Query Expansion/Reduction and its Impact on Retrieval Effectiveness

X. Allan Lu and Robert B. Keefer
Mead Data Central, Inc.
P.O. Box 933
Dayton, OH 45401
(alan,robk)@meaddata.com

1.0 Introduction

Query expansion should help improve information retrieval effectiveness. Reported studies [1-5] using the TREC data generally support this position, though earlier studies [6-8] using smaller test databases did not obtain any conclusive results. Three important research questions remain open. First, assuming that a thesaurus will be used, should a general or data specific thesaurus be used in query expansion? Second, at what point does query expansion cease to add value? Finally, to what degree does query expansion help improve retrieval effectiveness?

A general thesaurus has been found to have little effect on information retrieval results, and may even cause negative results. Prior studies [1-5] have observed that a thesaurus derived from the data on which the retrieval tasks will be performed tends to add useful terms to the query, and as a result, tends to improve retrieval effectiveness. To address this question an associative thesaurus was developed using the TREC data (disks 1 and 2) for our TREC3 query processing.

The second question, at what point does query expansion cease to add value, is a real challenge to those studying manual query expansion. The primary concern is that excessive expansion may dilute the original query and result in the retrieval of nonrelevant documents. The answer to this problem is beyond the scope of this report.

A proposed alternative to manual query expansion is automatic query expansion. This alternative is not a very viable option in the on-line service environment because automatic query expansion largely excludes the user from the query formulation process. This may cause the user some confusion since he does not know how the system has modified his query. Another alternative would be to combine manual and automatic query expansion. The user is an intelligent being, not merely a mechanical receptor of data. The user's intelligence can be channeled into the information retrieval system to strengthen the quality of the query and increase its effectiveness.

The primary focus of this report will be on the degree to which query expansion helps improve retrieval effectiveness. Previously published studies of TREC data [1-5] showed improvements ranging from 0% to 20%. We view these findings as inconclusive due to the characteristics of the TREC queries. To those working in the on-line information service industry, the TREC queries are unusually long, structured and descriptive. In other words, the current users of the commercial on-line services rarely, if ever, type in such lengthy natural language queries. To illustrate this point, Table 1 compares the TREC2 and TREC3 *ad hoc* queries to query statistics gathered over a 3 day period by FREESTYLE™, the non-Boolean on-line information retrieval service provided by Mead Data Central, Inc.

Query	Average Length	Longest	Shortest
TREC2	170	319	89
TREC3	105	180	49
FREESTYLE™	7	32	1

TABLE 1. TREC Ad Hoc Queries and FREESTYLE™ Query size (# words)

These basic descriptive statistics indicate the significant difference in query size between the TREC queries and those used in FREESTYLE™. Assuming that these statistics show an order of magnitude difference and that the FREESTYLE™ queries are typical of those processed by most on-line information services, two basic questions are raised. How should the TREC research results be interpreted and generalized? How much can query expansion help improve retrieval effectiveness? To address these questions, we designed and performed a series of experiments using four sets of shorter TREC2 *ad hoc* queries. A subset of these experiments were repeated in our TREC3 efforts.

2.0 TREC2 Query Reduction Experiments

The TREC2 topics (including summary and definition fields) were used as the ideal queries (baseline). Then four different generations of modifications were made to this baseline. First, the summary and definition fields were deleted. The second experiment had the concept field deleted from the base queries. The third experiment was conducted using the recomposed, short version of the queries. This version of the topics are queries consisting of one to three sentences. This process was completed by a professional data analyst who read the entire topic and then composed a shorter version of the topic. The goal was to condense the TREC2 topics down to a size that was closer to that of the FREESTYLE™ queries. The final set of queries contained only the description field of the TREC2 topics. Table 2 compares the sizes of the five generations of topics studied. The retrieval results from our experiments with the short topics may shed some light on the potential risks involved in generalizing the TREC results.

Query	Average Length	Longest	Shortest
Topics	170	319	89
Topics-Sum-Def	138	297	62
Topic-concept	135	273	64
ShortTopic	33	63	18
DescTopic	19	42	6

NOTE: The “-” should be viewed as deletion

TABLE 2. TREC2 Query Reduction

We used the SMART system developed by Cornell University [9] to perform our TREC2 experiments. This system was chosen for its general availability and acceptance over the past 30 years. Specifically, we used the default version of SMART. We did not utilize a customized stop list, phrase capability, sophisticated stemming, or thesaurus of any kind. We used a subset of automatic indexing strategies available in SMART to simulate different indexing systems. The selected indexing strategies are the top 10 ranked according to the 11-point averages in TREC2 (i.e., lnc.ltc, lnc.atc, lnc.mtc, ltc.ltc, mtc.ltc, ltc.atc, ltc.lnc, mtc.atc, ltc.mtc, ltc.nnn [10]). With this design, we could focus our research efforts on the two factors of interest: the size of the queries and the indexing/retrieval methods.

