–TREC-2 system treats all documents as collections of one or more sub-documents, operationalized as variable-sized units of approximately paragraph length. Such units are used as the basis for all statistical calculations and for measuring 'similarity' to a query. A full document is assigned the score (*e.g.*, for ranking) of the highest-scoring sub-document it contains.

2.2 Processing Method

Figure 1 offers a schematic overview of processing in the CLARIT–TREC-2 system. All topics were parsed for noun phrases. These, in turn, were either manually ("CLARTM") or automatically ("CLARTA") assigned weights (values "1", "2", or "3") for 'importance'. The terms for each topic were automatically supplemented with terms from a (pseudo-)thesaurus, automatically extracted from available known-relevant documents (in the case of routing topics) or from the top-ranked sub-documents returned in a first-pass querying of the TREC-2 collection (in the case of ad-hoc topics). All instances of retrieval took place over the applicable full set of documents, which had undergone an initial round of CLARIT processing (parsing).

The CLARIT–TREC-2 system incorporates a vector-space retrieval system that uses several CLARIT-specific techniques to improve retrieval results. The principal techniques involve the use of (1) natural-language processing to identify and normalize indexing terms, (2) fully automatic query augmentation based on CLARIT thesaurus discovery, and (3) simple text-analysis heuristics to approximate the effect of more sophisticated discourse analysis of texts. These techniques are described in greater detail in the following sections.

2.2.1 Natural-Language Processing

CLARIT natural-language processing (NLP) encompasses an inflectional morphological analyzer for word recognition and normalization and a deterministic rule-based parser for phrase identification. For TREC-2 processing, only simplex noun phrases (NPs) were used. Simplex NPs are phrasal constitutents that include the modifiers and head noun(s) of an NP but not the post-head prepositional phrases, relative clauses, or verb constructions. The CLARIT parser can provide a more complex linguistic analysis of texts, but such additional detail was not used in TREC-2 experiments.

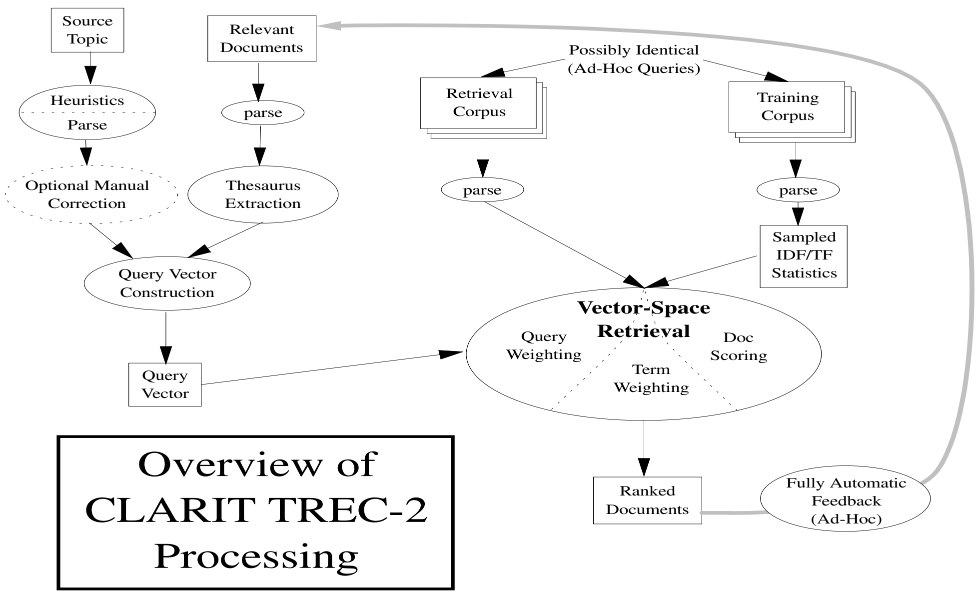

Figure 1: Overview of CLARIT–TREC-2 Processing

2.2.2 Query Augmentation

CLARIT thesaurus discovery is a process that can identify a set of core, representative terminology in a collection of documents. If the documents are relatively homogeneous topically, the resulting 'first-order' thesaurus can be regarded as a broad 'signature' of the topic of the collection. Typically, the procedure is most reliable when the collection is large—2 megabytes or more of text. In the CLARIT–TREC-2 system, however, thesaurus discovery was used on the relatively small collections of relevant documents for each topic. For each routing topic, the document collection used to establish a first-order thesaurus consisted of the training set of relevant documents for each topic; for each ad-hoc topic, a set of the top-ranked sub-documents from among the documents automatically retrieved via an initial pass of querying over the corpus. Terms in the discovered thesaurus were used to supplement the terms in the source query (topic). Thus, in the case of ad-hoc topics, the procedure represents an approach to query augmentation based on fully automatic feedback.

2.2.3 Text Analysis Heuristics

Texts can have complex and interesting discourse structure, which often provides clues about the 'topic(s)' and the 'important' information in a document. In general, it is very difficult to exploit text-structure information reliably in retrieval tasks over large collections of heterogeneous documents. Nevertheless, the CLARIT–TREC-2 system applied two simple processing techniques to topics and corpus documents to attempt to capture information encoded in rhetorical structure.

First, all training and corpus documents were divided into paragraph-sized units called "sub-documents". This procedure is sensitive both to the 'normal' demarcation of paragraphs (successive blank lines) and also to the total length of the text (measured in numbers of sentences). After documents are partioned into sub-documents, the sub-document is taken as the basic unit for subsequent processing, *viz.*, in collecting statistics and in scoring retrieval.

Second, terms extracted from a topic description were assigned importance coefficients based on their locations in the topic text. Terms found in the first paragraph are given a weight of 3; terms in the second paragraph, 2; and all other terms, 1.

Of course, both techniques exploit possibly idiosyncratic characterisitics of the TREC-2 processing task. The use of scoring over sub-documents is clearly sensitive to the TREC definition of document relevancy, *viz.*, that a document is relevant regardless of length if it contains a single relevant sentence. Furthermore, automatic term importance weighting is possible only because of the formal discourse structure of TREC topics; it would not neccessarily apply to other presentations of topics or queries and certainly would not apply to free text in general.

We observe, however, that there is no one set of techniques that will perform optimally in every information-retrieval (IR) situation. We emphasize, instead, that one important measure of a system's utility and adaptability is its ability to take advantage of exploitable features in a given IR task.

2.3 System Performance Notes

All CLARIT–TREC-2 processing took place on DEC 3000/400 (ALPHA/AXP) workstations running DEC OSF/1. One system had 128 megabytes of RAM, one 64 megabytes, and two 32 megabytes. Realized performance was considerably slower than would be expected given the clockrate (133.33 MHz) of the DEC 3000/400's CPU. In fact, because of suboptimal compilers for the 64-bit architecture, performance was perhaps two times slower than the maximum possible. One pass over the entire TREC-2 collection for 50 topics (processed simultaneously) required approximately four hours, running on the four machines in parallel. The processing of a single topic is proportionally faster, requiring approximately 20 minutes on a single machine.

3 Results

The CLARIT team processed all the topics in both the "routing" and "ad-hoc" categories and worked with the full set of data. Two sets of results were submitted for each category, corresponding to the "manual" ("CLARTM") and fully-"automatic" ("CLARTA") processing approaches taken with the topics.

3.1 General Summary of Official Results

Table 1 gives the official CLARIT–TREC-2 system routing results as reported by NIST. A graph of the precision–recall curves for the two sets of results is given in Figure 2. The total number of documents retrieved under the routing task was 6,785 (CLARTM) and 6,811 (CLARTA), representing, respectively, 64.69% and 64.93% of the total known relevants (10,489).