2.1 Impact of Query Size on Retrieval Effectiveness

Accepting the hypothesis that query expansion is beneficial to information retrieval, we may also hypothesize that query reduction would be detrimental to retrieval results. In other words, by demonstrating that query reduction has a negative effect on the retrieval results, we may infer that query expansion would have a positive effect on the retrieval results.

Associated with the query reduction hypothesis is the hypothesis that query reduction has a different degree of negative effect on individual indexing and retrieval systems. In other words, some information retrieval systems that perform effectively with large queries may not perform as well with small queries. In addition, the deterioration rates among the particular indexing and retrieval systems are different in processing a series of gradually smaller queries.

The three tables in this section summarize the experimental results. The numbers were generated by the SMART system based on the top 1000 retrieved documents for each TREC2 test query. The measure in Table 3, 11-point recall average precision, is a recall-biased measure. The measures in Tables 4 and 5 are precision at the fixed rank points, top 5 documents and top 15 documents, respectively, and are precision-biased measures. The deterioration rates, computed using the Topic column as the baseline, are in parentheses next to each precision measure.

In Table 3, deleting the summary and definition fields did not have any noticeable effect on the measure of 11-point recall average. However, when the concept field was deleted, retrieval effectiveness was reduced by an average of 23%. Use of the recomposed, short

TREC2 topics reduced the results by another 10%. Finally, use of the description field as the query reduced the results by still another 18%. The results in Table 3 support the hypothesis that a shorter query is strongly related to significantly lower retrieval performance.

Indexing	Topic	Top-Sum-Def	Topic-concept	ShortTopic	DescTopic
lnc.ltc	0.3405	0.3413(0%)	0.2611(-23%)	0.2208(-35%)	0.1526(-55%)
lnc.atc	0.3199	0.3208(0%)	0.2527(-21%)	0.2093(-35%)	0.1536(-52%)
lnc.mtc	0.3099	0.3127(1%)	0.2464(-20%)	0.2142(-31%)	0.1504(-51%)
ltc.ltc	0.3040	0.3055(0%)	0.2319(-24%)	0.2099(-31%)	0.1588(-48%)
mtc.ltc	0.2953	0.2958(0%)	0.2187(-26%)	0.1921(-35%)	0.1511(-49%)
ltc.atc	0.2952	0.2957(0%)	0.2278(-23%)	0.2093(-29%)	0.1583(-46%)
ltc.lnc	0.2858	0.2926(2%)	0.2057(-28%)	0.1790(-37%)	0.1108(-61%)
mtc.atc	0.2849	0.2839(0%)	0.2168(-24%)	0.1902(-33%)	0.1527(-46%)
ltc.mtc	0.2750	0.2797(2%)	0.2181(-21%)	0.2045(-26%)	0.1567(-43%)
ltc.nnn	0.2662	0.2760(4%)	0.2002(-25%)	0.1727(-35%)	0.1081(-59%)
Average	0.2977	0.3004(1%)	0.2279(-23%)	0.2002(-33%)	0.1453(-51%)

TABLE 3. Impact of Query Reduction: TREC2 11-Point Recall Average Precision

Analysis of the precision biased measures in Tables 4 and 5 produces similar observations to those in Table 3. The absence of summary and definition fields did not have observable impact on either of the two precision results. However, by removing the concept field or by using the two shortened topics, both of the precision percentages showed significant deterioration. Tables 4 and 5 further support the hypothesis that query reduction is detrimental to retrieval effectiveness.

Indexing	Topic	Top-Sum-Def	Topic-concept	ShortTopic	DescTopic
lnc.ltc	0.6200	0.6040(-3%)	0.5320(-14%)	0.3960(-36%)	0.3600(-42%)
lnc.atc	0.5760	0.5640(-2%)	0.5160(-10%)	0.4360(-24%)	0.3480(-40%)
lnc.mtc	0.5320	0.5520(4%)	0.4800(-9%)	0.4040(-24%)	0.3520(-34%)
ltc.ltc	0.5200	0.5240(1%)	0.4320(-17%)	0.3960(-24%)	0.2840(-45%)
mtc.ltc	0.5520	0.5480(-1%)	0.4880(-12%)	0.4480(-19%)	0.3920(-29%)
ltc.atc	0.5240	0.5160(-2%)	0.4560(-12%)	0.3920(-24%)	0.2880(-45%)
ltc.lnc	0.5200	0.5320(2%)	0.4480(-14%)	0.4000(-23%)	0.2920(-44%)
mtc.atc	0.5480	0.5440(-1%)	0.4880(-11%)	0.4440(-19%)	0.3840(-30%)
ltc.mtc	0.4720	0.4760(1%)	0.3960(-16%)	0.3760(-20%)	0.2800(-41%)
ltc.nnn	0.5000	0.4440(-11%)	0.4040(-19%)	0.3760(-25%)	0.2880(-42%)
Average	0.5364	0.5304(-1%)	0.4640(-13%)	0.4068(-24%)	0.3268(-39%)

TABLE 4. Impact of Query Reduction: TREC2 Precisions at top 5 Documents

Indexing	Topic	Top-Sum-Def	Topic-concept	ShortTopic	DescTopic
lnc.ltc	0.5680	0.5667(0%)	0.4933(-13%)	0.4147(-27%)	0.3307(-42%)
lnc.atc	0.5653	0.5693(1%)	0.4853(-14%)	0.4133(-27%)	0.3387(-40%)
lnc.mtc	0.5213	0.5240(0%)	0.4573(-12%)	0.3947(-24%)	0.3173(-39%)
ltc.ltc	0.4947	0.5013(1%)	0.4107(-17%)	0.3800(-23%)	0.2800(-43%)
mtc.ltc	0.5240	0.5160(-2%)	0.4507(-14%)	0.4227(-19%)	0.3493(-33%)
ltc.atc	0.4880	0.4893(0%)	0.4240(-13%)	0.3800(-22%)	0.2827(-42%)
ltc.lnc	0.5147	0.5240(2%)	0.4067(-21%)	0.3520(-32%)	0.2440(-52%)
mtc.atc	0.5027	0.5013(0%)	0.4480(-11%)	0.4213(-16%)	0.3533(-30%)
ltc.mtc	0.4587	0.4653(1%)	0.3787(-17%)	0.3547(-23%)	0.2800(-39%)
ltc.nnn	0.4440	0.5080(14%)	0.3680(-17%)	0.3320(-25%)	0.2347(-47%)
Average	0.5081	0.5165(2%)	0.4323(-15%)	0.3865(-24%)	0.3010(-40%)

TABLE 5. Impact of Query Reduction: TREC2 Precisions at top 15 Documents

2.2 Impact of Query Size on Performance Consistency

Associated with the query reduction hypothesis is the hypothesis that query reduction has a varying degree of negative effect on individual indexing and retrieval systems. In other words, some information retrieval systems that perform effectively with large queries may not perform as well with small queries. To test the hypothesis that information retrieval systems perform inconsistently on different sized queries, a series of simple correlation analyses was performed using the results in Tables 3, 4 and 5. If an information retrieval system performs consistently over an array of gradually smaller queries, performance on large queries should be a good predictor of performance on smaller queries as measured by linear correlation. Stated another way, one could expect a high linear coefficient of determination, i.e. high R^2 value. Conversely, performance of an information retrieval system would be considered inconsistent if it produced a low R^2 value. To illustrate this point, the weak correlations between the retrieval performance of the topic and the recomposed topics in Table 6 suggest that the selected indexing/retrieval systems did not consistently handle the large and small TREC2 queries. The results, however, are preliminary and to some extent inconclusive due to the small sample size and the SMART system constraints.

Correlation	11-Point Avg	Precision at top 5	Precision at top 15
Topic vs. Top-Sum-Def	R^2=0.97	R^2=0.79	R^2=0.75
Topic vs. Top-Concept	R^2=0.85	R^2=0.90	R^2=0.88
Topic vs. ShortTopic	R^2=0.53	R^2=0.25	R^2=0.57
Topic vs. DescTopic	R^2=0.23	R^2=0.47	R^2=0.43

TABLE 6. TREC2: Linear Correlation analysis of indexing/retrieval consistency

3.0 TREC3 Query Expansion and Reduction Experiments

One of the two submitted TREC3 entries, ASSCTV1, was created to further test the two hypotheses that query expansion in general is beneficial to information retrieval and that different retrieval systems receive different levels of help from query expansion. Indexing and retrieval conditions identical to the TREC2 query reduction experiments were maintained, i.e. the default version of SMART and the same set of automatic indexing methods were used. Unlike the TREC2 experiments, the TREC3 *ad hoc* queries were expanded using an associative thesaurus. The expanded queries were loaded into the SMART system to retrieve the top 1000 ranked documents. We then waited for the query relevance information in order to conduct further experiments.