Table 2 gives the official CLARIT–TREC-2 system ad-hoc query results as reported by NIST. A graph of the precision–recall curves for the two sets of results is given in Figure 3. The total number of documents retrieved under the ad-hoc query task was 8,229 (CLARTM) and 8,109 (CLARTA), representing, respectively, 76.30% and 75.19% of the total known relevants (10,785).

The graph in Figure 4 shows the average precision score for each process at *N* documents, for selected values of *N*. It should be noted that the maximum possible precision score at 500 and 1,000 documents is less than 100%. In particular, the average number of relevants per routing topic is 209.78; this corresponds to a maximum precision of 41.96% at 500 documents and 20.98% at 1,000 documents. The average number of relevants per ad-hoc query topic is 215.70; this corresponds to a maximum precision of 43.14% at 500 documents and 21.57% at 1,000 documents.

Tables 3 and 4 provide another view of total performance. The numbers in each cell give the number of times the CLARIT–TREC-2 system produced results above, equal to, or below the median for all TREC-participant systems. Numbers in brackets give the instances of 'extreme' performance—best and worst—among all systems. For the routing topics, for example, CLARIT retrieval results at 1,000 documents were better than the median 36 times in both "manual" and "automatic" modes; CLARIT scored the maximum 10 and 11 times, respectively. For the ad-hoc query topics, CLARIT retrieval results at 1,000 documents were better than the median 44 times in "manual" mode and 42 times in "automatic" mode; CLARIT scored the maximum 4 and 9 times, respectively.

3.2 CLARIT Automatic vs. Manual Modes of Processing

In both tasks (routing and ad-hoc querying), CLARIT–TREC-2 automatic processing results are virtually identical to manual results. This confirms our hypothesis that the principal contribution to performance derives from (1) the base-level CLARIT process (using linguistic phrases as information units) and (2) the effect of query augmentation via thesaural terms. On this latter point, we note that, on average, the final query vector for a topic will contain many more terms that derive from thesaurus extraction than terms that derive from the source topic. In general, then, when reliable information is available (as in sample known relevants or highly-likely relevants returned in a first-pass retrieval), the CLARIT process will succeed in finding good supplemental terminology for a topic and the overall effects of manual intervention will be minimized.[1]

Figures 5 and 6 illustrate the relative absence of a positive effect for manual intervention in the selection and weighting of query terms. There are approximately as many instances of decreased performance as there are instances of increased performance. Most topics show very little percentage difference in numbers of documents returned;[2] this is especially underscored in the results for routing topics at 1,000 documents.

4 Analysis

4.1 CLARIT Precision

As in TREC-1, CLARIT precision–recall curves demonstrate very high precision at low percentages of recall. The first few documents returned by the system are extremely likely to be relevant for the given topic. This fact of CLARIT processing was successfully exploited in the TREC-2 processing method: query augmentation was possible because there was, in general, a good concentration of topic-relevant information among the sub-documents of the first-pass returned documents.

As shown in Figure 4, precision remains quite stable for all methods across the first 30 documents retrieved and is relatively high across the full retrieved set of 1,000 documents.

5 Query Augmentation Experiments

A distinguishing feature of the CLARIT–TREC-2 system is the use of fully-automatic query augmentation. As noted above, the selection of terms for query augmentation depends on (1) the selection of a source set of known- or nominated-relevant documents and (2) the application of the CLARIT thesaurus-discovery procedure. Since the size (and quality) of the source set of documents can vary and since CLARIT thesaurus-discovery processing can be adjusted to nominate relatively greater or fewer numbers of terms, the 'query-augmentation' facet of the CLARIT process is a natural source of potential variation in system performance.

[1] Of course, there may be some forms of manual intervention—not utilized in the CLARIT 'manual' process—that would have effects dramatically better than the CLARIT automatic process. We know of no such process that can be applied efficiently to arbitrary topics and databases, however.

[2] Indeed, even in the absolute number of documents returned for each topic—not shown in the figures—there is very little difference.

Official TREC-2 Results

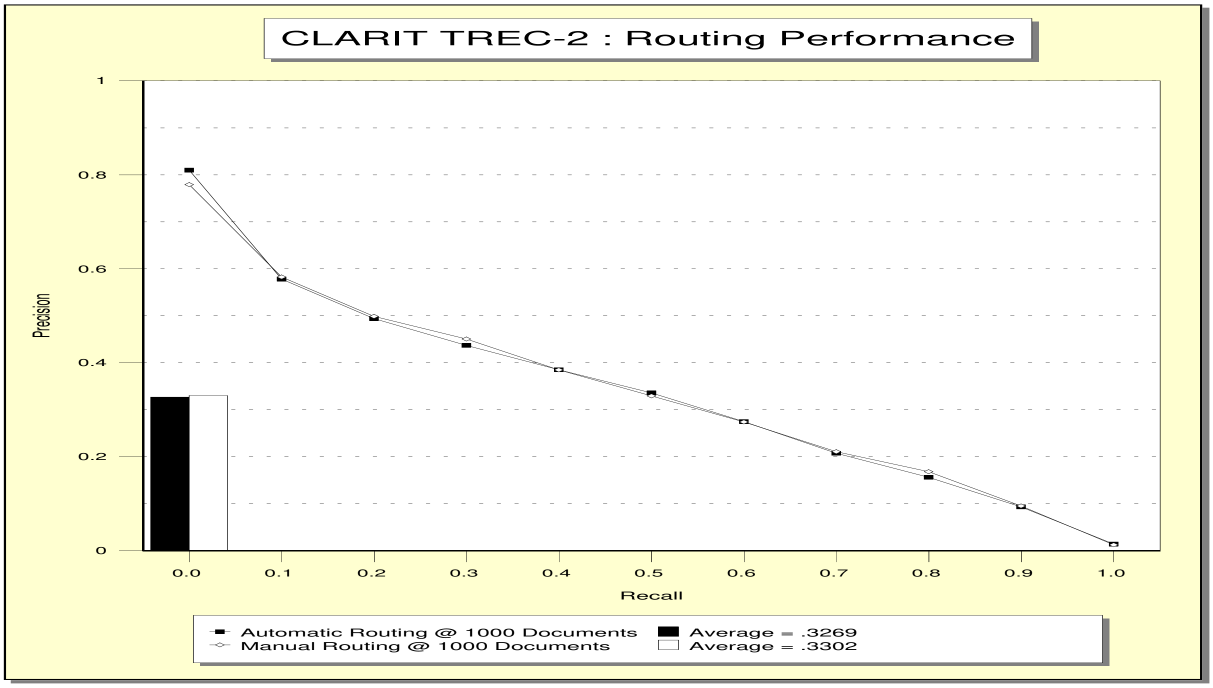

Figure 2: P–R Curve for Routing Over 1000 Docs—CLARTM and CLARTA

CLARTM Routing Queries, 51–100	
Total number of documents over all queries	
Retrieved:	50000
Relevant:	10489
Rel_ret:	6785
Interpolated Recall–Precision Averages:	
at 0.00	0.7791
at 0.10	0.5825
at 0.20	0.4987
at 0.30	0.4505
at 0.40	0.3848
at 0.50	0.3295
at 0.60	0.2739
at 0.70	0.2109
at 0.80	0.1683
at 0.90	0.0951
at 1.00	0.0126
Average Precision (non-interpolated)	
over all rel docs:	0.3302
Precision:	
At 5 docs:	0.6120
At 10 docs:	0.5940
At 15 docs:	0.5800
At 20 docs:	0.5700
At 30 docs:	0.5527
At 100 docs:	0.4396
At 200 docs:	0.3436
At 500 docs:	0.2176
At 1000 docs:	0.1357
R-Precision (precision after R retrieved)	
Exact	0.3642

CLARTA Routing Queries, 51–100	
Total number of documents over all queries	
Retrieved:	50000
Relevant:	10489
Rel_ret:	6811
Interpolated Recall–Precision Averages:	
at 0.00	0.8101
at 0.10	0.5781
at 0.20	0.4941
at 0.30	0.4368
at 0.40	0.3851
at 0.50	0.3358
at 0.60	0.2749
at 0.70	0.2074
at 0.80	0.1561
at 0.90	0.0933
at 1.00	0.0141
Average Precision (non-interpolated)	
over all rel docs:	0.3269
Precision:	
At 5 docs:	0.6200
At 10 docs:	0.6060
At 15 docs:	0.5893
At 20 docs:	0.5730
At 30 docs:	0.5460
At 100 docs:	0.4346
At 200 docs:	0.3416
At 500 docs:	0.2184
At 1000 docs:	0.1362
R-Precision (precision after R retrieved)	
Exact	0.3646

Table 1: Offical Results, Routing Topics 51–100

Official TREC-2 Results

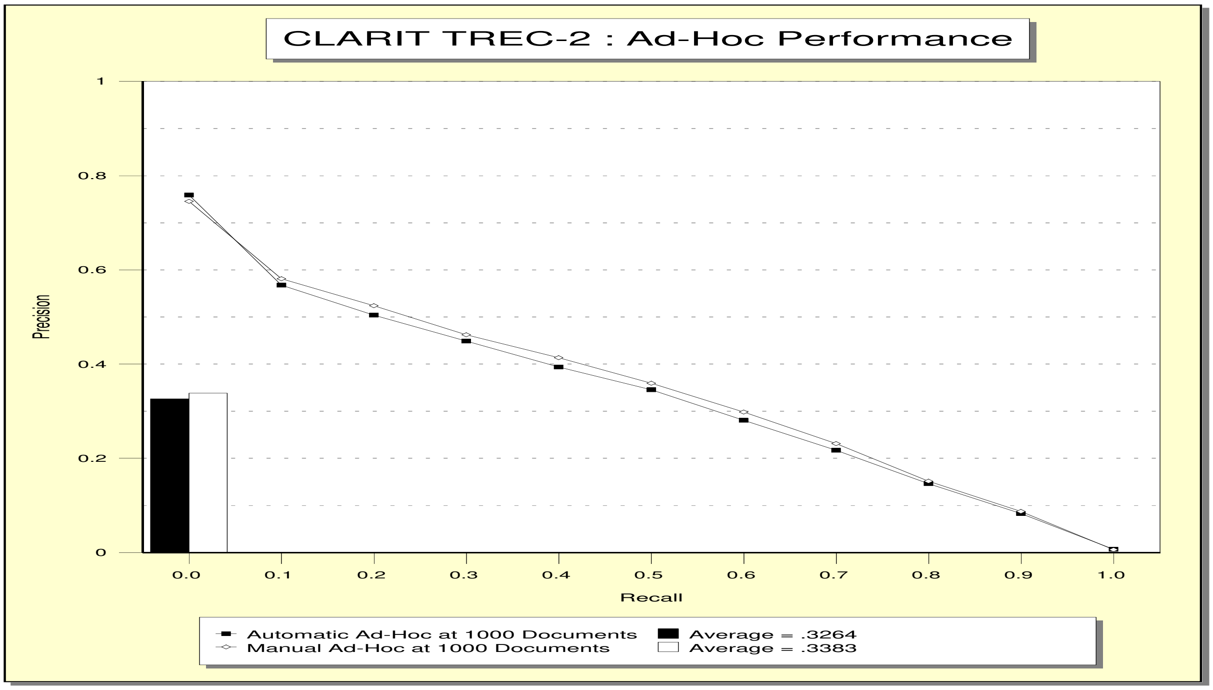

Figure 3: P–R Curve for Ad-Hoc Queries Over 1000 Docs—CLARTM and CLARTA

CLARTM Ad-Hoc Queries, 101–150	
Total number of documents over all queries	
Retrieved:	50000
Relevant:	10785
Rel_ret:	8229
Interpolated Recall–Precision Averages:	
at 0.00	0.7455
at 0.10	0.5811
at 0.20	0.5240
at 0.30	0.4622
at 0.40	0.4135
at 0.50	0.3593
at 0.60	0.2983
at 0.70	0.2312
at 0.80	0.1516
at 0.90	0.0874
at 1.00	0.0062
Average Precision (non-interpolated)	
over all rel docs:	0.3383
Precision:	
At 5 docs:	0.5840
At 10 docs:	0.5740
At 15 docs:	0.5640
At 20 docs:	0.5590
At 30 docs:	0.5433
At 100 docs:	0.4846
At 200 docs:	0.3975
At 500 docs:	0.2601
At 1000 docs:	0.1646
R-Precision (precision after R retrieved)	
Exact	0.3741

CLARTA Ad-Hoc Queries, 101–150	
Total number of documents over all queries	
Retrieved:	50000
Relevant:	10785
Rel_ret:	8109
Interpolated Recall–Precision Averages:	
at 0.00	0.7587
at 0.10	0.5676
at 0.20	0.5038
at 0.30	0.4489
at 0.40	0.3938
at 0.50	0.3455
at 0.60	0.2806
at 0.70	0.2172
at 0.80	0.1463
at 0.90	0.0827
at 1.00	0.0070
Average Precision (non-interpolated)	
over all rel docs:	0.3264
Precision:	
At 5 docs:	0.5800
At 10 docs:	0.5680
At 15 docs:	0.5547
At 20 docs:	0.5490
At 30 docs:	0.5253
At 100 docs:	0.4644
At 200 docs:	0.3874
At 500 docs:	0.2542
At 1000 docs:	0.1622
R-Precision (precision after R retrieved)	
Exact	0.3645