Three additional experiments were performed: one using the TREC3 topics; one using the recomposed, shorter TREC3 queries; and one using only the description field of the TREC3 queries. Table 7 provides the statistics of the original, expanded and recomposed TREC3 *ad hoc* queries. The Expanded Topics results shown in Table 7 reflect an actual query expansion experiment, as opposed to a query reduction experiment. In expanding the TREC3 topics, we tried to maintain the rule to not expand to more than 150% of the original topic size. This rule is merely a guide to prevent over expansion.

Query	Average Length	Longest	Shortest
TREC3 Topics	105	180	49
Expanded Topics	135	194	73
ShortTopics	24	41	16
DescTopics	23	43	9

TABLE 7. TREC3 Query Expansion and Reduction

Tables 8 through 10 describe the impact of query expansion and reduction on the TREC3 *ad hoc* retrieval results. Note that the baseline in these three tables is composed of the searches using the original TREC3 *ad hoc* topics. The heading “Expanded Topic” is for the expansion experiment; the headings “ShortTopic” and “DescTopic” are for the two reduction experiments. In Table 8 the expanded queries enhanced the recall-oriented performance an average of 33%, while the two sets of short queries reduced the performance an average of 34% and 39%, respectively. Moreover, Tables 9 and 10 show the expanded queries and the short queries enhanced or reduced the precision-oriented performances an average of approximately 30%.

Indexing	Topic	Expanded Topic	ShortTopic	DescTopic
lnc.ltc	0.2930	0.3685(26%)	0.1775(-39%)	0.1738(-40%)
lnc.atc	0.2835	0.3551(25%)	0.1787(-37%)	0.0915(-68%)
lnc.mtc	0.2624	0.3375(29%)	0.1737(-34%)	0.1708(-35%)
ltc.ltc	0.2343	0.3181(36%)	0.1564(-33%)	0.1527(-35%)
mtc.ltc	0.2279	0.3001(32%)	0.1674(-27%)	0.1630(-28%)
ltc.atc	0.2261	0.3086(36%)	0.1582(-30%)	0.1557(-31%)
ltc.lnc	0.2067	0.3140(52%)	0.1172(-43%)	0.1149(-44%)
mtc.atc	0.2203	0.2930(33%)	0.1663(-25%)	0.1619(-26%)
ltc.mtc	0.2243	0.2895(29%)	0.1549(-31%)	0.1510(-32%)
ltc.nnn	0.2000	0.2877(44%)	0.1140(-43%)	0.1121(-44%)
Average	0.2379	0.3172(33%)	0.1564(-34%)	0.1447(-39%)

TABLE 8. Impact of Query Expansion/Reduction: TREC3 11-Point Recall Average Precision

Indexing	Topic	Expanded Topic	ShortTopic	DescTopic
lnc.ltc	0.5640	0.7080(26%)	0.3720(-34%)	0.3480(-38%)
lnc.atc	0.5840	0.7000(20%)	0.3840(-34%)	0.1880(-68%)
lnc.mtc	0.4560	0.5520(21%)	0.3560(-22%)	0.3320(-27%)
ltc.ltc	0.3920	0.5680(45%)	0.2400(-39%)	0.1527(-61%)
mtc.ltc	0.5080	0.6720(32%)	0.3760(-26%)	0.3840(-24%)
ltc.atc	0.3960	0.5920(50%)	0.2560(-35%)	0.2600(-34%)
ltc.lnc	0.3920	0.6360(62%)	0.2360(-40%)	0.2160(-45%)
mtc.atc	0.4600	0.6400(39%)	0.3800(-17%)	0.3880(-16%)
ltc.mtc	0.3360	0.4600(37%)	0.2480(-26%)	0.2440(-27%)
ltc.nnn	0.3120	0.4600(47%)	0.2200(-30%)	0.2000(-36%)
Average	0.4400	0.5988(36%)	0.3068(-30%)	0.2712(-38%)

TABLE 9. Impact of Query Expansion/Reduction: TREC3 Precisions at top 5 Documents