Table 2: Offical Results, Ad-Hoc Topics 101–150

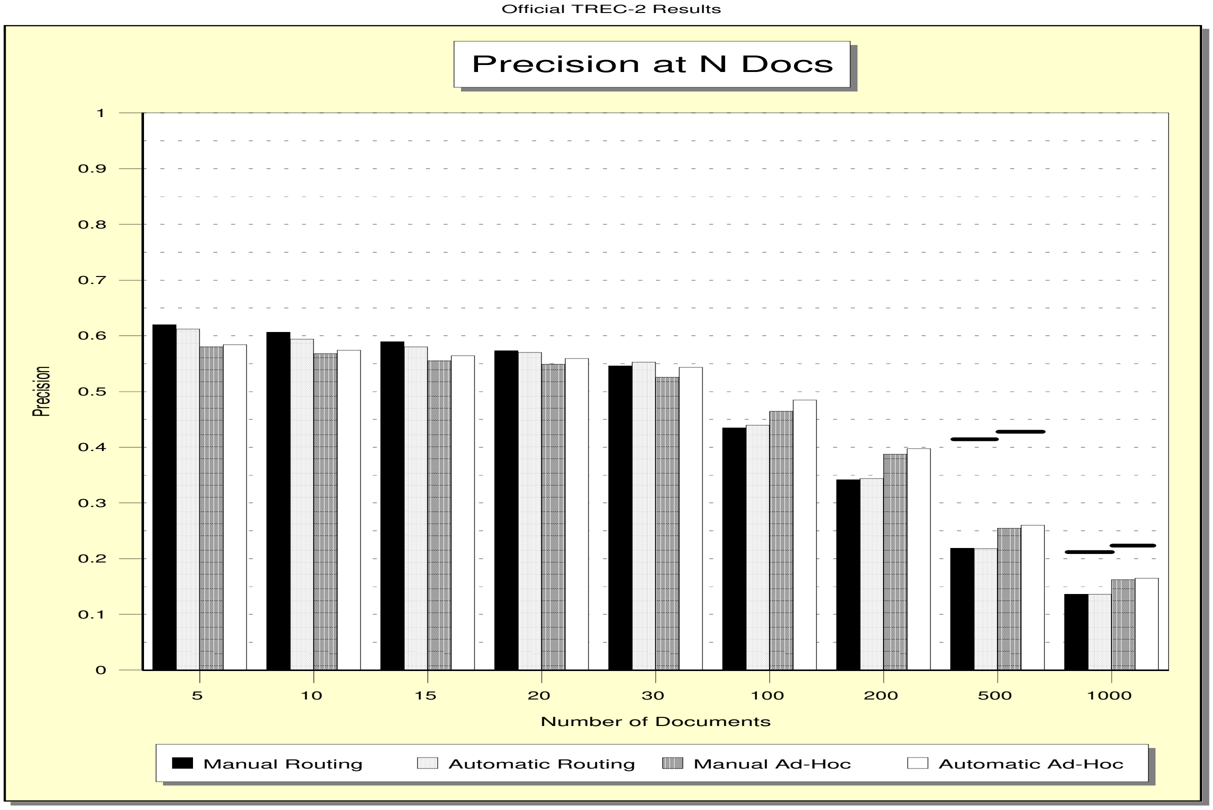

Figure 4: Comparative Precision at N Documents

	>Median		=Median		<Median	
	CLARTM	CLARTA	CLARTM	CLARTA	CLARTM	CLARTA
Precision	37 [1]	34 [0]	0	4	13 [0]	12 [0]
Rels in Top 100	32 [2]	29 [2]	6	8	12 [1]	13 [1]
Rels in Top 1000	36 [10]	36 [11]	5	6	9 [0]	8 [0]

Table 3: Summary of Results for Routing (51–100)

	>Median		=Median		<Median	
	CLARTM	CLARTA	CLARTM	CLARTA	CLARTM	CLARTA
Precision	38 [0]	34 [3]	4	7	8 [0]	9 [0]
Rels in Top 100	39 [1]	31 [0]	3	2	9 [0]	17 [0]
Rels in Top 1000	44 [4]	42 [9]	2	2	4 [0]	6 [0]

Table 4: Summary of Results for Ad-Hoc Queries (101–150)

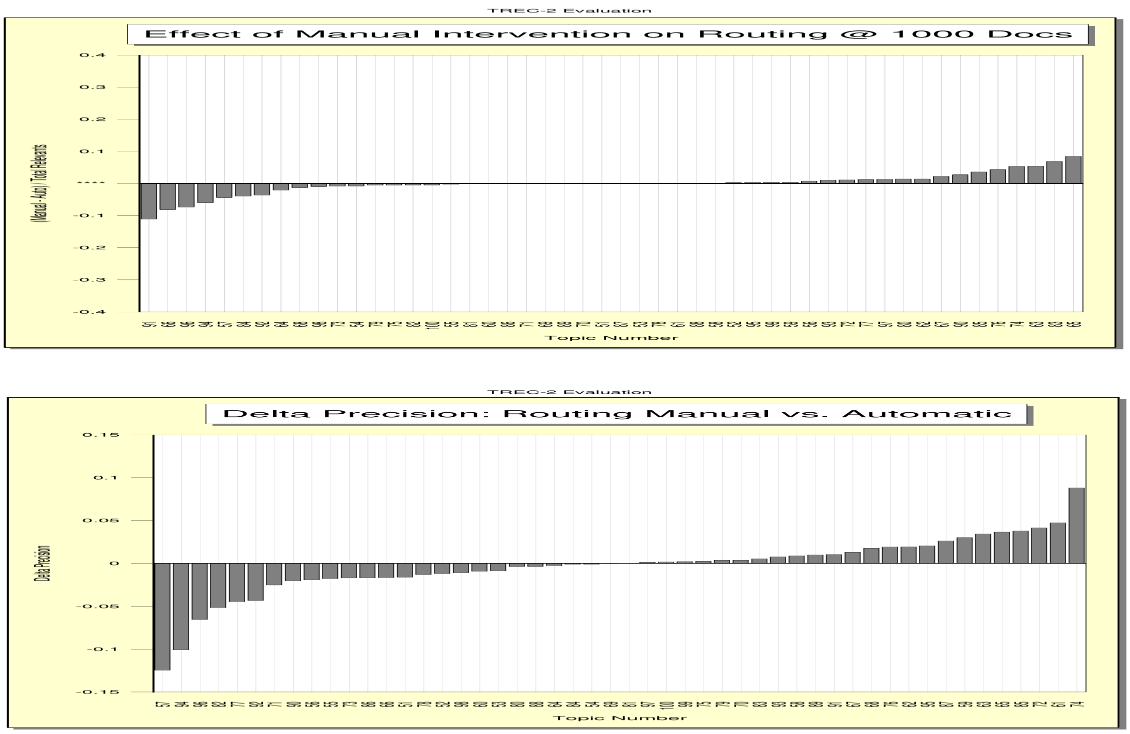

Figure 5: Manual vs. Automatic Preparation of Routing Queries

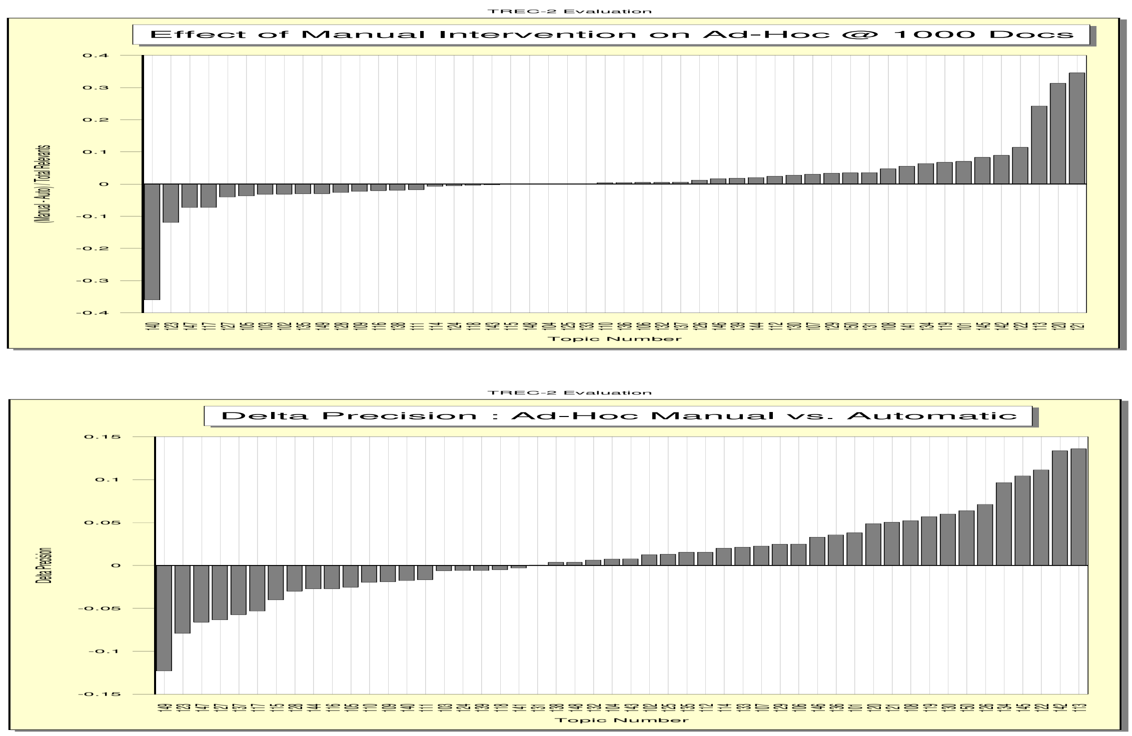

Figure 6: Manual vs. Automatic Preparation of Ad-Hoc Queries

Using TREC-2 processing results as a baseline, we have begun exploring several of the parameters of query augmentation in the CLARIT–TREC-2 system. The first set of experiments focuses on three parameters:

- The use of query augmentation compared to a procedure involving no query augmentation;
- The size of relevant document samples as source collections for thesaurus extraction; and
- The 'threshold' chosen for the thesaurus discovery process.

5.1 Query Augmentation vs. No Augmentation

In order to verify that CLARIT query augmentation techniques do have a positive effect on retrieval results, we have compared the official TREC-2 results against experimental runs that do not take advantage of any augmentation. For the routing queries, we re-ran the routing task using only the raw queries as vectors (consisting of terms from the topic statement only). The effects for manual and automatic modes are shown in Figure 7. In the case of the ad-hoc queries, the unaugmented results are simply the results of the initial round of querying, before automatic feedback has taken place. The effects for both modes are shown in Figure 8.

Query augmentation using known-relevant documents, as in the routing experiment, has a dramatic effect on system performance. The manual routing task showed a 69% improvement with augmentation, while automatic routing improved by 76%. Figure 9 shows the effect of query augmentation for individual routing topics, calculated as the difference in average precision between augmented and unaugmented queries. Note that very few queries do not improve with augmentation; most show great improvement.

Query augmentation is not as effective when used for automatic feedback in processing ad-hoc queries, but it still results in substantial improvements. For manual ad-hoc queries, we found a 21% improvement when using query augmentation; for automatic ad-hoc queries, a 22% improvement. Obviously, the use of known-relevant documents in the case of routing topics has an impact on query augmentation, but it is not the only important factor. The query-by-query effect of augmentation for ad-hoc queries is shown in Figure 10. While the effect is not as great as with routing topics, most queries show improvement with augmentation. By comparing the query-by-query results for automatic and manual modes, one can see that manual intervention in query formulation improves the results of augmentation in specific cases, such as query 121, but the positive effect is difficult to predict. Even without user review, we can have confidence in the ability of query augmentation to improve query results, given reasonably accurate initial query formulation.

We believe that the techniques we used in CLARIT–TREC-2 automatic feedback can be refined to give better documents as input to query augmentation. First, due to engineering and time constraints in TREC-2 processing, we did not select the best sub-documents from the entire corpus. Instead, we selected the best sub-documents from a pool of the best relevant documents, which might not always correspond to the optimal set of sub-documents for query augmentation. Second, the TREC-2 process used an absolute number of sub-documents—the top N—in query augmentation, regardless of the 'similarity' scores of those sub-documents to the source query. While it may not be possible to determine 'absolute' relevance on a query-by-query basis, minimum thresholds might be applied to exclude clearly irrelevant sub-documents, should such be within the N otherwise to be used in augmentation.

5.2 Source Sample Size

In the case of the submitted results of CLARIT–TREC-2 processing, the thesaurus for each routing query was extracted from the set of all its known-relevant training documents. However, we can imagine that many such documents contain some—perhaps a great deal of—information that is not relevant to the topic at hand. (This is especially expected with long documents.) Therefore, we have experimented with alternative approaches to sampling text from known-relevant documents. In particular, we have used a ranking of sub-documents (paragraphs) to nominate candidate 'good' relevant texts and have used variable numbers of sub-documents as source text in thesaurus extraction.

In practice, to select source text for the thesaurus-discovery process for a topic, we run the raw topic vector as a query over the collection of relevant documents, which are partitioned into sub-documents, and return only the top N relevant sub-documents. (In cases where the topic has fewer than N sub-documents in the collection of relevants, all sub-documents are selected.) Our hypothesis is that thesauri extracted from relevant sub-documents will contain greater numbers of 'true-positive' terms related to the topic. We have run the experiment for N = 100, 200, 300, 500, and 1000 as the cutoff for selected sub-documents. Figure 11 gives a sample of the results for routing topics, including the best case of 500 sub-documents.

While the differences for different N are not great, certain trends are evident. First, it is clear that using only the more similar sub-documents from a collection of relevants gives better results than using all relevant full documents. (This represents an important improvement over the method employed in the CLARIT–TREC-2 system.) Second, larger sample sizes seem to perform better than smaller ones. In the experiments that we ran, this effect peaks at 500 sub-documents, but such results probably interact with the average number of relevant sub-documents available. As noted previously, we would like ideally to use a variable number of sub-documents for each query, but such an approach requires an accurate measure of 'absolute' relevance.

5.3 Thesaurus Size

CLARIT thesaurus discovery techniques allow for the selection (sampling) of greater or fewer numbers of terms from a collection. In practice, the size of the sample is determined by a real-number threshold between 0.00 and 1.00. A larger threshold results in the inclusion of more general terminology from the document collection; a smaller threshold results in selection of terms that are more specific to the collection. (Intuitively, such variation correlates with the 'breadth' or 'narrowness' of the thesaurus.) The set of terms selected at a larger threshold will always properly include all of the terms that would be selected from the same collection at a smaller threshold. In CLARIT–TREC-2 processing, the threshold was set at 0.50. Note, however, that in the document-sample-size experiments reported above, we used a threshold of 0.75.

Using sub-document samples of relevant documents with N = 300 (reflecting the 'second-best' performance obtained in the experiments on sample size), we have explored the effects of using different thesaurus extraction thresholds—at 0.50, 0.75, 0.85, and 0.95 levels. Results for routing topics are given in Figure 12.

As with sample-size variation, differences in performance between increments are not dramatic, but trends do appear. First, the 0.50 thesaurus is clearly inferior and actually performs very much like the baseline CLARIT–TREC-2 system. Such a result indicates that much of the variation observed between baseline CLARIT–TREC-2 processing and our current 'best' technique is due to changes in thesaurus thresholds, rather than changes in the document sample size. Second, while all of the 0.75, 0.85, and 0.95 thesauri have the similar precision at low recall levels, the 0.75 thesaurus performs slightly better in the average case. From this we might hypothesize that the 0.75 value is close to the optimal threshold for thesaurus discovery in the context of query augmentation.

Figure 13 gives the results for the automatic routing task for the optimal number of sub-documents and thesaurus threshold compared to TREC-2 reported results, compared to the unaugmented baseline. Here we can see that a simple refinement to parameter setting yields a 5.5% overall improvement in average precision and 9.1%, 8.8%, and 7.6% improvement in precision at 10%, 20%, and 30% recall levels, respectively. In addition, there is a 6.3% improvement in total relevant documents (7,241 vs. 6,811). Furthermore, our experiments suggest mechanisms, such as measures of absolute relevance, that might result in further significant gains.

6 Conclusion

The CLARIT–TREC-2 system has successfully demonstated the ability to operate as a fully-automatic IR system. Since the performance differences between CLARIT manual and automatic processing modes are negligible, one can use CLARIT in fully-automatic mode and expect high precision and very good recall on retrieval tasks.

The TREC-2 results also demonstrate the efficacy of the CLARIT technique of automatic query augmentation. It is generally difficult for a user to predict whether the addition of terms to a query will have a positive or negative effect on performance. CLARIT query augmentation, using CLARIT thesaurus-discovery techniques, however, shows positive effects. Because the technique is fully automatic, it can be applied either at the time of query formulation (if exemplary relevant texts are known) or at the time of 'first-pass' retrieval. In either case, final results will be improved.