Indexing	Topic	Expanded Topic	ShortTopic	DescTopic
lnc.ltc	0.5227	0.6253(20%)	0.3333(-36%)	0.3267(-38%)
lnc.atc	0.5280	0.6147(16%)	0.3427(-35%)	0.1947(-63%)
lnc.mtc	0.4480	0.5267(18%)	0.3173(-29%)	0.3120(-30%)
ltc.ltc	0.3827	0.5347(40%)	0.2827(-26%)	0.2733(-29%)
mtc.ltc	0.4587	0.5840(27%)	0.3573(-22%)	0.3533(-23%)
ltc.atc	0.3960	0.5333(35%)	0.2893(-27%)	0.2800(-29%)
ltc.lnc	0.3827	0.5560(45%)	0.2187(-43%)	0.2253(-41%)
mtc.atc	0.4320	0.5800(34%)	0.3640(-16%)	0.3587(-17%)
ltc.mtc	0.3467	0.4573(32%)	0.2747(-21%)	0.2667(-23%)
ltc.nnn	0.3200	0.4493(40%)	0.2067(-35%)	0.2133(-33%)
Average	0.4218	0.5461(29%)	0.2987(-29%)	0.2804(-34%)

TABLE 10. Impact of Query Expansion/Reduction: TREC3 Precisions at top 15 Documents

Correlation analysis was conducted using the results in Tables 8, 9 and 10. Table 11 summarizes the analysis results. Similar to the TREC2 results, the selected indexing/retrieval methods were inconsistent in processing the large and small queries as indicated by the low R^2 values. Note that the "predictor" in Table 11 is the expanded TREC3 topic instead of the original TREC3 topic. These results are preliminary and to some extent inconclusive due to the small sample size and the SMART system constraints.

Correlation	11-Point Avg	Precision at top 5	Precision at top 15
ExpTop vs. Topic	R^2=0.78	R^2=0.69	R^2=0.74
ExpTop vs. ShortTopic	R^2=0.20	R^2=0.29	R^2=0.18
ExpTop vs. DescTopic	R^2=0.21	R^2=0.25	R^2=0.23

TABLE 11. TREC3: Correlation analysis of indexing/retrieval consistency

The associative thesaurus used in these experiments was compiled automatically using the data that was to be used for the *ad hoc* query portion of TREC3. For this particular version of the thesaurus, the computation took approximately 80 clock hours on a shared SUN Sparc 1000 machine. The actual computation time depends on the number of natural language processing tasks to be performed on the data set. The constructed thesaurus contains uncontrolled index terms. A user interface was built for accessing the thesaurus.The associative thesaurus was also used in the query expansion for another TREC3 entry (ASSCTV2) which was an internal study on search engines.

4.0 Summary

The results of the TREC2 and TREC3 query reduction experiments support the concept that reducing queries is detrimental to information retrieval results. Moreover, the results of the TREC3 *ad hoc* query expansion experiments support the concept that expanding queries using a thesaurus derived from the local data tends to improve retrieval effectiveness. Depending on the length of the original query, the improvement could be very significant. The optimal level of manual query expansion is still an open question. In addition, both sets of experiments suggest that different indexing/retrieval systems perform inconsistently in processing the various sizes of queries. This observation suggests a cautious attitude toward efforts to generalize the current TREC results.

ACKNOWLEDGMNTS: The authors of this paper would like to thank Rita Freese for her analysis work. The hours spent recomposing the queries for this research did not go unnoticed.

REFERENCES:

1. Callan, J. P. and Croft, W. B. “An evaluation of query processing strategies using the TIPSTER collection.” Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.347-356, 1993.
2. Evans, D. A. and Lefferts, R. G. “Design and evaluation of the CLARIT-TREC-2 system.” The Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215, Edited by D. Harman, pp.137-150, 1993.
3. Voorhees, E. M. “On expanding query vectors with lexically related words.” The Second Text REtrieval Conference (TREC-2), NIST Special Publication 00-215, Edited by D. Harman, pp.223-231.
4. Efthimiadis, E. N. and Biron, P. V. “UCLA-Okapi at TREC-2: Query expansion experiments.” The Second Text REtrieval Conference (TREC-2), NIST Special Publication 00-215, Edited by D. Harman, pp.279-289.
5. Buckley, C., Salton, G. and Allan, J. “The effect of adding relevance information in a relevance feedback environment.” Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.292-300.
6. Lesk, M. E. “Word-word associations in document retrieval systems.” American Documentation, Vol.20, pp.8-36, 1969.
7. Sparck Jones, K. and Barber, E. O. “What makes an automatic keyword classification effective.” JASIS, Vol.22, pp.166-175, 1971.
8. Harman, D. “Towards interactive query expansion.” Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.?, 1988.
9. Buckley, C. “Implementation of the SMART information retrieval system.” Technical Report 85-686, Computer Science Department, Cornell University, Ithaca, New York, May 1985.
10. Salton, G. and Buckley, C. SMART 11.0, Computer Science Department, Cornell University, Ithaca, New York, July 1992.