In several experiments, we have already identified two simple adjustments to CLARIT parameters that will improve performance beyond the CLARIT–TREC-2 system baseline. The system is not yet optimized; we expect to make other straightforward improvements.

Many text processing functions currently available in the CLARIT system or near completion were not used on TREC-2 documents. In future evaluations, we plan to utilize some of the more sophisticated functionality in the system. For example, we have been developing grammars for recognizing complex tokens such as proper names, dates, times, monetary values, *etc.*, but did not use token recognition modules in CLARIT–TREC processing. We believe that such token recognition will improve the results for queries involving specific persons or time intervals. Finally, we have also been experimenting with generating sub-corpus-derived equivalence classes for words and terms. We expect to use equivalence classes selectively to supplement thesaurus terms in query augmentation.

In sum, we believe that CLARIT–TREC-2 processing results demonstrate the power of CLARIT tools to solve IR tasks. The CLARIT–TREC-2 system represents only one of many possible configurations of CLARIT modules. In subsequent work, we plan on exploring other configurations.

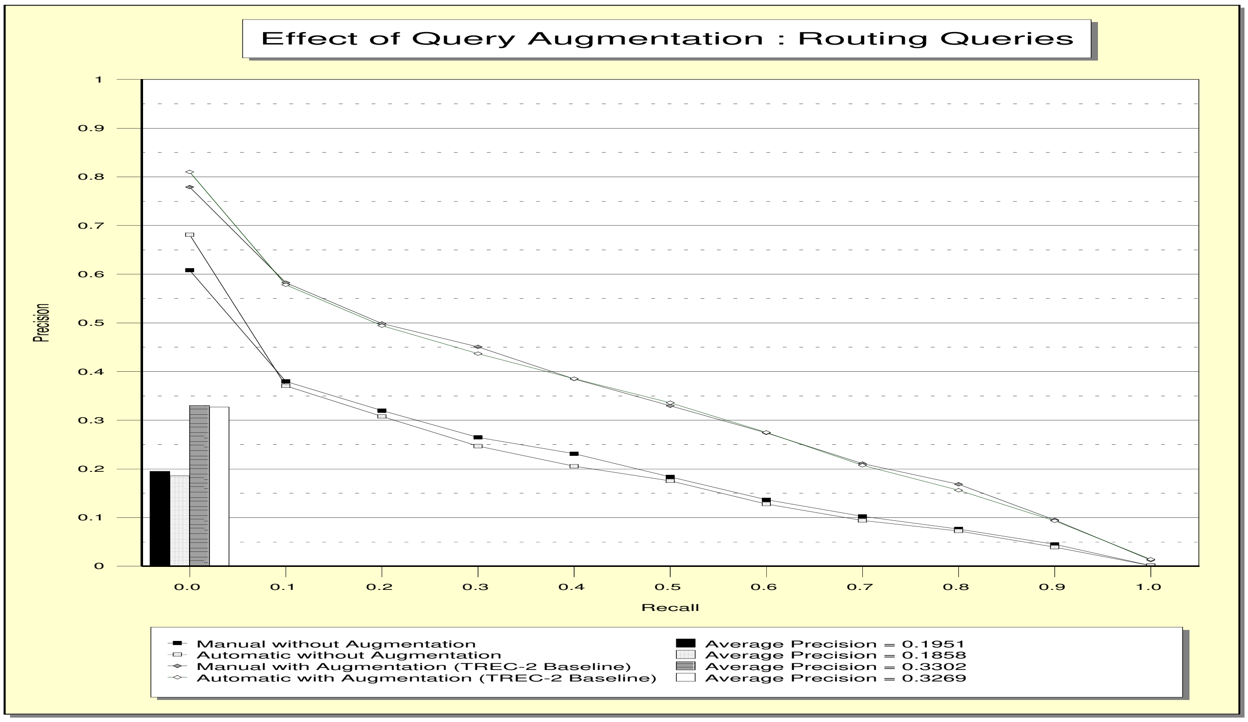

Figure 7: Effect of Query Augmentation on Routing Queries

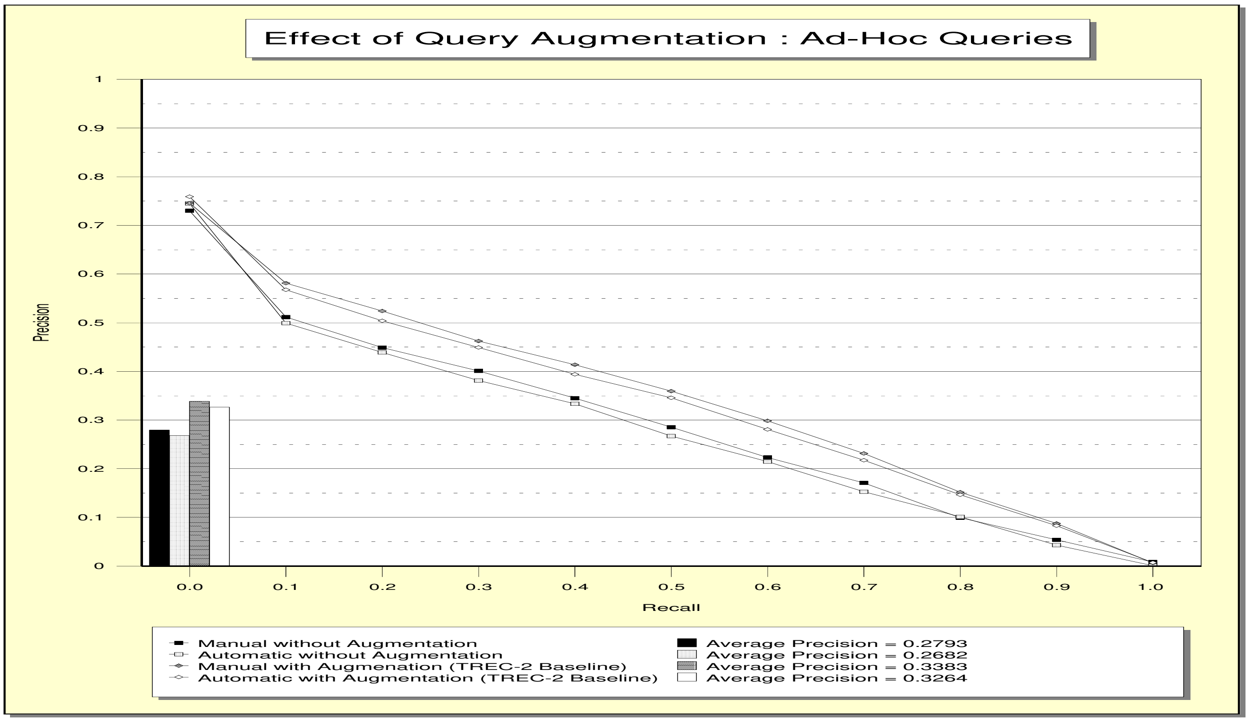

Figure 8: Effect of Query Augmentation on Ad-Hoc Queries

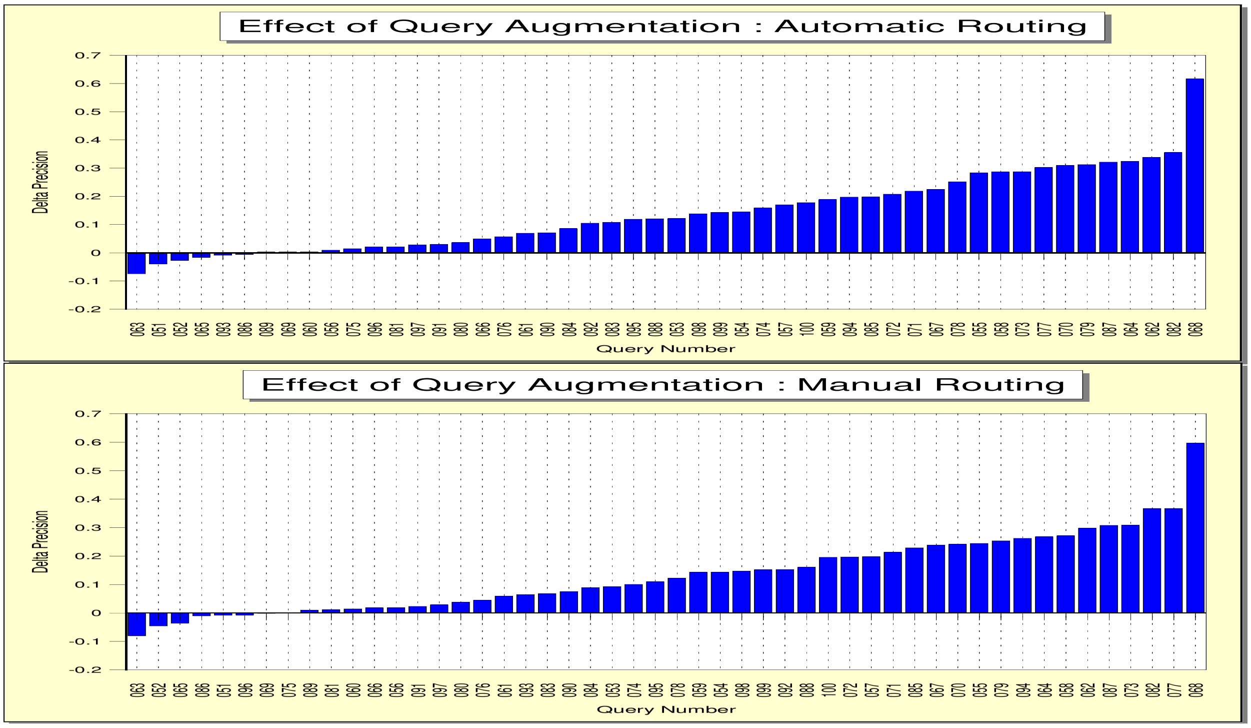

Figure 9: Query-by-Query Analysis of Effect of Query Augmentation on Routing Queries

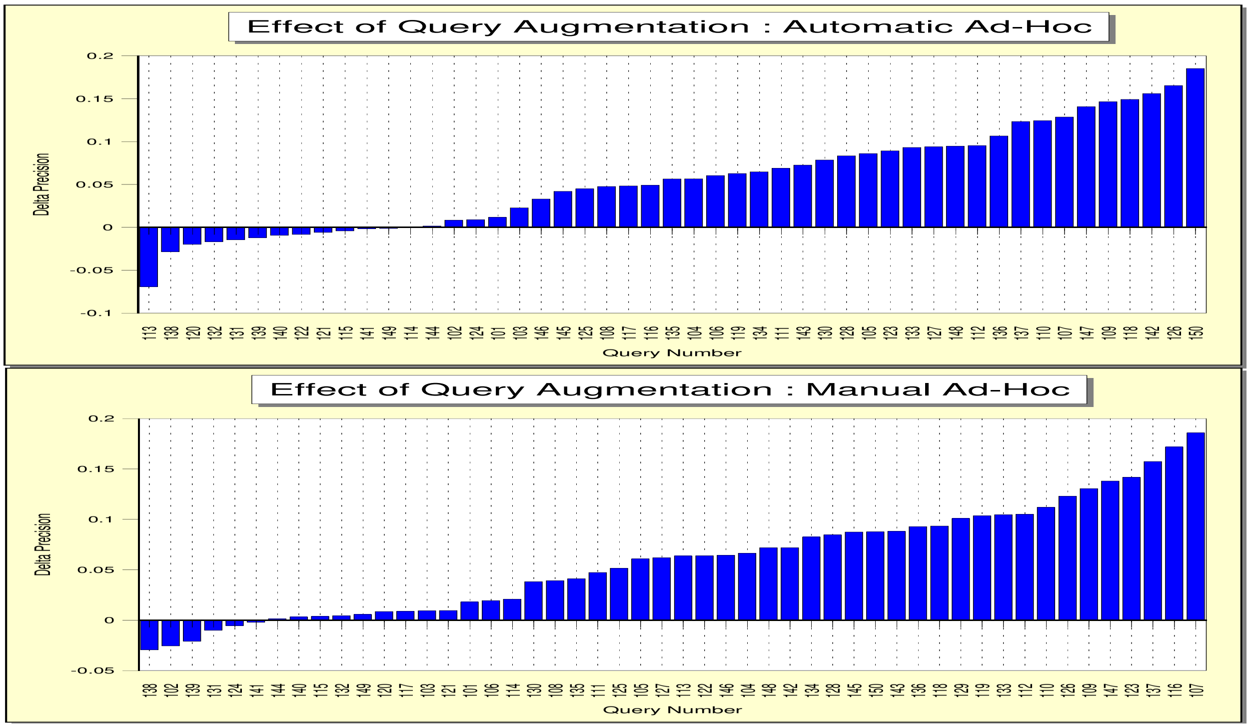

Figure 10: Query-by-Query Analysis of Effect of Query Augmentation on Ad-Hoc Queries

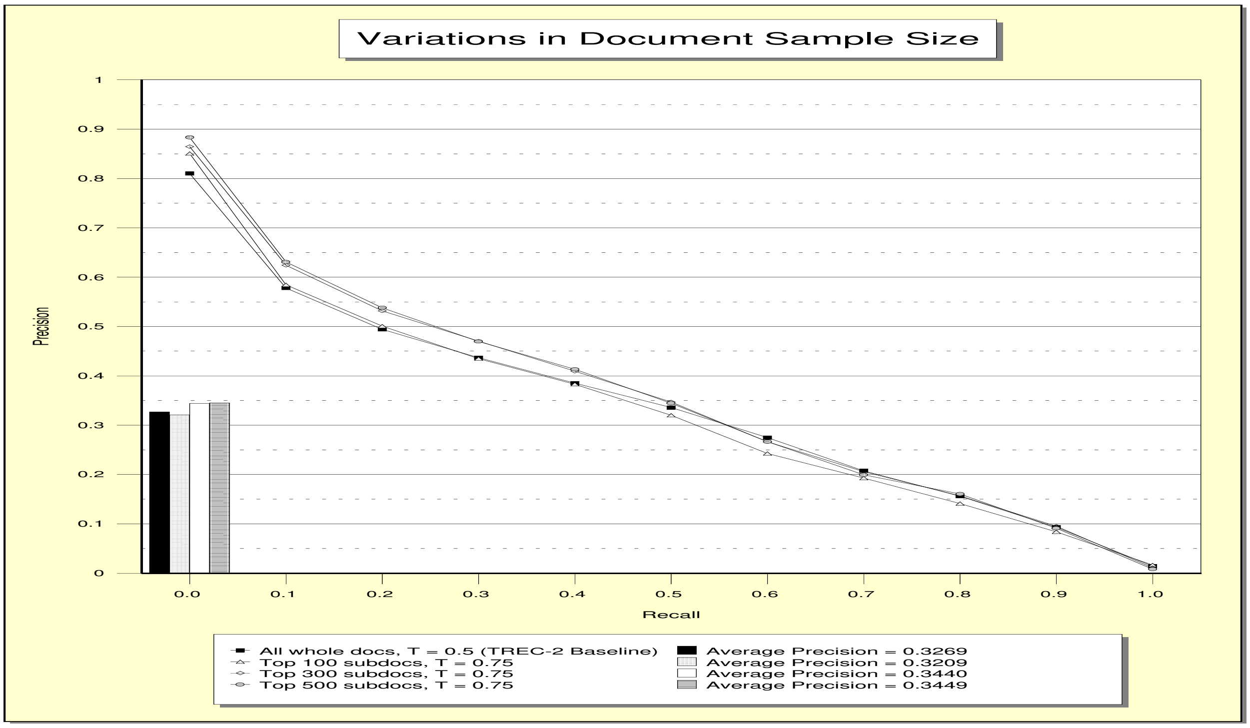

Figure 11: Effect of Sample Size (Routing Topics)

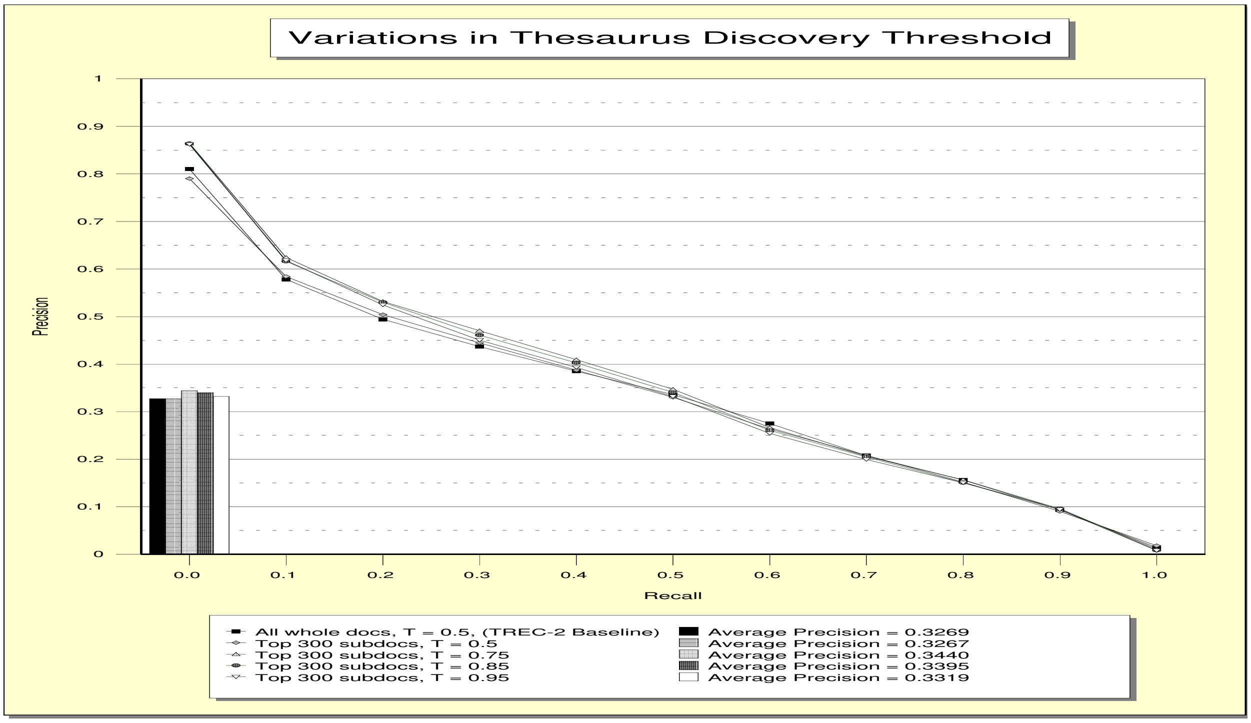

Figure 12: Effect of Thesaurus Size (Routing Topics)

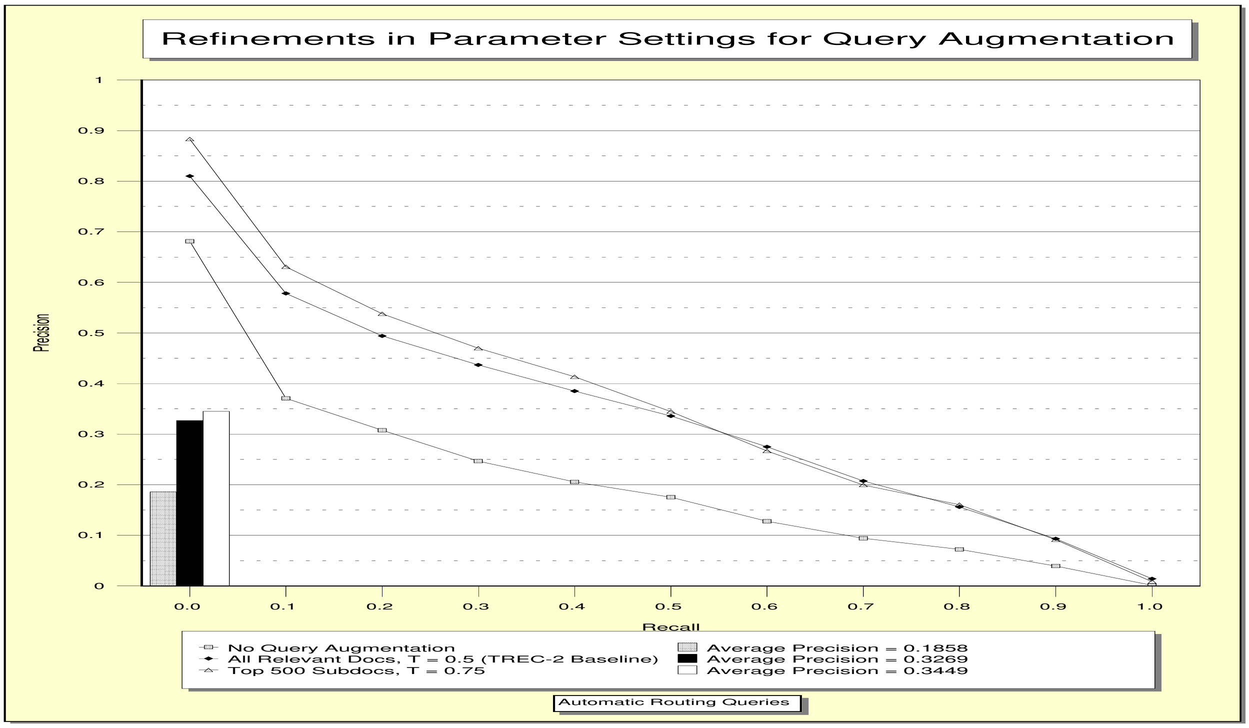

Figure 13: Optimized Sample Size and Thesaurus Threshold Results for Automatic Routing

7 Acknowledgements

CLARIT team participation in the TREC activities was made possible, in part, by grants from ARPA/NIST and CLARIT Corporation. We received valuable assistance from CLARIT-Project-team members Greg Grefenstette, Steve Handerson, and William R. Hersh and from CLARIT Corporation staff Armar A. Archbold, Michael McInerny, and Nataša Milić.

8 References

[Evans *et al*. 1993] Evans, David A., Lefferts, Robert G., Grefenstette, Gregory, Handerson, Steven K., Hersh, William R., and Archbold, Armar A., "CLARIT TREC Design, Experiments, and Results". In Donna Harman (Editor), *The First Text REtrieval Conference (TREC-1)*. NIST Special Publication 500-207. Washington, DC: U.S. Government Printing Office, 1993, 251–286; 494–501.